Predicting Radiation Belt Electron Flux with Adaptive Multi-Input Linear Filters

by

E. Joshua Rigler

B.S., Northern Arizona University, 1994

M.S., University of Colorado at Boulder, 2000

A thesis submitted to the

Faculty of the Graduate School of the

University of Colorado in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

Department of Aerospace Engineering Sciences

2004

This thesis entitled:
Predicting Radiation Belt Electron Flux with Adaptive Multi-Input Linear Filters
written by E. Joshua Rigler
has been approved for the Department of Aerospace Engineering Sciences

_____
Daniel N. Baker

_____
Alexander J. Klimas

Date _____

The final copy of this thesis has been examined by the signatories, and we find that both the content and the form meet acceptable presentation standards of scholarly work in the above mentioned discipline.

Rigler, E. Joshua (Ph.D., Aerospace Engineering)

Predicting Radiation Belt Electron Flux with Adaptive Multi-Input Linear Filters

Thesis directed by Professor Daniel N. Baker

The broad goal of the research and results described in this thesis is to improve our ability to make short-term space weather forecasts, and in particular, to predict relativistic electron flux variations over a broad range of altitudes based on measured solar wind inputs. Previous efforts using single-input time-stationary finite impulse response (FIR) linear prediction filters have enjoyed limited success, but generally fail to account for various non-stationary (statistically speaking), and ultimately non-linear dynamical behavior. Two different methods designed to account for the non-stationary radiation belt behavior are studied.

The first assumes poor predictions result from failing to account for all relevant system inputs. Designing linear filters that operate on multiple simultaneous inputs eliminates much bias caused by single-input filters due to simple time-correlations of unmodeled solar wind inputs. FIR profiles (as functions of geomagnetic equatorial altitude, or L-shell) subsequently transition from somewhat smeared functions of space and time-lag, to a set of more specific and impulsive responses that can distinguish between different types of input events. An added benefit is the improvement in prediction efficiencies (PEs) over single-input filters at all altitudes.

The second method for accounting for non-stationary radiation belt behavior involves the adaptive identification of the linear prediction filter coefficients. The well-known Kalman Filter is used to recursively update filter coefficients and minimize prediction error on the fly. This gives a simple linear filter the flexibility to track considerably NON-linear dynamics with time. Not surprisingly, prediction efficiencies improve dramatically at nearly all altitudes for both single-input and multi-input filters. This

improvement is not uniform across all L-shells, but rather tends to scale with regions of the radiation belts that exhibit increasingly persistent behavior. This is expected since persistence allows the Kalman Filter more time to adjust the FIR coefficients to match the current conditions.

One drawback is that the Kalman Filter adapts at different rates for different inputs based on their magnitude and dynamic variability. V_{sw} response functions are therefore able to adapt more quickly to changing conditions, and generally experience more improved PEs than either of the other inputs, when used in a single-input configuration. When adaptive multi-input filter coefficients are calculated, the fact that V_{sw} filter coefficients adapt more quickly causes it to account for an undue share of the electron flux variability. Very good predictions are still provided across a broad range L-shells, but definitive statements regarding the physical cause of the various responses are ill-advised. Modifications to the standard Kalman Filter identification algorithm are suggested in the thesis that address this problem, but they have not been implemented for the studies discussed in this thesis.

Acknowledgements

To start, I must thank Dr. Daniel Baker; my advisor, my teacher, my publicist. My amazement at his ability to subtly guide my research efforts to benefit both my own career and space science in general, is exceeded only by my appreciation for the freedom he gave me to pursue my own interests, and more importantly, to learn from my own mistakes. Unpayable debts of gratitude are owed to Scot Elkington and Janet Green for their infinite patience with often repeated questions regarding fundamental physics of the magnetosphere; and beyond infinite patience with ever-distracting social commentary I so often brought to the office after NPR "parking lot moments" spent yelling at the AM radio in the car on my way to work. Thank you Bill Peterson for teaching me that science need not be devoid of politics or passion, only my results. Perhaps a little more than most, Dimitris Vassiliadis, Alex Klimas, Robert Weigel, and Xinlin Li share my appreciation for the difficulty and thrill of bridging gaps between scientific theory and application. I look forward to the continued honor or working with you all for a long time to come. I am eternally grateful to my father Gedd, my mother Sylvan, my sister Jessica, my best friend Marcia, and the rest of my family and friends for feigning such genuine interest in the Earth's radiation belts. The ability to impress my loved-ones with tales of solar eruptions and radiation storms in the real world kept me going more than anything. Finally, I must acknowledge the role of my favorite "girl" Maggie. Without her unwavering devotion, unconditional love, and cold, wet nose waking me early on dark winter mornings so she could chase the evil squirrels from her yard, I don't know if this day would have ever come. Good girl!

Contents

Chapter

1 Background and Motivation 1

1.1 Introduction 1

1.2 Solar and Geospace Environment 3

1.2.1 The Sun and Solar Wind 3

1.2.2 The Terrestrial Magnetosphere and Solar Wind Coupling . . . 6

1.2.3 Van Allen Radiation Belts 11

1.3 Existing Radiation Belt Electron Models 17

1.3.1 Physical Models 17

1.3.2 Empirical Models 22

1.3.3 Data Assimilation 29

1.4 Non-Linear Dynamics, Local-Linear Filters, and Adaptive System ID . 32

1.5 Thesis Organization 35

2 Solar Wind and Radiation Belt Data 37

2.1 OMNI Solar Wind Parameters 37

2.2 SAMPEX 2-6MeV Electrons 39

2.3 Data Pre/Post-Processing 39

3 Linear System Identification 43

3.1 Time-Stationary Linear Systems 44

3.1.1 Linear Least-Squares Parameter Estimation 45
3.1.2 Recursive Least Squares (RLS) 49
3.2 Time-Varying Linear Systems 50
3.2.1 RLS with Exponential Forgetting 51
3.2.2 Kalman Filter 53
3.3 Chapter Summary 60

4 Electron Radiation Belt Response 62
4.1 Single-Input Models 63
4.1.1 Response Function Profiles 63
4.1.2 Predictions vs. Observations 69
4.2 Multi-Input Models 81
4.2.1 Response Function Profiles 82
4.2.2 Predictions vs. Observations 88
4.3 Chapter Summary 92

5 Future Directions 96
5.1 Data 96
5.1.1 Solar Wind 97
5.1.2 Radiation Belts 99
5.2 Improved Model Structures 101
5.2.1 Dynamic Feedback and Noise Filters 101
5.2.2 3-D Cartesian Coordinates 104
5.2.3 State Space Models and the Extended Kalman Filter 107
5.3 Chapter Summary 110

6 Summary and Conclusions 112

Bibliography 116

Appendix

A Training/Validation Data 126

B Predictions and Residuals 131

C Polynomial-based Linear Prediction Filters 183

C.1 Combined Deterministic/Stochastic Prediction Filters 183

C.2 Dynamical Feedback and Transfer Functions 185

C.3 Common Black-Box Models 187

C.3.1 Pure Stochastic Models 187

C.3.2 Purely Deterministic Models 189

C.3.3 Combined Deterministic and Stochastic Models 190

Figures

Figure

1.1 Ulysses Solar Wind Speed and IMF Direction 5

1.2 Solar Wind and Parker Spiral 5

1.3 Magnetospheric Current Systems and Particle Populations 7

1.4 Antiparallel Magnetic Reconnection and Magnetospheric Convection . 9

1.5 Charged Particle Characteristic Motions in Earth's Magnetic Field . . . 11

1.6 Inner and Outer Electron Radiation Belts (Scematic) 13

1.7 Electron Damage to Space Technology 16

1.8 Static (AE-8) and Quasistatic (CRRESELE) Electron Flux Models . . . 24

1.9 Geostationary 2-6 MeV Electron Log-Flux Response to V_{sw} Impulse . . 27

1.10 NOAA Relativistic Electron Forecast Model (REFM) 27

1.11 Li et al. (2001) Electron Flux Predictions 30

1.12 State and Time-Dependent Local Linearization of Non-Linear Function 34

2.1 OMNI Solar Wind Data Availability 38

2.2 OMNI Solar Wind and SAMPEX Electron Flux Training Data 42

3.1 Stationary Single Input, Single Output (SISO) System ID Block Diagram 44

3.2 Stationary Multi-Input, Single-Output (MISO) System ID Block Diagram 48

3.3 Adaptive Single Input, Single Output (SISO) System ID Block Diagram 51

3.4 Test-Case Stationary FIR System ID with the Kalman Filter 58

3.5 Test-Case Non-Stationary FIR System ID with the Kalman Filter 59

4.1 Stationary SISO FIR profiles (V_{sw}, $|\mathbf{B}|$, and ρ_{ion}) 64
4.2 Representative Solar Wind Inputs and Time Scales 65
4.3 Seasonally Averaged Non-Stationary SISO FIR Profiles 70
4.4 Annually Averaged Non-Stationary SISO FIR Profiles 71
4.5 Stationary and Adaptive SISO FIR Electron Flux Predictions 73
4.6 Stationary and Adaptive SISO FIR Electron Flux Prediction Residuals . 76
4.7 SISO FIR Prediction Residual Auto-Correlation Functions 78
4.8 SISO FIR Prediction Residual Cross-Correlation Functions 79
4.9 Persistence Prediction Residual Cross- and Auto-Correlation Functions 80
4.10 Stationary MISO FIR profiles ([V_{sw}, $|\mathbf{B}|$, ρ_{ion}]) 83
4.11 Seasonally Averaged Non-Stationary MISO FIR Profiles 86
4.12 Seasonally Averaged Non-Stationary MISO FIR Profiles 87
4.13 Stationary and Adaptiive MISO FIR Electron Flux Predictions 89
4.14 Stationary and Adaptive MISO FIR Electron Flux Prediction Residuals 91
4.15 MISO FIR Prediction Residual Auto-Correlation Functions 93
4.16 MISO FIR Prediction Residual Cross-Correlation Functions 93

5.1 Non-Deterministic Stationary SISO System ID Block Diagram 103
5.2 Frequency Response of Electron Flux AR Prediction Filter 103
5.3 Sub-Daily Resolution Electron Log-Flux Response to Solar Wind Inputs 105

A.1 Solar Wind Speed 1994-1999 (Expanded) 127
A.2 IMF Magnitude 1994-1999 (Expanded) 128
A.3 Solar Wind Ion Density 1994-1999 (Expanded) 129
A.4 SAMPEX 2-6 MeV Electron Log-Flux 1994-1999 (Expanded) 130

B.1 Persistence Predictions and Residuals (1994-1999) 132

B.2 Persistence Predictions and Residuals (1994) 133
B.3 Persistence Predictions and Residuals (1995) 134
B.4 Persistence Predictions and Residuals (1996) 135
B.5 Persistence Predictions and Residuals (1997) 136
B.6 Persistence Predictions and Residuals (1998) 137
B.7 Persistence Predictions and Residuals (1999) 138
B.8 Persistence Residuals Correlation Functions 139
B.9 Stationary $V_{sw} \rightarrow \log(\text{flux})$ Predictions and Residuals (1994-1999) . . . 140
B.10 Adaptive $V_{sw} \rightarrow \log(\text{flux})$ Predictions and Residuals (1994-1999) 141
B.11 Stationary / Adaptive $V_{sw} \rightarrow \log(\text{flux})$ Predictions and Residuals (1994) 142
B.12 Stationary / Adaptive $V_{sw} \rightarrow \log(\text{flux})$ Predictions and Residuals (1995) 143
B.13 Stationary / Adaptive $V_{sw} \rightarrow \log(\text{flux})$ Predictions and Residuals (1996) 144
B.14 Stationary / Adaptive $V_{sw} \rightarrow \log(\text{flux})$ Predictions and Residuals (1997) 145
B.15 Stationary / Adaptive $V_{sw} \rightarrow \log(\text{flux})$ Predictions and Residuals (1998) 146
B.16 Stationary / Adaptive $V_{sw} \rightarrow \log(\text{flux})$ Predictions and Residuals (1999) 147
B.17 Stationary / Adaptive $V_{sw} \rightarrow \log(\text{flux})$ Residuals Correlation Functions . 148
B.18 Stationary $|\text{B}| \rightarrow \log(\text{flux})$ Predictions and Residuals (1994-1999) . . . 149
B.19 Adaptive $|\text{B}| \rightarrow \log(\text{flux})$ Predictions and Residuals (1994-1999) 150
B.20 Stationary / Adaptive $|\text{B}| \rightarrow \log(\text{flux})$ Predictions and Residuals (1994) 151
B.21 Stationary / Adaptive $|\text{B}| \rightarrow \log(\text{flux})$ Predictions and Residuals (1995) 152
B.22 Stationary / Adaptive $|\text{B}| \rightarrow \log(\text{flux})$ Predictions and Residuals (1996) 153
B.23 Stationary / Adaptive $|\text{B}| \rightarrow \log(\text{flux})$ Predictions and Residuals (1997) 154
B.24 Stationary / Adaptive $|\text{B}| \rightarrow \log(\text{flux})$ Predictions and Residuals (1998) 155
B.25 Stationary / Adaptive $|\text{B}| \rightarrow \log(\text{flux})$ Predictions and Residuals (1999) 156
B.26 Stationary / Adaptive $|\text{B}| \rightarrow \log(\text{flux})$ Residuals Correlation Functions . 157
B.27 Stationary $\rho_{ion} \rightarrow \log(\text{flux})$ Predictions and Residuals (1994-1999) . . . 158
B.28 Adaptive $\rho_{ion} \rightarrow \log(\text{flux})$ Predictions and Residuals (1994-1999) . . . 159

B.29 Stationary / Adaptive $\rho_{ion} \rightarrow \log(\text{flux})$ Predictions and Residuals (1994) 160
B.30 Stationary / Adaptive $\rho_{ion} \rightarrow \log(\text{flux})$ Predictions and Residuals (1995) 161
B.31 Stationary / Adaptive $\rho_{ion} \rightarrow \log(\text{flux})$ Predictions and Residuals (1996) 162
B.32 Stationary / Adaptive $\rho_{ion} \rightarrow \log(\text{flux})$ Predictions and Residuals (1997) 163
B.33 Stationary / Adaptive $\rho_{ion} \rightarrow \log(\text{flux})$ Predictions and Residuals (1998) 164
B.34 Stationary / Adaptive $\rho_{ion} \rightarrow \log(\text{flux})$ Predictions and Residuals (1999) 165
B.35 Stationary / Adaptive $\rho_{ion} \rightarrow \log(\text{flux})$ Residuals Correlation Functions 166
B.36 Stationary Multi-Input Predictions and Residuals (1994-1999) 167
B.37 Adaptive Multi-Input Predictions and Residuals (1994-1999) 168
B.38 Stationary / Adaptive Multi-Input Component Predictions (1994) 169
B.39 Stationary / Adaptive Multi-Input Predictions and Residuals (1994) . . . 170
B.40 Stationary / Adaptive Multi-Input Component Predictions (1995) 171
B.41 Stationary / Adaptive Multi-Input Predictions and Residuals (1995) . . . 172
B.42 Stationary / Adaptive Multi-Input Component Predictions (1996) 173
B.43 Stationary / Adaptive Multi-Input Predictions and Residuals (1996) . . . 174
B.44 Stationary / Adaptive Multi-Input Component Predictions (1997) 175
B.45 Stationary / Adaptive Multi-Input Predictions and Residuals (1997) . . . 176
B.46 Stationary / Adaptive Multi-Input Component Predictions (1998) 177
B.47 Stationary / Adaptive Multi-Input Predictions and Residuals (1998) . . . 178
B.48 Stationary / Adaptive Multi-Input Component Predictions (1999) 179
B.49 Stationary / Adaptive Multi-Input Predictions and Residuals (1999) . . . 180
B.50 Stationary Multi-Input Prediction Residuals Correlation Functions . . . 181
B.51 Adaptive Multi-Input Prediction Residuals Correlation Functions 182

Chapter 1

Background and Motivation

1.1 Introduction

The walls separating math, science, engineering, and the various subdisciplines within each arena, have grown thinner and thinner with time. When is research considered “pure”; when is it “applied”? Who do we label “theorist”, “experimentalist”, “analyst” or “technician”? With the tremendous leaps made in computational expertise and processing power in recent years, how does numerical modeling fit into the grand scheme of things? Some would say (with good justification) that these distinctions are mostly self-imposed; that they are simply a product of researchers’ desires to stake claims in particular niches within their respective scientific communities and ensure their academic survival. That does not imply, however, that such divisions are non-existent. To the contrary, they are very real, and all too often they hinder scientific progress more than help scientific careers.

One arena in which these walls have actually begun to fall, or have at least been chipped at significantly, is that of terrestrial weather forecasting. This is not without good reason. Meteorology is a discipline in which our scientific understanding is far from complete, thereby justifying continued research into the basic physics. At the same time, however, the social and economic impacts of not being able to make regular and reliable weather predictions can be quite severe, thus justifying the use of (potentially less elegant) empirical specification and forecasting techniques to fill gaps

resulting from our incomplete scientific understanding. Engineers, scientists, and mathematicians, have been forced to delve into each other's scientific history and literature and, more and more frequently, to resort to interdisciplinary collaborations in order to find solutions to their respective problems.

More recently, due primarily to the combined phenomena of an increased number of geospace observation platforms and an increased societal dependence on technologies that are susceptible to the extreme temporal and spatial variations in observed geomagnetic properties, there has been a similar breaking down of walls separating subdisciplines in aerospace engineering, space physics, and meteorology. This has led to a new scientific classification popularly referred to as "space weather". Within this relatively new field, a premium has been placed on the accurate specification and short-term prediction of geomagnetic properties that is nearly equal to the desire for a full and self-consistent understanding of the underlying physics. Neither of these worthy goals will be fully realized, however, unless the interest and expertise available on all sides can be exchanged, enhanced, and ultimately extended to help answer as yet unforeseen scientific questions.

This thesis details one aspiring engineer/scientist/mathematician's humble efforts to knock a brick or two from walls separating those fields he finds so interesting. The first chapter provides background and motivation for the space weather research and modeling efforts discussed in subsequent chapters. Section 1.2 is a rudimentary description of the space physics and phenomenology associated with the solar, interplanetary, and geospace systems. Particular attention is paid to the Earth's Van Allen radiation belts and their space weather implications. Some of the limitations of our current physical understanding of this complex system are addressed, thus indirectly providing scientific motivation for much of the research presented herein. Section 1.3 reviews the various types of numerical models used thus far in efforts to characterize the Earth's electron radiation belts, and without whose prior example and motivation the current

research would be impossible. Section 1.4 introduces the concept of "local-linear" models, with an emphasis on distinguishing prior work from the adaptive models that are the primary subject of this thesis. Finally, Section 1.5 describes the structure and outlines of the remainder of this thesis, placing an emphasis on the clean separation of background material, methodology, results, and interpretation, so that the the document will be as broad of interest and utility as possible.

1.2 Solar and Geospace Environment

This section provides a scientific overview of all the major systems playing a role in the energization of relativistic electrons in the Van Allen radiation belts. It is neither all encompassing nor overly focused on any particular topic. It is fairly strong on phenomenology and heuristics, and deliberately light on mathematics beyond describing a few fundamental physical relationships. It might easily be skipped by those already indoctrinated into the field of space science, but should serve as motivation for the empirical modeling efforts described later in this thesis in that it makes clear the incredibly complicated nature of the physics most pertinent to space weather modeling.

1.2.1 The Sun and Solar Wind

The sun is the ultimate energy source driving all space weather. Its surface temperature is approximately 5700 K, producing a nearly black-body electromagnetic (EM) spectrum which peaks in the visible wavelengths near $\lambda \approx 500$ nm. The integrated solar spectrum, or total solar irradiance (TSI), provides over 1360 watts of EM radiation per square meter at 150,000,000 kilometers, the distance from the sun to the Earth, or 1 astronomical unit (AU). Since the Earth's orbit is not perfectly circular, the impinging EM radiation varies approximately 7-8% over the course of one year. Otherwise, the solar EM power flux is extremely stable, if not necessarily well established experimentally (estimates range from 1362 to 1368 watts per square meter). Sunspots

(darker regions on the sun's surface that are slightly cooler than the surrounding area) can cause ~0.2% short-term variations in the TSI as they transit the solar disk. These and related solar variations exhibit an 11 year variation in number and magnitude referred to as the solar cycle, culminating in a ~0.1% TSI variability (White and Rottman, 2000; Rottman, 2002).

From a space weather perspective, ultraviolet (UV) wavelengths (i.e. $\lambda < 380$ nm) are probably the most important solar EM radiation because they are responsible for most of the heating and photoionization of molecules in the Earth's upper atmosphere, dramatically affecting both its electrical and fluid/thermal characteristics. UV irradiances have similar periodic variations to the TSI, but they typically exhibit much larger relative changes in magnitude – sometimes on the order of 10-100% (depending on specific wavelength; Rottman et al., 2001). UV irradiance can also become extremely enhanced for very short durations associated with impulsive events like solar flares. On average, however, UV irradiance constitutes less than 7% of the TSI, and the wavelengths most pertinent to space weather (~10-100 nm) are less than 1%.

The sun also ejects over 10^7 tons of hot plasma (mostly ionized hydrogen and electrons) per year in the form of the supersonic solar wind. While this amount of material sounds impressive, the sun's nuclear furnace will run out of hydrogen fuel long before any significant reduction in the sun's total mass (~2×10^{30} tons) will ever be seen. Even though the exact physical mechanisms are still a topic of intense scientific study, the energy source responsible for accelerating this material to escape velocities and beyond can be found in the complex magnetohydrodynamic (MHD) interactions taking place in the solar corona, a region of super-heated plasma extending beyond the solar surface many millions of kilometers (Kivelson and Russell, 1995). In the sun's polar regions, magnetic field lines are effectively open and fairly straight, allowing the solar wind to escape easily at high velocity. Nearer to the solar magnetic equator, however, the fields are closed, thereby slowing the outflow to less than half its speed

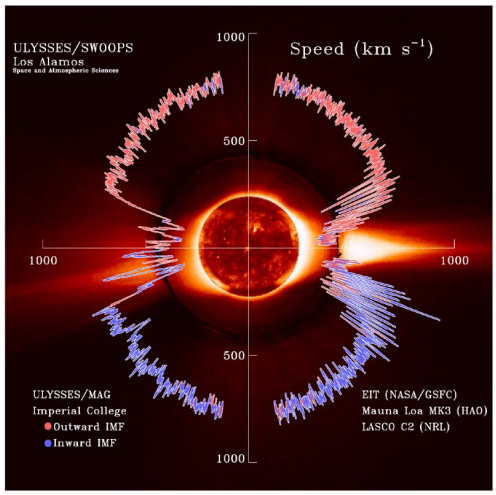


Figure 1.1: Solar wind speed and IMF direction as a function of solar magnetic latitude from the Ulysses spacecraft (from McComas et al., 1998).

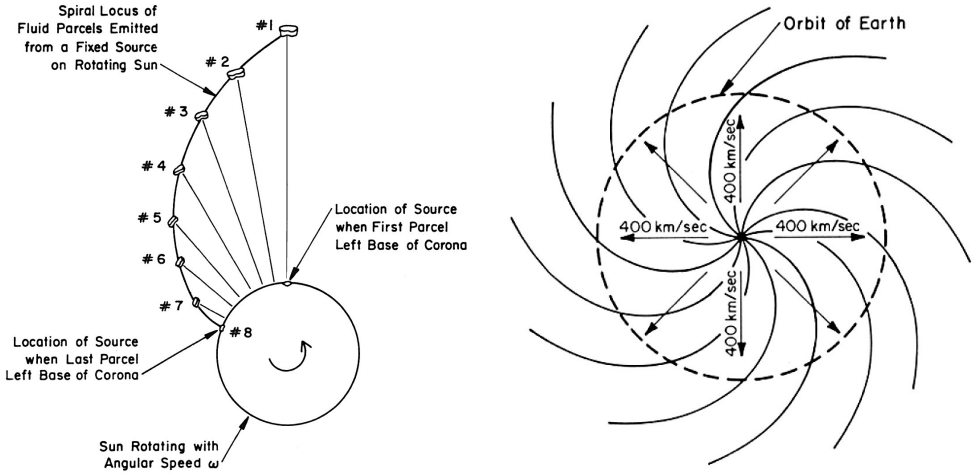


Figure 1.2: A succession of fluid parcels emitted at constant speed from a source fixed on the rotating sun is shown on the left. On the right is the resulting “Parker spiral” for moderate solar wind speeds (from Kivelson and Russell (1995)).

in the polar regions, as evidenced in the Ulysses data shown in Figure 1.1. The lost momentum of the equatorial plasma is actually transferred to the solar magnetic flux, pulling field lines far out into the solar system and creating the interplanetary magnetic field (IMF). Due to conservation of angular momentum of the plasma, the entrained field lines are stretched and twisted into a spiral pattern. As seen in Figure 1.2, the angle of the IMF is then a function of solar rotation period, orbital position, and solar wind speed. This is approximately 45° at 1 AU given moderate solar wind speeds.

With speeds averaging ~450 km/s, temperatures on the order of 10^5 K, ion number densities of ~7 cm^{-3}, and magnetic field intensities around 7 nT at 1 AU, the kinetic, thermal, and magnetic power flux levels found in the solar wind (approximately 6×10^{-4}, 2×10^{-5}, and 1×10^{-5} watts per square meter, respectively; Kivelson and Russell, 1995) are considerably less than those of solar UV radiation . It is generally agreed, however, that the terrestrial magnetosphere's dynamics are driven primarily by its energetic coupling with the solar wind and IMF. This apparent paradox is at least partly reconciled in the following section.

1.2.2 The Terrestrial Magnetosphere and Solar Wind Coupling

When the magnetized solar wind begins to interact with the Earth's magnetic field, what would otherwise be a roughly dipole configuration is transformed into a bullet-shaped cavity as a balance is achieved between the dynamic pressure of the solar wind plasma, and the magnetic pressure of the Earth's magnetic field. As shown in Figure 1.3, a multitude of different complicated and interacting particle populations, current systems, and interesting magnetic field configurations exist within this "magnetosphere". While the internal dynamics of this system are surely quite complicated, and warrant extensive study in their own right, the internal energy production is nowhere near sufficient to drive the tremendous variability observed at all spatial and time scales. The only two realistic external drivers of this system are solar EM radiation and the so-

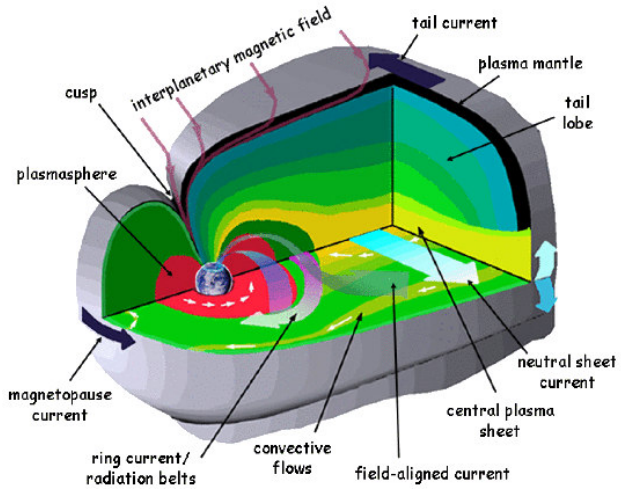


Figure 1.3: A schematic of the Earth's magnetosphere, including major current systems and plasma regions (courtesy S. Elkington).

lar wind, described previously.

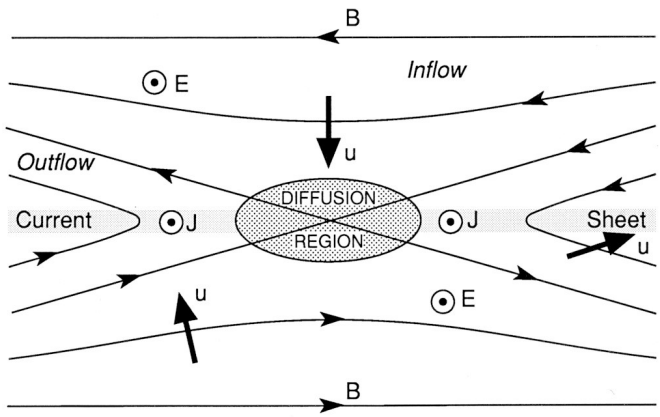
Despite the fact that, even alone, UV radiation possesses much more power than what is available in the solar wind, the general scientific consensus is that the solar wind is actually the primary power source driving magnetospheric dynamics. One reason for this is related to the relative areas of the interaction regions for each type of power input, or their coupling cross-section. Neglecting exceptions that are insignificant to the problem at hand, photons are effectively neutral particles, and unaffected by magnetic or electric fields. In fact, photons must either collide, or pass extremely close to, individual atmospheric molecules in order to transfer their energy. Therefore, the cross-section for UV energetic coupling with the Earth is only slightly larger than the radius of the Earth due to low particle density (average mean free path is ~160 km at 800 km altitude; Tascione, 1994) makes the probability of a photon hitting an atmospheric molecule almost zero beyond a few hundred kilometers altitude.

The solar wind, on the other hand, is comprised almost entirely of charged particles that have a much more extended area of interaction, due to their ability to interact

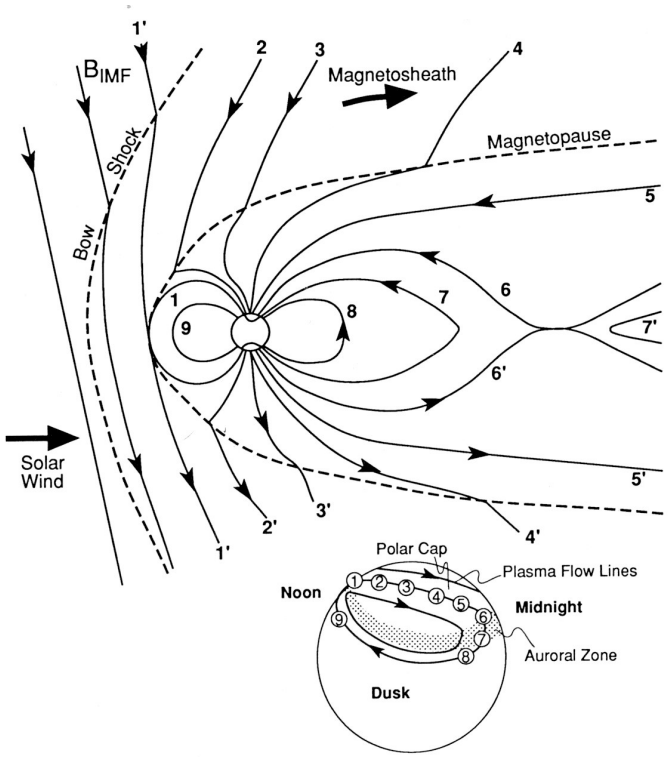
with electric and magnetic fields. This means that not only can solar wind particles interact with the fields found in the Earth's extended magnetosphere (effectively increasing the coupling cross-section by several orders of magnitude), but that their energy can be absorbed and transmitted across and along the field lines as waves that penetrate deeply into the magnetosphere, affecting the plasma properties there as well. These waves result, in part, from pressure variations inherent to the solar wind. There can also be viscous boundary-layer dynamics creating fluid instabilities in the form of Kelvin-Helmholtz waves, which, while primarily a function of solar wind speed, may be tied to its thermal properties as well (Vasyliunas et al., 1982).

Still, the general scientific consensus is that these fluid-like interactions are responsible for a small fraction of the total energy that is transferred from the solar wind to the magnetosphere, leaving only the magnetic energy flux, the least of the solar wind's available power sources, to drive nearly all observed geomagnetic activity. This perhaps unexpected conclusion is drawn from correlation studies designed to balance estimates of the various types of available solar wind power with estimates of the energy dissipation in the magnetosphere (e.g. Arnoldy, 1971; Perreault and Akasofu, 1978; Akasofu, 1979, 1981). These studies suggest that most of the solar wind-magnetospheric energetic coupling is actually in the form of Poynting flux associated with the day-side reconnection of antiparallel IMF and terrestrial magnetic field lines, so it is prudent to describe this process in some detail before proceeding.

First, consider antiparallel magnetic field lines in the presence of a uniform electric field $\mathbf{E}$ as shown in Figure 1.4(a). Charged particles will tend to move in the $\mathbf{E} \times \mathbf{B}$ direction, or toward the center of the figure. Frozen-in flux conditions require that the magnetic field lines be carried with the particles, so they will be pulled toward the center where they counteract the oppositely directed field line, reducing the local $\mathbf{B}$-field strength to near zero, and allowing a thin current sheet to form. In the so-called diffusion region, the $\mathbf{B}$-field strength falls to something that is effectively zero, allowing



(a) Anti-parallel magnetic fi eld reconnection.



(b) Time-sequence of plasma parcels in a reconnection driven magnetospheric convection pattern.

Figure 1.4: From Kivelson and Russell (1995).

the random motions of individual particles to carry them onto new field lines, thereby breaking and reconnecting what were previously antiparallel fields. A highly kinked magnetic field line now exists on either side of the diffusion region. Magnetic tension will force these field line to straighten, introducing a tremendous amount of free energy to the system.

Now, consider the IMF embedded in the solar wind impinging on the Earth's magnetic field in a similar fashion to Figure 1.4(b), so that the influx of **B** is continuous at the day-side magnetopause. If a southward component exists in the IMF there will be reconnection as just described. The reconnection results in IMF field lines merging with terrestrial magnetic field lines, providing solar wind plasma direct access to the Earth's magnetosphere. A combination of the release of the magnetic tension in these newly kinked field lines, and the fact that the IMF is being carried past the Earth by the solar wind, pulls these "open" field lines across the pole and causes them to pile up in the tail. If enough of these antiparallel field lines build up, another reconnection region develops in the magnetotail which serves to re-close terrestrial magnetic fields. These newly closed field lines are also highly kinked and want to straighten, but the simultaneous build-up of more newly-closed lines forces them to the side where they migrate back toward the day-side to start the whole process over again.

The bulk of observational evidence (both direct and using derived indices) is consistent with this reconnection-driven convective description of solar wind-magnetospheric interaction. It must be noted, however, as Vasyliunas et al. (1982) showed using a very thorough dimensional analysis, that the difference in energetic input between this kind of driving and more fluid-like viscous interactions is proportional to $(V/\rho)^{\frac{1}{6}+\frac{b}{3}}$ (V is the solar wind velocity; ρ is the mass density of the solar wind plasma; b is an empirically determined exponent that differentiates between viscous dissipation in ordinary hydrodynamic flows and magnetohydrodynamic flows). If $b = 0$, viscous interactions are not affected by the magnetic fields. Given the typically limited range

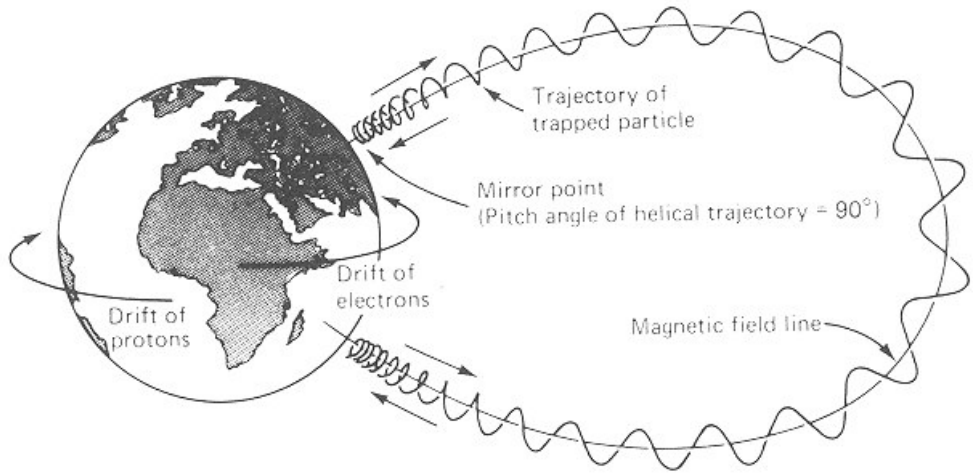


Figure 1.5: Schematic of the three characteristic motions of charged particles (gyration, bounce, and drift) within the Earth's magnetic field (from Stern and Ness, 1981).

of variation in $(V/\rho)^{\frac{1}{6}}$, the statistical separation of Poynting flux energetic input and viscous dissipation at the magnetopause can be quite difficult. Indeed, more recent studies, especially those of the Van Allen radiation belts, suggest that ULF wave activity driven by Kelvin-Helmholtz instabilities near the magnetopause plays a crucial role in enhancing, if not directly driving, the acceleration of various charged particle populations inside the magnetosphere (Rostoker et al., 1998; Baker et al., 1998, 2001).

1.2.3 Van Allen Radiation Belts

Charged particles exhibit three fundamental types of motion in the presence of a magnetic field **B**, as shown in Figure 1.5. First, the particle will experience a Lorentz force perpendicular to both its direction of motion and the direction of **B** that is directed by the particle's charge ($\mathbf{F_L} = q\mathbf{v} \times \mathbf{B}$). Since this acts to change its direction of motion, a singly-charged particle traces out a circular path in the plane perpendicular to **B** with a gyroradius that is proportional to its perpendicular momentum divided by the magnetic field strength and particle charge ($r_c = mv_\perp/|q\mathbf{B}|$). The gyrofrequency, or rate at which

the charged particle gyrates about a field line, is therefore independent of its velocity ($\omega_c = |q\mathbf{B}|/m$). Assuming an average strength of $\sim 10^{-5}$ T for the Earth's magnetic field, the gyroperiods are about 6.6ms and 3.6μs for protons and electrons respectively. An adiabatic invariant μ associated with this motion requires that the magnetic moment ($1/2(mv_{\perp}^2)/|\mathbf{B}|$) remain constant over a single gyration. This requirement is usually met within the Earth's relatively strong magnetic field, but high frequency electric or magnetic waves, and/or strong curvature in the **B**-field, can accelerate the particle on time-scales comparable to its gyroperiod ($2\pi/\omega_c$) and violate the conservation μ.

The second type of particle motion manifests when there are components of the gradient in magnetic flux ($\bigtriangledown\Phi$) that are parallel to **B**, and the charged particle has some velocity in the same direction. This implies a convergence of field lines about which the particle gyrates. At any given instant in time, the Lorentz force is directed toward the instantaneous center of gyration of the particle. Since the field lines are converging, however, the vector quantity $\mathbf{F_L}$ must be integrated over the the entire gyration, resulting in a component of the Lorentz force directed oppositely to the increasing magnetic flux. Another way to look at this is that the particle experiences an electromotive force, according to Faraday's law of induction (i.e. EMF $= -d\Phi/dt$), that serves to slow the particle's motion parallel to **B** and eventually reverse its direction (Tascione, 1994). In a dipole field, this "mirror" force causes a charged particle to bounce back and forth between the poles. An adiabatic invariant J is associated with this motion, and dictates that the parallel momentum integrated over one bounce period remains constant. If the particle experiences anything but perpendicular acceleration during a single bounce, J will not be conserved. Since bounce periods are usually much longer than gyroperiods for particle's with similar energy, this invariant is violated more easily than μ.

The third type of particle motion results from components of $\bigtriangledown\Phi$ that are perpendicular to **B**. This causes a particle's instantaneous gyroradius to be larger when it is in a region of lower field strength, and smaller when it is in a region of higher field

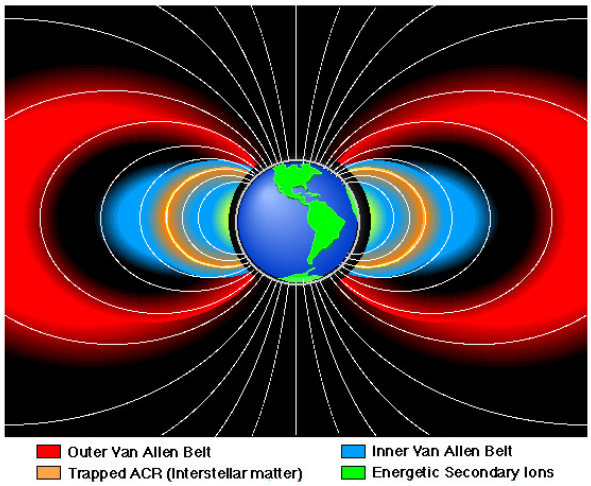


Figure 1.6: A schematic of the Earth's radiation belts in an idealized dipole magnetic field (from Mewaldt et al., 1994).

strength. The gross effect is that the particle drifts, according to the sign of its charge, in a direction perpendicular to both **B** and $\nabla\Phi$. For the roughly dipole field of the Earth this means that equatorial particles (i.e. particles with no momentum parallel to **B** that reside near the magnetic equator) trace out a circular path around the Earth. A third adiabatic invariant φ is associated with this drift motion that dictates that the total magnetic flux enclosed by the drift path remains constant, allowing the particle to return to its origin. The drift period is a function of particle energy, and is always much longer than either the gyro- or bounce-period for the same energy particle, usually on the order of several minutes in the Earth's magnetic field. The conservation of φ is therefore often violated since the Earth's field can change considerably on this time-scale.

A balance of time and spatially-dependent acceleration and loss mechanisms, combined with these three types of charged particle motion in an idealized dipole field, results in charged particle distributions that are organized by magnetic field lines into torus-like shells around the Earth called the Van Allen radiation belts (see Figure 1.6). This highly structured magnetic organization leads to the regular use of a reduced set of

geomagnetic coordinates, L and B/B_0. The so-called L-shell is defined as the distance, in Earth radii (R_E), that a given drift path is from the center of the Earth when the particle passes through the geomagnetic equator during a single bounce between mirror points. B/B_0 is simply the ratio of the local magnetic field intensity to the intensity of that same field-line at the geomagnetic equator.

Due to non-resonant gyroperiods, and in general to their large mass (relative to electrons), radiation belt protons are not easily perturbed from their nominal trajectories along magnetic field lines by high-frequency waves present in the magnetosphere. As a consequence, their populations are well organized into a single belt with a peak whose altitude is, for the most part, inversely energy-dependent due to the fact that the large gyroradii of very energetic protons tend to carry them beyond the confines of the Earth's magnetic field. For example, the peak in the distribution of >1 MeV protons is near L=3.0 R_E, while the peak for >100 MeV protons is nearer to L=1.5 R_E (Tascione, 1994). The proton radiation belts are very stable, exhibiting only slow and easily predicted changes over time.

Electrons, on the other hand, exhibit a two-peak distribution in altitude, at least at higher energies. The "inner belt" is typically confined to altitudes less than L=2 R_E, while the "outer belt" is spread broadly between $L \approx$3-7 R_E. This is not due to any inherent difference in the fundamental physics governing the dynamics of this particle population. In fact, a perfectly static magnetic field with no high-frequency waves present would result in an electron belt with a very proton-like single peaked distribution. However, the Earth's magnetosphere tends to exhibit resonant behavior that coincides with the gyrofrequencies of high-energy electrons at equatorial altitudes around two to three times its diameter. This accelerates the electrons, often in a direction parallel to the magnetic field (violating the second adiabatic invariant J), causing their mirror point to drop to lower altitudes where they interact with the denser ionosphere and neutral atmosphere and are lost from the radiation belts. This depletion zone is

relatively permanent, and the term “slot region” is often used to describe this boundary between the inner and outer electron radiation belt.

The inner belt high-energy electrons, like the protons, are relatively stable, with an altitude-dependent distribution defined almost entirely by their energy. The outer belt, however, is much more dynamic. The exact set of mechanisms responsible for heating outer belt electrons is still largely unknown. The direct acceleration of particles via resonant wave activity is a likely source of electron energization, although, as was already mentioned, such wave-particle interactions may also scatter electrons into the Earth’s atmosphere. On the other hand, while electron fluxes typically fall off as altitude increases, the phase space density is often seen to increase. Assuming stochastic motion in these particle distributions, a general radial diffusion of electrons inward will take place. This violates the third adiabatic invariant φ, but, if μ is still conserved, the particle populations necessarily gain energy as they move to regions of higher magnetic field intensity. The question remains, however, whether or not this diffusion is fast enough to explain the rapid changes observed in high-energy electron fluxes in the outer radiation belt. A growing body of theoretical and observational evidence indicates that ultra-low frequency (ULF) wave activity can enhance the radial diffusion rates. The fact that ULF wave activity is both theoretically and observationally related to solar wind parameters, combined with similar observed correlations between electron flux enhancements and changes in solar wind parameters, suggests that radial diffusion is indeed the primary acceleration mechanism for radiation belt electrons, at least at geostationary (L=6.6 R_E) altitudes. Loss mechanisms for these particles are of equal importance, but have only very recently received a similar level of scientific scrutiny.

In addition to the challenging scientific questions that might be addressed by studying this dynamic region, the outer electron belt is also of significant engineering interest. Free electrons pose a number of potential dangers to satellites and their various sub-systems and instrument payloads. Figure 1.7 illustrates three of these. Single-

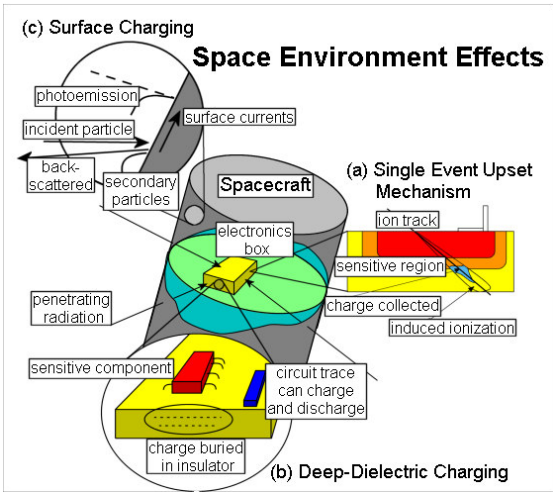


Figure 1.7: Three types of electron-related satellite damage: a) single-event memory upsets; b) deep-dielectric charging; and c) surface charging (from Baker, 2001).

event upsets are very common, but redundancy and/or error-correcting memory make this a relatively simple problem to address. Likewise, surface charging, whether due to photo-ionization or low-energy electron charge buildup, can be alleviated by carefully designing the satellite to quickly and evenly redistribute charge. Deep-dielectric charging, however, is very difficult to mitigate since it involves the embedding of electrons deep in insulating material that was actually designed to prevent redistribution of charge. If enough electrons embed themselves in this material, eventually the resistance of the material will break down, and a strong electrostatic discharge will occur, resulting in potentially permanent damage to vital electronic components. So-called “killer electrons”, with energies ranging from hundreds of KeV to many MeV, are the primary cause of deep-dielectric charging.

As it turns out, the outer electron belt envelops the orbits of some of the most societally and commercially important satellites flying today. This includes the United States Department of Defense’s Global Positioning System (GPS) satellites (L=4.2), which were originally of only military interest, but in the last decade or so a multi-

billion dollar industry related to GPS technology has emerged around the world. It also includes geostationary satellite orbits (L=6.6), including hundreds of communications, weather, and a defense/intelligence related satellites. The growing complexity and capability of modern spacecraft systems is striking. Increasingly extensive use of very large-scale integration in space electronics has occurred, and much faster and more powerful processors, controllers, and computer memory are likely to be employed in the future. It is clear that as capabilities of electronics go up, the susceptibility to radiation effects increases. In this sense, the decreasing "hardness" of electronic components in space is of great societal concern, and this trend only promises to continue.

1.3 Existing Radiation Belt Electron Models

The need for a high-quality forecast model for the outer electron radiation belt has certainly not gone unnoticed in recent years. On the contrary, there are probably more different models attempting to simulate this environment than just about any other type of space weather. This section discusses several different general types of electron radiation belt models, as well as specific implementations of each. Each type is presented roughly in order of their relative prediction capabilities to date, followed by a brief discussion of how various forms of data assimilation have been employed to address the shortcomings of each.

1.3.1 Physical Models

Arguably, physics-based computer models should provide both the best long and short-term forecasts of physical systems since they are ideally based on exact, or at least unchanging, mathematical descriptions of reality. It is strange then that they should be discussed first in a list of model-types with increasing prediction capabilities. One problem is that, although we have a very good understanding of the most fundamental physics of electricity and magnetism (i.e. Maxwell's equations), these relationships can

only be solved for simple isolated systems. Extending them in an effort to reproduce the dynamics of each and every particle and/or field becomes computationally prohibitive at anything beyond the smallest scales.

Simplifying assumptions need to be made when studying electromagnetic phenomena in complex systems, and these are usually only valid over a very limited range of operating conditions and/or spatio-temporal scales. The terrestrial magnetosphere exhibits such an incredibly broad range of operating conditions and scales that the complicated interplay of different dynamic processes makes it extremely difficult to include them all in the same model. This is especially true for radiation belt electrons which, because of their low densities and relativistic properties, have historically been deliberately neglected in more traditional theoretical plasma physics models (i.e. magnetohydrodynamics (MHD)).

Particle Tracing

Attempts to model radiation belt electrons have not, so far, included any major revisions of MHD theory. Rather they generally fall into one of two categories, both of which rely on a magnetic field model that is mostly if not entirely independent of the relativistic electron fluxes themselves. The first approach involves integrating the trajectories of a population of individual electrons through the magnetic field as if they were isolated particles. Often these particle tracing codes assume conservation of the first and/or second adiabatic invariant, simplifying the model by reducing the number of degrees of freedom involved in calculating the particle's trajectory. This leads to the so-called "guiding center" approximation which tracks the center of gyration of a charged particle about a field line only, rather than having to calculate its trajectory through each gyration.

Of course one drawback to a particle pushing approach to radiation belt modeling is that the number of particles that can be included in the simulation is always going to

be inadequate to match reality. The only solution, given current or foreseeable computational power, is to integrate the trajectories of a fixed number of particles, and apply a weighting factor. Weighting factors are designed to eliminate bias resulting from choice of initial particle distributions (for example, if particles are initially arranged in azimuthally symmetric configuration about the Earth, there will be a greater particle density nearer the center than farther away) and initial energy distributions, while simultaneously accounting for proximity to virtual detectors, conservation of phase space density (Liouville's theorem; Roederer, 1970), and initial flux as determined a priori by a static radiation belt model (Elkington et al., 2002).

Particle tracing radiation belt models have led to improved understanding of electron acceleration due to strong impulsive events in the solar wind. Li et al. (1993) used an analytic magnetic field and pulse model designed to mimic the results of unknown solar wind conditions that led to an intense enhancement in radiation belt electrons during the storm sudden commencement (SSC) of the March 24, 1991 geomagnetic storm. More realistic magnetic field models based on MHD output have since been used to model this event, as well as more "ordinary" sub-storm and storm-time electron dynamics (e.g. Elkington et al. 2002, 2004a,b). These studies were all restricted to equatorial mirroring electrons. More recent work by K. Perry at Dartmouth University, and X. Shao at NASA's Goddard Space Flight Center (GSFC) holds much promise in understanding the dynamics of off-equatorial relativistic electrons, providing a true three-dimensional description of radiation belt fluxes.

In addition to the weighting functions described above, the simulation results will also need to be scaled to match instrument viewing geometries and efficiencies for direct comparison with the observed fluxes. Graphical comparisons between model output and real electron fluxes are provided in some of the papers listed above, showing reasonable qualitative agreement between simulation and reality, no mean feat considering most simulations were performed during geomagnetically active periods. This

fact in itself, however (i.e. that simulations typically last only one to a few days) makes it difficult to conduct comprehensive and quantitative residual analyses and determine meaningful metrics for evaluating the operational value of such models.

Particle Diffusion

The second physics-based modeling approach assumes that radiation belt electrons belong to well-defined populations of particles that can be statistically organized according to their phase-space density (PSD). Perturbation theory then describes how these populations might become energized as they violate adiabatic invariants while being slowly moved through a gradient in the PSD. This is essentially a stochastic process and is analogous to diffusion in non-magnetic fluids. A typical scenario involves assuming that PSD gradients only exist perpendicular to magnetic field lines, and that the third adiabatic invariant is violated while the first and second are conserved. Since these field lines are assumed to map to the magnetic equator, this results in a type of "radial diffusion" that can energize electrons assuming that PSD increases as one moves to higher L-shells (Falthammer, 1965, 1966; Schulz and Lanzerotti, 1974).

Radial diffusion is usually sufficient for describing the stochastic acceleration of protons, which are more or less immune to pitch-angle scattering due to their relatively high mass (and corresponding low gyrofrequencies). Electrons on the other hand, are easily affected by high-frequency waves present in the magnetosphere, making them quite sensitive to pitch-angle diffusion. Four-dimensional diffusion theory (three dimensions corresponding to the three adiabatic invariants, and one a local-time dimension accounting for magnetospheric convection) has been used to produce a model capable of accounting for both radial and pitch-angle diffusion in the radiation belts (Salammbo; see Boscher et al., 1996; Bourdarie et al., 1997a,b).

Neither pure radial diffusion, nor the more complicated 4-D diffusion codes, are currently capable of reproducing the high time-variability typically observed in the ra-

diation belts. The possible causes are manifold. First, most existing models simply assume a constant, or at least simplified source that maintains PSD gradients that will result in diffusion toward regions of stronger magnetic fields, thereby always accelerating electrons. This in itself is somewhat unrealistic, as studies show that PSD can vary greatly in the outer regions of the magnetosphere, resulting in both positive and negative gradients depending on the current conditions and or the energy of the particles being studied (Brautigam and Albert, 2000; Green and Kivelson, 2004).

Another poorly understood component of any diffusion-based model are the losses that must be accounted for. The possible mechanisms include adiabatic changes, resulting from changed magnetic field configurations; magnetopause crossings, in which electrons drifting on higher L-shells are actually lost to open field lines as they drift from midnight toward noon; and finally, pitch-angle scattering that changes the trajectories of previously trapped electrons so that their mirror points fall inside the Earth's atmosphere and the particles are lost to collisions with neutrals. The first mechanism is not a true loss, as populations should return to normal once the magnetic field configuration relaxes to normal. Recent work by J. Green and T. Onsager has attempted to better determine which of the second two mechanisms is more responsible for the permanent loss of radiation belt electrons, and the published results are eagerly awaited by this researcher. Usually loss is parameterized in a diffusion model, and treated as an exponential decay with a realistic time constant (e.g., Shprits and Thorne, 2004).

A problem with any diffusion code is that the equations used to describe the evolution of the system can only be analytically derived using simple static magnetic field configurations to provide fixed diffusion coefficients. In real-world situations, these coefficients are not time stationary, and must be determined empirically. This has been done parametrically, based on the Kp index (Brautigam and Albert, 2000), and analytically, by making the diffusion coefficients functions of ULF wave power derived from MHD simulations (Elkington and Fei et al., personal communication).

1.3.2 Empirical Models

Empirical is defined here in a very broad sense. All the models discussed in this section share the common trait that a significant amount of data were collected and used to determine all or part of the model itself. Data were not simply used to establish initial or boundary conditions to a theoretical model, although there is no reason this could not be done as well. The line between physics-based, or otherwise theoretical models, and empirical models is at times quite fuzzy, as evidenced from previous descriptions of time-variable diffusion coefficients.

Static and Quasi-Static Statistical Models

Static statistical models are the simplest characterization of any system that can actually be used to estimate what that system will look like in the future. They make no use of the recent history of the system's state, nor do they employ external input to drive changes in the system. In other words, they exhibit absolutely no dynamical behavior whatsoever. They simply attempt to determine the mathematical expectation of a particular system variable based on observations. This is not to say that a sophisticated set of conditions cannot be externally imposed on the mathematical expectation that allow it to give different estimates at different times. Conditional probabilities (explicit or implied) are often employed in the creation of so-called quasi-static models that can reproduce time variations in the system variable with considerable accuracy. The primary difference between a static or quasi-static model and a truly dynamic model is that the dynamic model is at least theoretically capable of producing output that falls outside the bounds of the existing measurements of the system, while even a quasi-static model remains constrained by the limits of the data used to train it.

The most widely used model for the radiation belt electrons is NASA's AE-8 trapped electron radiation belt model. This is the most recent incarnation of a series

of AE models that started in the mid-1960s and has evolved to make use of improved mathematical techniques, physical understanding, and available satellite data. It is comprised of three parts, each based on a different set of assumptions used to map pertinent observations into a three-dimensional L vs. B/B_0 vs. energy coordinate system (Vette, 1966, 1991). These parts include an inner zone covering a range of L-shells from 1.2 to 2.4, a transition region covering L-shells from 2.4 to 3.0, and an outer zone covering L-shells from 3.0 to 11. Local time mapping functions have been determined in order to account for local time variations in the fluxes, but these are only valid in a highly time-averaged sense due to the extreme short-term variability seen at each local time (factors of 10 to 50 are not uncommon; Spjeldvik and Rothwell, 1985). A very limited attempt to account for absolute time variations in the radiation belts is found in the two versions of the AE-8 model corresponding to solar minimum and solar maximum respectively. These are typically considered separate static models, however, since there exists no means for accounting for the current state of the system.

Brautigam et al. (1992) took advantage of over a year's worth of electron measurements from the Combined Release and Radiation Effects Satellite (CRRES) to create a quasi-static electron flux model of the outer radiation belt. The CRRES High Energy Electron Fluxmeter (HEEF) measured electrons with energies between 0.8 and 8 MeV in ten differential energy channels. This instrument was a telescope by design, with an acceptance half-angle of 10.5°, and it was mounted perpendicular to the spin axis of the satellite, thus allowing it to resolve pitch angle distributions. Finally, the satellite traveled in a moderately inclined (18°) geostationary transfer orbit that carried it from a perigee of 350km out to an apogee of over 33,500km, providing near-equatorial coverage of the radiation belts from 0.1 R_E to beyond 6 R_E. To this day, CRRES constitutes one of the most comprehensive radiation belt experiments ever flown. The data was ordered according to eight levels of geomagnetic activity as measured by the Ap_{15} index, a 15 day running average of the ground-based magnetometer

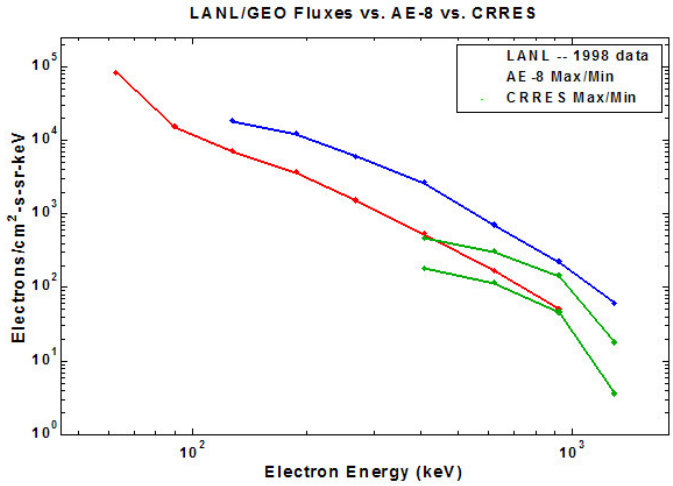


Figure 1.8: Comparisons of CRRESELE, AE-8 output, and LANL geostationary electron flux observations are shown (from Moorer and Baker, 2001).

derived Ap index provided by the NOAA-USAF Space Environment Services Center (now simply Space Environment Center, or SEC). The Olson-Pfitzer tilt-dependent magnetic field model (Olson and Pfitzer, 1977) was used to determine the L parameter for each measurement, as well as map observations to other geomagnetic latitudes.

This first incarnation of what would become the CRRES electron module (CRRESELE) from the Air Force Research Laboratory (AFRL) Geospace package (Brautigam and Bell, 1995) provided an immediate and substantial improvement in electron flux specification ability over the AE-8 model (both SOLARMIN and SOLARMAX versions) by addressing the inability of the AE-8 to adequately track time-variations in radiation belt electron fluxes. It also showed that the AE-8 model tended to regularly overestimate electron fluxes, even during the most geomagnetically active periods, confirming a variety of anecdotal accounts regarding the same phenomenon (see Figure 1.8). This was (and remains) especially true at higher energies, and a brief examination of AE-8 internals by Brautigam et al. showed that, indeed, the AE-8 assumed a much harder energy spectrum for radiation belt electrons than was measured by the CRRES

satellite at any point during its lifetime. The continued popularity of the NASA radiation models is therefore somewhat of a mystery. The most likely explanation is simply that, as an engineering tool, the AE-8 model provides a substantial margin of error for satellite mission and system designs.

Linear Prediction Filters

Linear prediction filters have been used to help understand dynamic coupling between the solar wind and the geospace environment for well over 20 years. Arnoldy (1971) first noted the strong linear relationship between the $\mathbf{B}_z$ component of the IMF and the AE index, a measure of auroral electrojet activity. Iyemori et al. (1979) were among the first to calculate discrete linear impulse responses of several geomagnetic indices to $\mathbf{B}_z$. A number of similar studies followed in the early-to-middle 1980s using various different inputs and outputs (e.g., Clauer et al., 1981, 1983; McPherron et al., 1984; Bargatze et al., 1985).

The linear prediction filters used for these studies were typically comprised of simple vectors of linear coefficients that were convolved with an input time series to provide a zero-lag prediction (a "nowcast") of the linear response in a given output time series. This type of model can be thought of as a finite impulse response (FIR) function, since the response becomes zero as soon as there is no longer input to be filtered. This is represented numerically in Equation 1.1, where b_τ comprises the filter coefficients, $u(t)$ is the input time series, and $y(t)$ is the output of the system. The index τ represents the discrete time-lag relative to the desired prediction time-step, with p and n signifying the most positive and most negative lags (note that it is not necessary for n to be negative or p to be positive, just that $n > p$).

$$y(t) = \sum_{\tau=p}^{n} b_\tau u(t-\tau) \tag{1.1}$$

Representative filter coefficients describing the discrete impulse response of relativistic

radiation belt electrons at geostationary altitudes to changes in the solar wind speed are shown in Figure 1.9. A detailed description of the techniques used to determine these optimal filter coefficients is given in Chapter 3.

Due to increased operational spacecraft use, a strong interest in practical radiation belt electron predictions developed in the middle-to-late 1980s. Not surprisingly, FIR prediction filters were one of the first techniques employed. Electron flux predictions for geostationary altitudes were made using both causal (i.e., $p \geq 0$;Nagai, 1988) and acausal (Baker et al., 1990b) filters designed to operate on several different types of input data. Typical prediction efficiencies from these models were around 50%, meaning that roughly 50% of the variance observed in the output was accounted for by the model.

Encouraged by this success, the technique was implemented in an operational setting at the Space Environment Center (SEC) at NOAA. The Relativistic Electron Forecast Model (REFM) uses an FIR linear prediction filter estimated using real-time data from the Advanced Composition Explorer (ACE) and particle detectors on their own geostationary GOES satellites to provide one to eight day forecasts of relativistic electrons at geostationary altitudes. Figure 1.10 is an example of the daily output generated by the REFM.

A more recent study by Vassiliadis et al. (2002) used observations of 2-6 MeV electrons from the Solar, Anomalous, and Magnetospheric Particle EXplorer (SAMPEX) to study the spatially distributed (in L-shell) radiation belt response to changes in solar wind speed, inspiring much of the work that led to the results presented in Chapter 4. Significant time-variability was noted at solar cycle, semi-annual, and seasonal time scales, and was discussed in terms of the well-known equinoctial and axial hypothesis for solar wind variability (see, for example, Boller and Stolov, 1970; Russell and McPherron, 1973; Cliver et al., 2000), demonstrating clearly the scientific value of linear filters, in addition to their proven operational value. Prediction efficiencies

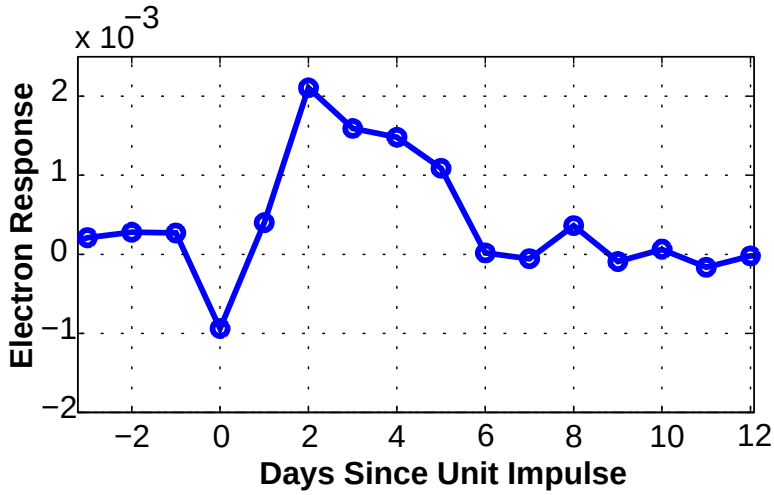


Figure 1.9: FIR filter coefficients describing the time-invariant response of the log(flux) of 2-6 MeV electrons at $L = 6.5$ to a discrete unit impulse in the solar wind bulk speed.

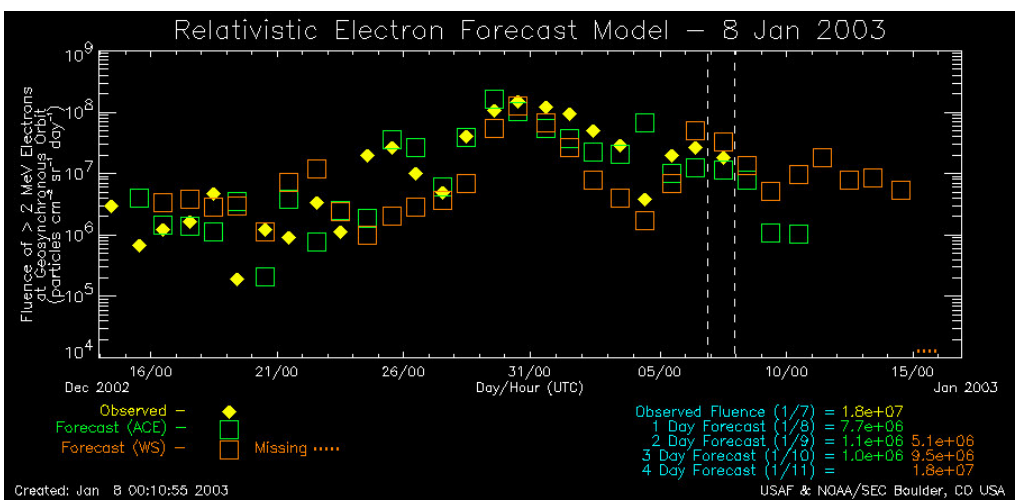


Figure 1.10: NOAA's Relativistic Electron Forecast Model. 1-3 day forecasts are based on FIR filters optimized for 1-3 day predictions. Forecasts beyond 3 days are from an FIR filter driven by artificial solar wind input generated by the Wang-Sheeley (WS) model SEC.

were typically lower than those reported in earlier papers, but this is most likely a consequence of using less processed electron data from the SAMPEX satellite. The prediction efficiencies were mostly consistent with those presented in Chapter 4.

Semi-Empirical Models

One important radial diffusion-based model that was not mentioned in the physics-based model section is the radiation belt prediction model of Li et al. (2001b). It is based on the same standard radial diffusion theory as those models mentioned previously, but uses highly parameterized non-linear functions of solar wind inputs to determine the time-variability of its diffusion coefficient. It assumes a PSD profile whose outer boundary is fixed at a higher level than the inner boundary, providing an arbitrary source that maintains an outward PSD gradient given than no internal acceleration mechanisms are provided. It also allows the time-constant τ to be adjusted, and finally, adiabatic effects are partially accounted for using an ad hoc function of a modeled D_{st} index, itself a parameterized function of solar wind inputs.

A potential drawback to this modeling approach lies in its physical and computational complexity. This is especially true from an operational perspective, where the advanced physical understanding that was required to develop the various components of the model are unlikely to be shared by typical forecasters, who are often called upon to judge whether or not the model output is appropriate for the current operating conditions. This is in contrast to the relatively simple and standardized engineering theory used to develop linear prediction filter prediction models, which offer an easily-understood baseline upon which to make more qualitative assessments of the current space weather conditions.

As presented in Li et al. (2001b), the parameters were fixed to best match observations made in 1995 and 1996, giving in-sample prediction efficiencies exceeding 80%. Prediction efficiencies for the entire period portrayed in Figure 1.11 hover

around 60%. This significant improvement in predictive ability over other radial diffusion codes, and even static FIR linear filters, is achieved because a large part of the dynamics of non-considered physics (like pitch-angle diffusion) actually gets wrapped into the adjustable parameters. The difference between this and the empirical functions used to describe sources, loss mechanisms, and time-dependent diffusion coefficients mentioned previously, is mostly a matter of the degree to which physical constraints have been relaxed, resulting in what are often considered unrealistic values for the parameters in the radial diffusion equations.

To conclude, high prediction efficiencies, at worst, make this an excellent engineering model. However, their consistency over time, and more importantly, the model's ability to predict well in different spatial regions, may provide strong empirical evidence that radial diffusion does indeed play a very significant role in the acceleration of radiation belt electrons.

1.3.3 Data Assimilation

The models mentioned thus far exhibit a range of predictive abilities, but none reproduce observations perfectly, even during periods of relative geomagnetic calm. There are two general reasons for this. First, there is invariably uncertainty associated with any measurement made of a system's state. All too often this shortcoming is ignored because the measurement errors are assumed to be small and unpredictable (i.e. white noise), and can be alleviated by simple smoothing or some other such filtering process. Second, and even less likely to experience concerted efforts to account for potential degradation in prediction quality, is the possibility that the model itself is insufficient. More often, as described above, simple parameterizations and/or adjustable initial and boundary conditions are used to force a particular model to reproduce unintended dynamics. Data assimilation (DA), in the most general terms, is any technique that actually acknowledges the errors associated with both observations and the model

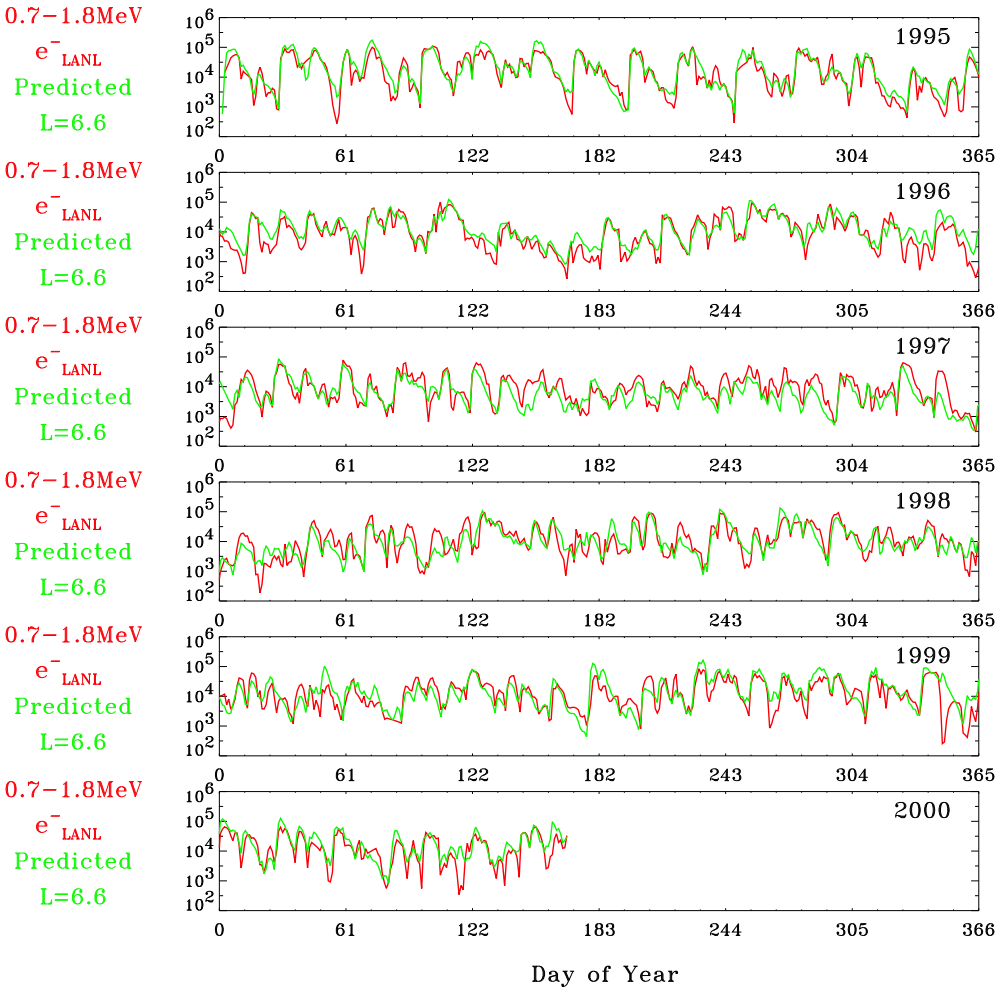


Figure 1.11: A comparison of over five years of daily-averaged MeV electron fluxes at geostationary altitudes (red lines and text) with results from the Li et al. (2001) parameterized radial diffusion electron flux prediction model (green lines and text; courtesy X. Li).

itself, and purposefully uses estimates of these uncertainties to make predictions that are closer to the true state of the system than either model or observations are capable of providing alone.

DA has been used for decades in terrestrial weather studies, and is becoming more popular in space weather models in recent years, as evidenced by a number of ad hoc attempts to use recent system observations to improve subsequent forecasts made by a number of the models mentioned above. For instance, NOAA's REFM actually keeps track of its own residuals, and uses them to determine a correction for the next linear prediction. This adds a tremendous amount of dynamical flexibility to the traditional FIR model structure used in the REFM. Appendix C discusses a number of variants of the linear prediction filter. Although the REFM documentation does not explain exactly how their offset correction is calculated, it is clear that the REFM would better be described as either a moving average filter with exogenous input (MAX), or an autoregressive filter with exogenous input (ARX) model than a simple FIR filter.

For more sophisticated, non-filter based models, such well-identified standard model structures do not exist. DA can still add significant flexibility to otherwise inflexible models by acknowledging and taking advantage of the expected errors in model output. Moorer and Baker (2001) used the CRRESELE model as a baseline which could be subsequently corrected to better match recent observations made by operational spacecraft like NOAA's GOES, and Los Alamos National Laboratory's geostationary and GPS satellites. In essence, the CRRESELE model actually served a nonlinear empirical mapping function that allowed current measurements to be extended throughout the region of interest, namely the Earth's electron radiation belts. More recent work by S. Bourdarie and R. Friedel is under review for publication, but promises similar results using the Salammbo code as a physical baseline model, thus adding dynamical flexibility and extending the operational relevance of the model beyond the region traversed by the CRRES satellite more than a decade in the past.

It is important to note that for these cases, and indeed others that are based on more formalized DA techniques, the underlying baseline models are almost always assumed to be perfect, and any model residuals employed to make necessary corrections are attributed to uncertainties in the observations alone. This means that DA has no inherent ability to improve estimates of future space weather conditions beyond the fact that it provides a better set of initial conditions for a dynamical model capable of propagating the system's state forward in time. If one assumes that some uncertainty can be attributed to the model itself, however, similar techniques to those just mentioned can be used to adjust the model parameters in order to help minimize the residuals. In a robust estimation algorithm like the well-known Kalman Filter, there is little inherent difference between adjusting the model output directly, or adjusting the model parameters to affect a similar change. We refer to the latter as adaptive system identification, a field with a proven theoretical and operational track record.

1.4 Non-Linear Dynamics, Local-Linear Filters, and Adaptive System ID

The primary goal of the research described in this thesis is to develop and validate an electron flux model that is of greater general utility, in both practical and scientific terms, than those already discussed. Purely statistical models simply lack the dynamic flexibility to be of tremendous use in a space weather forecasting sense. Physical models arguably provide the best short- and long-term forecasting capability because they are, ideally, a true representation of real-world phenomena. However, limitations imposed by finite precision computer arithmetic, and simply the computational intractability of full and complete physics models, almost always require approximations to be used or assumptions to be made. These usually degrade the prediction quality of the model. Semi-empirical (or semi-physical, if you prefer) models can usually improve and extend predictions by parameterizing known physics to account partially for these approximations and assumptions. However, they can be scientifically and/or philo-

sophically biased in such a way that the dynamical contribution of a particular physical mechanism may be overstated.

Linear prediction filters suffer from similar drawbacks to semi-empirical prediction models, and are additionally burdened by the reality that few natural systems exhibit time-stationary linear behavior except in very limited or controlled situations. This was made abundantly clear for the case of solar wind to magnetospheric coupling by Bargatze et al. (1985). In that work the FIR filter coefficients describing 30 partially overlapping sets of increasing geomagnetic activity data (AL index) were compared and demonstrated to exhibit very distinctive variations in their dynamical characteristics. In contrast to their lack of dynamical flexibility, however, linear prediction filters have the advantage that they are computationally simple, and that their mathematical properties are well understood. This allows the relatively easy manipulation of their parameters in order to better match current input/output environmental conditions, or the "state" of the system. Equation 1.2 is a generalized form of the discrete FIR convolution equation. This locally, or piece-wise, linear approximate model (hereafter referred to as a "local-linear" filter) is capable of effectively reproducing non-linear and/or non-stationary dynamics by making its coefficients state and/or time dependent.

$$\hat{y}(t) = \sum_{\tau=p}^{n} b_\tau(u,y,t)\, u(t-\tau) \tag{1.2}$$

When a deterministic system is being modeled (i.e. there is a one-to-one correspondence between the inputs and outputs), the time-dependence of the coefficients b_τ can be tied down firmly to the state, and should be dropped. This state-dependent local-linear filter provides a time-stationary, non-linear dynamical model of the system in question. The functional form of this state dependence is quite flexible, and variations of this type of model based on sorted look-up tables have been used in numerous studies of space- and geophysical systems (e.g. Codrescu et al., 1990; Vassiliadis et al.,

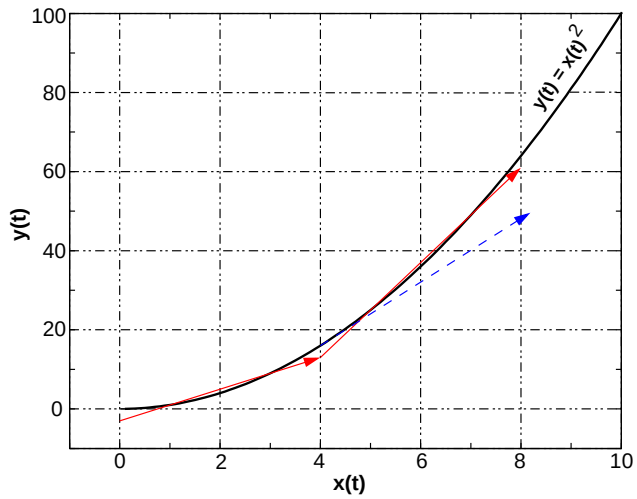


Figure 1.12: Local linearizations of a simple non-linear function $y(t) = x(t)^2$ (where x defines the "state", and can be input, output, or a combination of both). A two-piece state-dependent linearization is shown in red; a particular non-stationary (i.e. time-dependent) linearization in blue. Note the tendency of the non-stationary model to diverge from the true model as the state evolves, even though it is clearly the better approximation of the true dynamics when $x(t) = 4$.

1995, 1999).

If the coefficients vary as arbitrary functions of time (i.e., there is no known dependence on the state variables) the local-linear filter can still be used to track non-linear dynamics that manifest themselves as non-time-stationary time series. Klimas et al. (1997) determined time-varying local-linear filter coefficients using a fixed-width moving window of input/output data centered about the desired prediction step. The end objective of the Klimas et al. study was to functionalize the time-variability of the coefficients and derive an analytic non-linear model from these results, so they were not very concerned with the lack of causality in their system identification scheme. It is more common in engineering applications, however, to use time-varying linear models with efficient recursive identification algorithms which process only the latest system measurement. This requires no knowledge of future measurements to update the filter coefficients, but the state of the system cannot change too quickly relative to

the observation sampling time, or else the local-linear approximation will diverge from the true non-linear dynamics (see Figure 1.12). This approach is often referred to as adaptive linear system identification, and is the primary topic of this thesis.

1.5 Thesis Organization

In order to remain as accessible to as broad an audience as possible, every attempt was made to organize this thesis so that each chapter can stand alone if necessary, but will undoubtedly contribute to the document as a whole. Chapter 1 has already provided much of the historical and theoretical background necessary to motivate the research presented in the remainder of the thesis.

Chapter 2 details the OMNIWeb solar wind database and the SAMPEX 2-6 MeV radiation belt electron flux data, the two primary data sets used to develop and validate the models presented and utilized in subsequent chapters. This includes a detailed description of the satellites and instruments used to measure these complex geospace environments, as well as the pre- and post-processing techniques and algorithms used to massage the data into more appropriate formats. Finally, the chapter concludes with a discussion of the limitations inherent to these two data sets, and what kind of data might be better suited for real-time operational forecast models.

A more detailed explanation of general recursive/adaptive linear system identification techniques is presented in Chapter 3. This includes a review of the traditional least-squares approach to linear filter parameter estimation, in both off-line, or “batch” configurations, and recursive online configurations based on the recursive least-squares (RLS) algorithm. An adaptive form of the RLS, capable of tracking time-varying linear parameters by implementing an exponential “forgetting factor”, is described. Finally, a scheme for using the well-known Kalman Filter algorithm in a system identification configuration is described in detail and compared to the RLS in terms of its flexibility and numerical robustness.

Despite its somewhat ad hoc approach to “forgetting”, the Kalman Filter is shown to be more flexible and numerically robust than the traditional RLS. In addition, it lends itself better to state-space based model structures for use in real-time forecasting situations. Chapter 4 therefore presents real results from adaptive FIR linear models derived with the Kalman Filter using the data sets described in Chapter 2. This includes visual comparisons between data and model output, as well as rigorous statistical validation based on detailed residual analyses. Single and multi-input models are considered and the results are interpreted in terms of both known physics and heretofore unidentified dynamical responses.

Chapter 5 addresses some known short-comings of, and possible improvements to, the data and model structures presented in previous chapters, with particular attention being paid to the question of how one might build a useful adaptive model for real-time operational space-weather purposes. This includes: a) enhancing the simple FIR linear filter to include dynamical feedback and/or separable noise models; b) extending the output state-space to cover a 3-D spatially distributed region with potentially complicated multi-output dynamical feedback; and c) the transition to a true state-space representation of the linear dynamical equations, allowing much easier implementation of both points a) and b).

Finally, Chapter 6 concludes by summarizing the results and conclusions from the rest of the thesis, confirming that the motivating issues presented in Chapter 1 were properly addressed, and bringing closure to one young researcher’s first steps into the complex but exciting world of multi-disciplinary space science research.

Chapter 2

Solar Wind and Radiation Belt Data

This chapter provides a brief description of the data sets used for development and validation of the adaptive linear system identification and prediction algorithms presented in later chapters. These data were chosen for their easy access, their completeness, and their consistency. They are not necessarily optimal for near real-time specification and forecasting, mostly because they are not available in anything near real time. Section 2.1 describes the OMNI database of solar wind parameters maintained by the National Space Science Data Center (NSSDC), at NASA's Goddard Space Flight Center (GSFC). Section 2.2 describes the electron flux data taken from the Solar Anomalous and Magnetospheric Particle Explorer (SAMPEX) satellite. Finally, Section 2.3 discusses the techniques used to synchronize the two data sets and deal with data gaps during both the model training and prediction stages of development and testing.

2.1 OMNI Solar Wind Parameters

The OMNI database is a compilation of hourly resolution interplanetary magnetic field (IMF) data, solar wind plasma data, and energetic particle fluxes taken from several different satellites over the past few decades. All similar data is cross-normalized to provide multi-source uniformity (National Space Science Data Center, 2003). No correction is made for the solar wind propagation time from the satellite to the Earth's

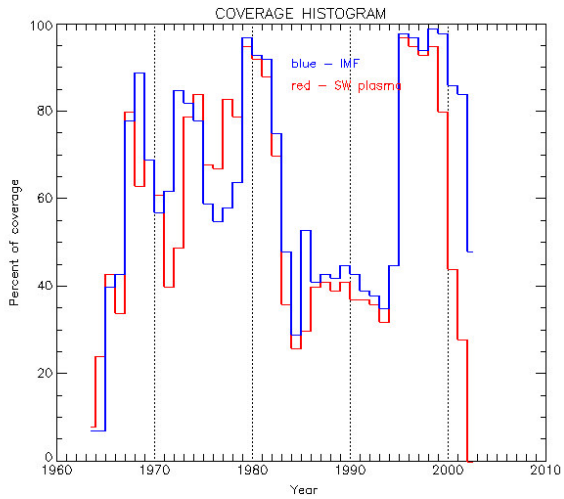


Figure 2.1: Data-density for the OMNI solar wind database. Data used in the research presented in this thesis was taken from 1994 through 1999.

position, since this is normally much less than one hour[1] . Figure 2.1 describes the density of the OMNI solar wind data as a function of time. The data used for this study was taken from the years 1994 through the end of 1999. Despite its relatively low data density, 1994 was included in order to evaluate the effects of poor data on the estimation and prediction phases of our study.

It was decided early on that a limited number of the OMNI parameters would be used in this study. These included measurements of the solar wind bulk speed V_{sw}, the magnitude of the IMF $|\mathbf{B}|$, and the solar wind ion density ρ_{ion}. Vector quantities associated with V_{sw} and $|\mathbf{B}|$ were available from the OMNI database, but were deliberately not included in the spirit of keeping the baseline models simple in order to explore various operational characteristics of adaptive linear system identification algorithms, especially the Kalman Filter (refer to Chapter 3).

[1] At the time these data were originally collected, the OMNI database did not include measurements from either the Advanced Composition Explorer (ACE), or the Wind satellite, two of the newest and most technologically advanced solar wind observation platforms. This is because propagation times for observations made by these two satellites can be comparable to the one hour resolution of the OMNI database, and a robust correction factor had not yet been developed that would allow these data to be consistent with the existing OMNI data. This has since been remedied with the OMNI2 dataset.

The six years of data used for this study are shown in the top three panels of Figure 2.2 in Section 2.3. Higher resolution versions of these plots can be found in Appendix A.

2.2 SAMPEX 2-6MeV Electrons

The Solar, Anomalous, and Magnetospheric Particle EXplorer (SAMPEX) satellite was launched in 1992 into a low-altitude (~600 km), high-inclination (~82°) orbit, with a period of approximately 100 minutes (Baker et al., 1993). As a result of its low altitude, it is possible to use the International Geomagnetic Reference Field (IGRF), an internal terrestrial magnetic field model, to bin SAMPEX particle measurements according to a nominal geomagnetic L-shell. The ELO channel of the Proton Electron Telescope (PET; Cook et al., 1993) was used for the studies described in subsequent chapters, providing flux measurements of 2-6 MeV electrons (electrons/cm$^2 \cdot$ sec $\cdot$ steradian), a particle population with important space weather implications (refer to Section 1.2).

The six years of data used for this study are shown in the bottom panel of Figure 2.2 in Section 2.3. A higher resolution version of this plot can be found in Appendix A.

2.3 Data Pre/Post-Processing

A significant amount of preprocessing of both data sets was necessary before they could be incorporated into the adaptive system identification and prediction algorithms discussed in later chapters. First, the SAMPEX electron flux measurements were provided by the SAMPEX data team as orbit-averages with a cadence of approximately 100 minutes. These orbit-averaged fluxes typically exhibit a strong diurnal variation that has been attributed to the rotation of the Earth's imperfect dipole with respect to the relatively fixed orbit plane of the SAMPEX satellite. As a result, the South Atlantic

Anomaly (SAA) causes a spike in fluxes across a broad range of L-shells once per day. In order to avoid undesirable statistical effects resulting from the auto-correlations associated with these diurnal spikes, it was decided that daily averages of all measurements would be used. In order to synchronize these two data sets at a daily cadence, a 24 hour moving average was calculated for each. The resulting smoothed time series was then resampled at the same instants in time for both the OMNI and SAMPEX data sets. This was perhaps unnecessary since daily-averaged versions of both data sets have actually been made publicly available. However, neither of these were well documented with regard to the method used to generate the daily averages, nor were any kind of confidence estimates provided, making it difficult to set a rejection criterion for bad data points.

The SAMPEX data was fairly complete, and the small data gaps that did exist could have easily been filled using a simple linear interpolation scheme. Unfortunately, the data gaps in the OMNI solar wind data tended to span much longer periods, such that linear, and in fact any geometrically based, interpolation scheme would not only result in missed dynamics, but would likely bias the estimation of the model parameters as the gaps constituted a statistically significant portion of the training data. An empirical study using time-stationary filters as test-cases indicated that replacing missing data with the average of the whole data set was statistically benign. This can be seen in the prediction residual time series presented in Chapter 4 by noting that the residuals during the data gaps were effectively zero for all the time-stationary models.

As it turns out, it may have been easier to deal with gaps in the SAMPEX data. This is due to the fundamental nature of the FIR linear filter, which requires a series of time-lagged inputs to predict a single output. The existence of even a single missing input data point will affect the prediction of a number of output data points equal to the length of the discrete filter. In the end, it was decided to replace not only the missing data points, but each neighboring point corresponding to the FIR time-lags. This had

to be done for both the OMNI input data as well as the SAMPEX electron flux outputs.

The data shown in Figure 2.2 have undergone the treatment just described. The resulting time series are missing data from most of 1994, not because none existed, but simply because it was impossible to go more than fifteen days (the length of the filters used in our studies) without encountering a missing data point. This was unfortunate, since it excluded a significant portion of the declining phase of the last solar cycle. In addition, there are subsequently occasional shorter, but still significant, periods where the historical averages were substituted for missing data.

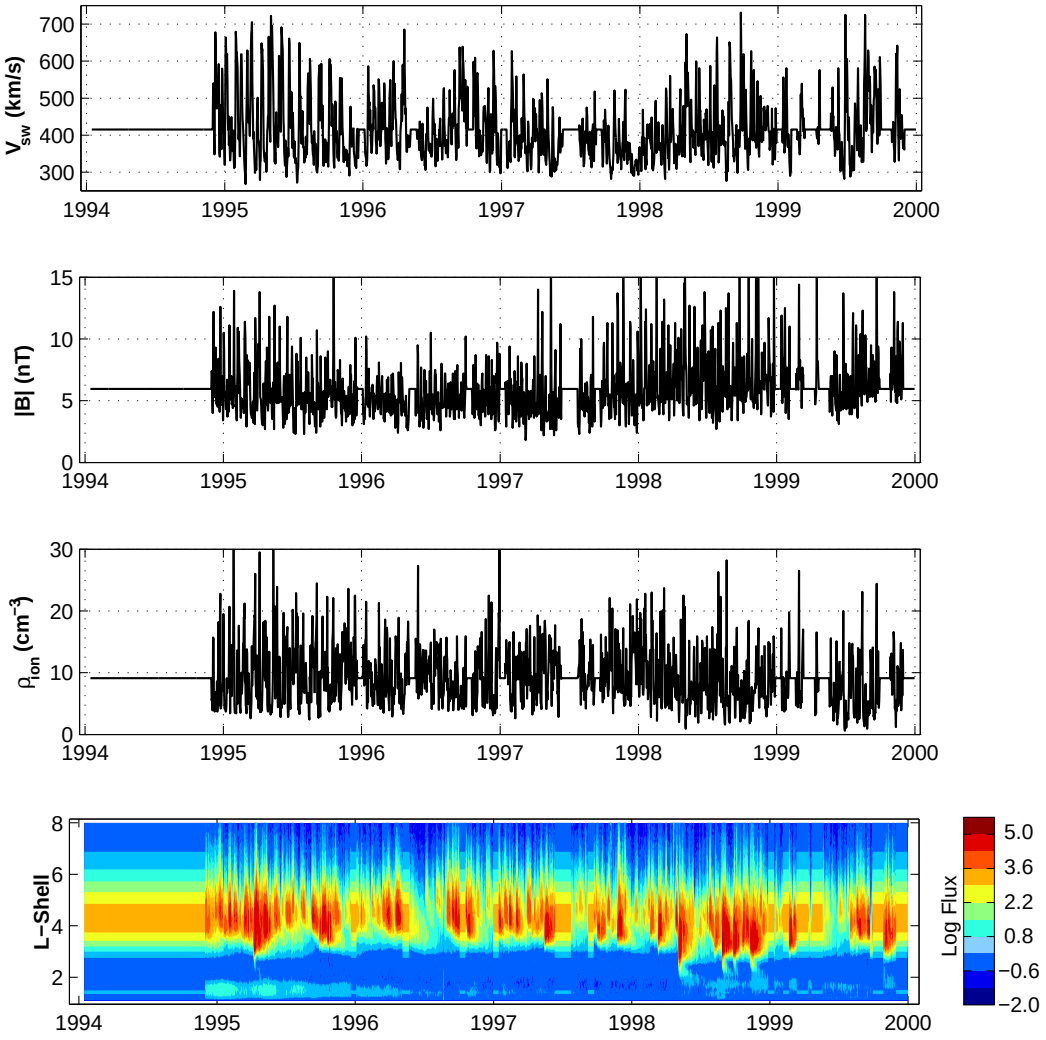


Figure 2.2: Six years of solar wind inputs plus L-shell dependent profile of SAMPEX 2-6 MeV electron log-fluxes. Input data gaps are easily recognizable as flat lines at values corresponding to the particular parameters time average, or as horizontal contours in the flux map, throughout most of 1994, and occasional subsequent shorter periods.

Chapter 3

Linear System Identification

System identification is the process of deriving a mathematical model from observed data in accordance with some predetermined criterion (Johansson, 1993), often the minimization of an error function describing a prediction model's residuals. Such models should be distinguished from physical theory, although in mature form they may be incorporated into the deductive process as a form of data-reduction tool. As such, the model's basic structure must be thoroughly understood in terms of its strengths and limitations before definitive statements about the system of interest are made based on model output. It is for this reason that linear filters were chosen to represent dynamics for the work presented in this thesis. While it is generally accepted that most real-world systems are inherently non-linear to some degree, linear approximations are often capable of recreating a significant fraction of the observed dynamic output of a given system. In addition, linear filters are mathematically simple, making them easy to implement and, perhaps more importantly, to modify in code.

This chapter describes the basic structure of the finite impulse response (FIR) linear filter as well as techniques for determining coefficients that provide optimal predictions based on a given input time series. Time-stationary linear filters (i.e., filters with fixed coefficients) are examined first, including discussion of both batch and infinite-memory recursive identification algorithms. This discussion is designed to establish standard notation and terminology for subsequent use. Next, assumptions about time-

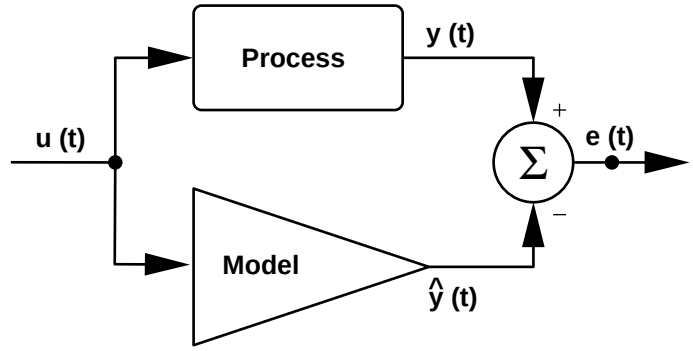


Figure 3.1: A generalized single input, single output system identification configuration for a time-stationary model. Deterministic model output and observations are compared, and the residuals $e(t)$ are minimized for an optimal prediction model.

stationarity are relaxed as finite-memory adaptive identification algorithms capable of tracking a non-linear system's smoothly-varying linear parameters through time, are presented. Relatively simple problems are solved and presented for illustrative purposes, with particular attention paid to the Kalman Filter identification scheme. A validation study, using the real-world observations described in Chapter 2, is presented in Chapter 4.

3.1 Time-Stationary Linear Systems

The FIR linear filter is the most fundamental component of all the models developed and analyzed in this thesis. Briefly described in Chapter 1, it is comprised of a simple vector of coefficients that can be convolved with a discrete input time series in order to reproduce discrete output. Despite the fact that many previous studies allowed acausal positive lags (i.e., $\tau < 0$), we will proceed to assume from this point forward that $p = 0$. This establishes a causal dynamic relationship between the input time series ($u(t)$) and the output time series ($y(t)$) which can be described by the nth-order filter $B = b_\tau$.

$$\hat{y}(t) = \sum_{\tau=0}^{n} b_\tau u(t-\tau) \tag{3.1}$$

Equation 3.1 provides deterministic output. In other words, there is a one-to-one correspondence between system input and model output, and any prediction error must constitute a white-noise time series (since, by definition, white noise perfectly random and impossible to predict).

There is no explicit feedback of system output included in Equation 3.1. This requires that the number of coefficients be sufficient to capture the system's entire impulse response, posing a potential problem in numerical complexity when dealing with systems sampled frequently relative to their dynamical time-scales. On the other hand, the lack of feedback has several advantages: 1) the filter is "linear in its parameters". In other words, the FIR filter in no way depends on previous model output to make its current prediction, thus allowing its optimal parameters to be determined using linear least-squares minimization; 2) the estimation of model parameters for multi-input systems is simplified by eliminating any cross-coupling feedback of the internal dynamics (see Chapter 4); and 3) since the model output is identically zero when the input is zero, it is immune to stability problems encountered in models with output feedback (alternatives to the FIR model incorporating dynamic feedback are discussed in Chapter 5).

3.1.1 Linear Least-Squares Parameter Estimation

Since the FIR filter is linear in its parameters, the convolution just described for a single input, single output system can be fully represented by the single matrix multiplication shown in Equation 3.2, provided the filter coefficients are time-stationary:

$$\hat{y}(t) = \Phi\theta \quad \text{...where} \tag{3.2}$$

$$\hat{y}(t) = \begin{bmatrix} \hat{y}_0 \\ \hat{y}_1 \\ \vdots \\ \hat{y}_t \end{bmatrix}, y(t) = \begin{bmatrix} y_0 \\ y_1 \\ \vdots \\ y_t \end{bmatrix}, \theta = \begin{bmatrix} b_0 \\ b_1 \\ \vdots \\ b_n \end{bmatrix}, \text{ and } \Phi = \begin{bmatrix} u_0 & u_{-1} & \cdots & u_{-n} \\ u_1 & u_0 & \cdots & u_{1-n} \\ \vdots & \vdots & \ddots & \vdots \\ u_t & u_{t-1} & \cdots & u_{t-n} \end{bmatrix}$$

Before continuing, it is prudent to clarify some of this notation for later reference. First, as stated previously $y(t)$ represents a time series of output observations. In other words, y is a function of time. Likewise, $\hat{y}(t)$ is a time series of predicted output, also a function of time. Subsequently, the variable t should be considered a time index when it is used as a subscript, so y_t corresponds to the tth row of whatever vector or matrix it is found.

In contrast, we have actually dropped the variable τ from any of our matrix equations since it is assumed from this point forward that the maximum time-lag is equal to zero, and that the minimum time-lag corresponds to the order of the system, n. Therefore, the variable n should be considered a positive time-lag if it is typed as a subscript, and u_{-n} corresponds to the nth column of the input regression matrix Φ. There is no compelling reason to assign the nth-order vector of FIR filter coefficients $[b_0\, b_1 \cdots b_n]^{\mathsf{T}}$ to the variable θ at this point, except that additional coefficients can be included in this generic parameter vector later on when the identification of more sophisticated model structures is addressed.

Moving on; if Φ is a non-singular square matrix, it can be easily inverted and multiplied by the measured output $y(t)$ in order to determine the set of filter coefficients required to make $\hat{y}(t) = y(t)$ exactly. If there are more observations than the filter has time-lags, the system is said to be over-determined, and it becomes necessary to define a loss function based on prediction residuals ($e(t) \equiv [e_1\, e_2 \cdots e_{i=t}]$) that can be minimized in order to determine the optimal set of filter coefficients:

$$J = \frac{1}{2}\sum_{i=1}^{t} e_i^2 \tag{3.3}$$

This loss function can be minimized by calculating the pseudo-inverse of the non-square Φ and multiplying it by the measured outputs:

$$\hat{\theta} = \left(\Phi^{\mathsf{T}}\Phi\right)^{-1}\Phi^{\mathsf{T}} y \tag{3.4}$$

This is known as "least-squares" parameter estimation. It is assumed that $e(t)$ is white noise, or uncorrelated with itself. If this is not true, it is possible that the estimated

parameters are sub-optimal, even though the loss function J was minimized. One last point; it is necessary that the matrix $\left(\Phi^{\mathrm{T}}\Phi\right)$ be non-singular, which will generally be the case if the input $u(t)$ is persistently exciting (i.e. it is non-zero and exhibits a dynamic variability that can be related to the output).

Weighted Least Squares

It is often the case that some observations used in the least-squares estimation procedure are more reliable than others, and should be weighted accordingly. This requires a modified form of the loss function in which each residual is given an associated weight, w. This can be represented in matrix form:

$$J_w = \frac{1}{2} e(t)^{\mathrm{T}} \mathbf{W} e(t) \tag{3.5}$$

where:

$$\mathbf{W} = \begin{bmatrix} w_{11} & 0 & \cdots & 0 \\ 0 & w_{22} & \cdots & \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & w_{tt} \end{bmatrix} \tag{3.6}$$

The pseudo-inverse and solution gives the so-called weighted least-squares (WLS) optimal parameter estimate, and takes the following form:

$$\hat{\theta} = \left(\Phi^{\mathsf{T}} \mathbf{W} \Phi\right)^{-1} \Phi^{\mathsf{T}} \mathbf{W} y \tag{3.7}$$

Clearly, if w_{11}=w_{22}=$\cdots$=w_{tt}, the standard least-squares solution is recovered since $\mathbf{W}^{-1}\mathbf{W} = \mathbf{I}$. Sometimes these weights are simply a measure of one's confidence in the training data. Often, however, they are made functions of either the input or output state, allowing the user to intentionally bias the model toward more or less extreme events. The weighted estimation of a time-stationary prediction filter designed to emphasize extreme electron enhancements, for example, might serve as a valuable space weather tool.

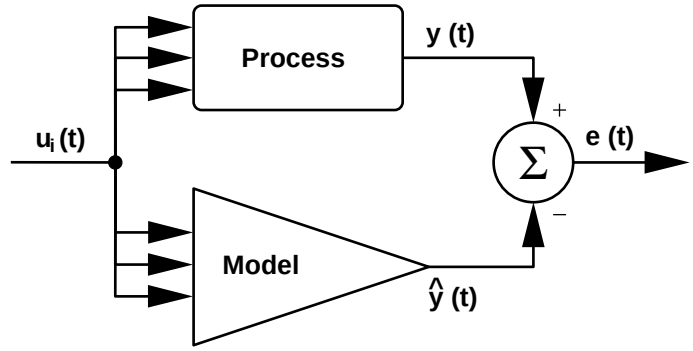


Figure 3.2: A generalized multi-input, single output system identification configuration for a time-stationary model. Deterministic model output and observations are compared, and the residuals $e(t)$ are minimized for an optimal prediction model.

Multiple Inputs

Many of the more interesting results presented in Chapter 4 involve the response of radiation belt electrons to multiple simultaneous solar wind inputs. Figure 3.2 illustrates conceptually a multi-input linear model, while Equation 3.8 shows what a three-input linear predictor would look like.

$$\hat{y}(t) = \sum_{\tau=0}^{n} b_{1,\tau} u_1(t-\tau) + \sum_{\tau=0}^{n} b_{2,\tau} u_2(t-\tau) + \sum_{\tau=0}^{n} b_{3,\tau} u_3(t-\tau) \tag{3.8}$$

Extending the standard least-squares identification to include multiple inputs is quite simple in theory, if often complicated in practice as one must keep careful track of matrix indices. The multi-input FIR identification method is essentially identical to that described in Equation 3.2, with minor changes to the parameter vector and regression matrix:

$$\theta = \begin{bmatrix} b_{0,1} \\ b_{0,2} \\ b_{0,3} \\ \vdots \\ b_{n,i} \end{bmatrix}, \text{ and } \Phi = \begin{bmatrix} u_{0,1} & u_{0,2} & u_{0,3} & \cdots & u_{-n,i} \\ u_{1,1} & u_{1,2} & u_{1,3} & \cdots & u_{1-n,i} \\ u_{2,1} & u_{2,2} & u_{3,3} & \cdots & u_{2-n,i} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ u_{t,1} & u_{t,2} & u_{t,3} & \cdots & u_{t-n,i} \end{bmatrix} \tag{3.9}$$

3.1.2 Recursive Least Squares (RLS)

The previous description of the least-squares parameter estimation procedure is often referred to as a "batch" process (e.g., Nelles, 2001). It is assumed that all of the data are immediately available to form the various matrices and calculate the optimal parameters in one pass, and is it is therefore only useful for offline applications (i.e. data mining) and identification of time-stationary linear filters.. Often this is not the case. For instance, in the not-too-distant past it was not uncommon for computer memory to limit the size of the regression matrices, and a recursive algorithm that operated on only a single observation at a time served to conserve expensive computer resources.

While the cost and reliability of computer memory is not as relevant as it once was, a recursive least-squares (RLS) algorithm is still quite valuable for "online" applications in which the system measurements, arriving in near-real-time, are processed one at a time to improve the model with each new observation. In this way, the parameter vector $\hat{\theta}$ becomes a function of time, and the RLS serves to recompute $\hat{\theta}(t)$ at each time instant t by adding a correction factor to the previous parameter estimate $\hat{\theta}(t-1)$ (Nelles, 2001). Equations 3.10 are the most common final form of the RLS, employing a well-known matrix lemma to remove an otherwise costly matrix inversion and reducing the computational complexity from order $O(n^3)$ to $O(n^2)$ for each time-step. The full derivation can be found in any one of numerous standard texts (e.g., Johansson, 1993; Ljung, 1999; Nelles, 2001).

$$\begin{aligned}
\hat{\theta}(t) &= \hat{\theta}(t-1) + \gamma(t)\left(y(t) - \phi(t)\hat{\theta}(t-1)\right) \\
\gamma(t) &= \left(\phi(t)\mathbf{P}(t-1)\phi^{\mathsf{T}} + 1/w(t)\right)^{-1}\mathbf{P}(t-1)\phi^{\mathsf{T}} \\
\mathbf{P}(t) &= (\mathbf{I} - \gamma(t)\phi(t))\mathbf{P}(t-1)
\end{aligned} \tag{3.10}$$

In words, the correction factor $\gamma(t)$ is multiplied by the residual of the current time step (assuming $\phi(t)$ is the corresponding row-vector from the regression matrix Φ) and added to the previous parameter vector to produce an updated set of FIR pa-

rameters. In the limit $t \to \infty$, $\mathbf{P}$ and $\gamma \to 0$ (assuming the input is persistently exciting), and the algorithm converges on the optimal weighted least-squares estimation of the FIR parameters. It is worth noting that the incorporation of a simple scalar weighting function at each time-step offers an additional advantage of a recursive identification procedure over the batch process described previously, which can take up a tremendous amount of memory if the full weighting matrix $\mathbf{W}$ is used.

It should be pointed out that individual observation weighting, whether in the form of a weighting matrix for batch estimation, or simple scalar weight w in the RLS, was not applied to any of the models developed and analyzed for this thesis. It is discussed here because of its relevance to, and distinction from, the concept of "forgetting", or the ability of an identification algorithm to gradually disregard observations made in the past, allowing the model to adapt to a changing state. This is explained in more detail in the following section.

3.2 Time-Varying Linear Systems

It has been noted that a non-linear dynamical system can be approximated with a non-stationary linear model if the parameters vary smoothly in state-space over time. This is the fundamental premise behind using adaptive filters in this thesis (see Chapter 1). A means for emphasizing a particular observation in the identification process has already been described, and may be thought of as a kind of adaptation to a particular set of operating conditions. However, once this weighting function has been applied, its influence remains for all time. A truly adaptive linear filter must generally be influenced less and less by older observations than those made more recently. In other words, a finite memory is given to the model.

Since this time-variability is not tied explicitly to the system state, such a model is not truly non-linear. It is not technically a deterministic model because there is no guarantee that the model parameters will return to the same values if the system returns

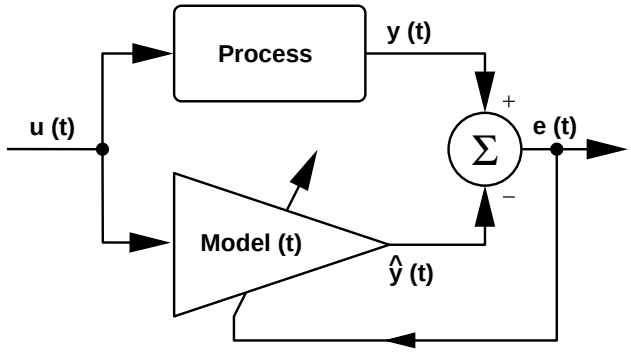


Figure 3.3: A generalized single input, single output system identification configuration for a time-varying, or adaptive model. A finite "memory" allows the oldest observations to be forgotten over time, resulting in a variable error loss function and model parameters that track non-stationary process through time. Note that residuals are not explicitly incorporated into the model.

to a previous state, and consequently there is no one-to-one correspondence of input to output. However, since the model will always give the same output if its parameters are the same, we will not refer to it as "non-deterministic". This term is reserved for models which can produce different output given identical states at different times, even with fixed parameters. These are discussed more in Chapter 5.

3.2.1 RLS with Exponential Forgetting

Finite memory in a recursive identification algorithm is achieved by preventing the correction factor $\gamma(t)$ from approaching zero. To do this, a forgetting factor $(0 < \lambda \leq 1)$ is introduced into the RLS that ensures that the covariance matrix $\mathbf{P}(t)$ remains significant. This requires a modified form of the loss function for recursive least-squares estimation.

$$J_\lambda = \sum_{i=1}^{t} \lambda^{t-i} \left(y_i - \phi_i \hat{\theta}(t)\right)^2 \tag{3.11}$$

This results in the following modified RLS algorithm which "remembers" training data exponentially as roughly λ^j for observations made j time-steps ago (Johansson, 1993).

$$\begin{aligned}
\hat{\theta}(t) &= \hat{\theta}(t-1) + \gamma(t)\left(y(t) - \phi(t)\hat{\theta}(t-1)\right) \\
\gamma(t) &= \left(\phi(t)\mathbf{P}(t-1)\phi^{\mathsf{T}} + \lambda/w(t)\right)^{-1}\mathbf{P}(t-1)\phi^{\mathsf{T}} \\
\mathbf{P}(t) &= 1/\lambda\left(\mathbf{I} - \gamma(t)\phi(t)\right)\mathbf{P}(t-1)
\end{aligned} \tag{3.12}$$

Typically λ is chosen as a compromise between noise sensitivity and tracking ability. If $\lambda = 1$ the algorithm recovers the original RLS without forgetting, and all data are weighted equally. This time-stationary model is effectively insensitive to noise, a desirable trait if assumptions about white residuals are warranted. On the other hand, if $\lambda = 0.9$, a typical value, new data are 3, 8, 24, and 68 times more significant than old data 10, 20, 30 and 40 samples prior. If $\lambda = 0.8$, new data are 9, 87, 808, and 7523 times more significant than old data at similar lags. The smaller the value for λ, the faster it tracks changing dynamics, but noisy outliers will be overly emphasized and may lead to instabilities in the algorithm if they are not otherwise accounted for.

One serious drawback to an approach based on exponential forgetting is that if the input is not sufficiently exciting (i.e., it is zero, or exhibits no significant dynamic variability) there may be problems related to the update of the covariance matrix $\mathbf{P}(t)$. For one, it will grow exponentially if $\phi(t) \approx 0$ due to the $1/\lambda$ term, destabilizing the RLS. This covariance matrix "blow-up" can be dampened somewhat by incorporating appropriate time-varying values for λ and/or w, however the exponential character of the growth still remains and may be exacerbated if these values are chosen poorly. Additionally, if $\phi(t)$ remains sufficiently large but unchanging, $\mathbf{P}(t)$ and $\gamma(t)$ will quickly approach zero, regardless of the forgetting factor, since the RLS takes this to mean there is little or no useful information remaining in the input signal. This can force the algorithm into a state that may not be adequately responsive when $\phi(t)$ once again exhibits dynamic variability. The well-known Kalman Filter is less susceptible to these issues, and possesses additional benefits that are discussed next.

3.2.2 Kalman Filter

The Kalman Filter is an approximate non-linear filter based on a first-order linearization of the state equation (Ljung, 1979), and is briefly described in general as follows. Assume the non-linear, discrete-time system

$$\begin{aligned} \xi(t+1) &= f(t,\xi(t)) + v(t) \\ y(t) &= h(t,\xi(t)) + e(t) \end{aligned} \tag{3.13}$$

where the function f represents the dynamics that map the system state ξ forward in time, the function h is a mapping function to transform the state into an observable in order to calculate prediction errors. There is an uncertainty associated with both operations that can be represented as random variables $v(t)$ and $e(t)$. The optimal estimate of the state at the next time step based on observations up to and including the current step is denoted by $\hat{\xi}(t+1)$, and given according to

$$\hat{\xi}(t+1) = f(t,\xi(t)) + K(t)\left(y(t) - h(t,\xi(t))\right)$$

where the calculation of $K(t)$, the Kalman Gain, is described below for the specific case of linear system identification (a more general description is provided by Ljung, 1979).

There is no inherent difference between state estimation and parameter estimation (Ljung and Gunnarsson, 1990). Therefore we can rewrite Equation 3.13 as

$$\begin{aligned} \theta(t+1) &= \theta(t) + v(t) \\ y(t) &= \phi(t)\theta(t) + e(t) \end{aligned} \tag{3.14}$$

Equation 3.14 defines the "state" of the system as the same set of FIR filter parameters $\theta(t)$ used in the previous identification schemes. The regression row-vector $\phi(t)$ serves to map the state to an actual observable. The random variable $v(t)$ represents the process noise of the system, or in this case, the error associated with the model parameters themselves. The variable $e(t)$ represents the actual observation errors if the model $\phi(t)\theta(t)$ predicted perfectly at time t. Both $v(t)$ and $e(t)$ are usually assumed

to be independent white noise sequences, but they are not actually known a priori in online applications. They are therefore represented as stochastic processes that can be defined by their respective noise covariance matrices: $\mathbf{Q}^v = E\left(vv^\mathsf{T}\right)$ and $\mathbf{Q}^e = E\left(ee^\mathsf{T}\right)$ ("E" is the statistical expectation). These are then incorporated into a recursive update algorithm as follows:

$$\begin{aligned}
\hat{\theta}(t+1) &= \hat{\theta}(t) + K(t)\left(y(t) - \phi(t)\hat{\theta}(t)\right) \\
K(t) &= \left(\phi(t)\mathbf{P}(t)\phi^\mathsf{T} + \mathbf{Q}^e\right)^{-1}\mathbf{P}(t)\phi^\mathsf{T} \\
\mathbf{P}(t+1) &= \left(\mathbf{I} - K(t)\phi(t)\right)\mathbf{P}(t) + \mathbf{Q}^v
\end{aligned} \tag{3.15}$$

Equations 3.15 are the Kalman Filter interpretation of recursive linear system identification. The best estimate of the model's current parameters is the sum of the parameters from the previous time-step and the prediction error multiplied by the Kalman gain, $K(t)$. If the model perfectly predicts the last observation, $\hat{\theta}$ will not change since $K(t) \cdot 0 = 0$. If, on the other hand, there was an error in the prediction $\phi(t)\theta(t)$, then the model is adjusted in a direction determined by the sign of that error, and its magnitude weighted according to $K(t)$.

The remainder of Equation 3.15 acts to update the elements of $K(t)$ so that there is a balance between the magnitude of the lagged inputs used to excite the model and the inverse of the uncertainty associated with the observations (i.e., if one isn't sure the observations represent reality, the model should not be adjusted very much to match these observations). The Kalman gain is also dependent on the actual confidence in the model. Since each new measurement processed by the algorithm should, in theory, improve the quality of the model, $\mathbf{P}(t)$, and subsequently $K(t)$, necessarily grow smaller and smaller, approaching zero as time progresses. A finite $\mathbf{Q}^v$ ensures that these never actually reach zero, and that the algorithm remains capable of changing the FIR parameters if necessary to better match current observations.

If it is assumed that $\mathbf{Q}^v = 0$ and that $\mathbf{Q}^e$ is equal to the inverse of the weighting

function in the RLS, the Kalman Filter and the WRLS (without forgetting) are equivalent; both simply return the weighted least-squares estimate for the parameters. When $\mathbf{Q}^v \neq 0$, the Kalman Filter is allowed to track time-variations in the parameters in a manner that it is more resilient to the previously mentioned covariance matrix blow-ups encountered when using the RLS with forgetting. This is because $\mathbf{P}(t)$ only grows at a linear rate when $\phi(t) = 0$. In addition, $\mathbf{P}(t)$ does not necessarily approach zero as $t \to \infty$ for a non-zero constant inputs. These two traits help the Kalman Filter remain stable, even when operating on rather extended sequences of bad input data.

One additional benefit afforded by the Kalman Filter is more flexibility in defining the effective memory for different model parameters. The RLS is stuck with the same exponential forgetting factor regardless of which parameter is being estimated. It is possible to control the forgetting individually for each estimated parameter in the Kalman Filter by adjusting the corresponding entry (usually on the diagonal) in the $\mathbf{Q}^v$ matrix. In general, the entries in $\mathbf{Q}^v$ should correspond to the strength of the time-variance of the individual parameters. In other words, if the parameter is expected to vary quickly, its associated value in $\mathbf{Q}^v$ should be relatively large, and if it is assumed to be relatively fixed in time, its $\mathbf{Q}^v$ entry should be small. Unfortunately this kind of forgetting is somewhat ad hoc, meaning that deterministic estimates of the length and nature (i.e. linear, exponential, etc.) of the effective memory are impossible. Some trial and error is usually necessary before the desired output can be achieved.

Test Cases

Several test-case scenarios were run to ascertain the Kalman Filter's ability to determine parameters for both time-stationary and time-varying systems. The model structure used to produce the results shown in Figure 3.4 was a simple third order moving average filter comprised of the arbitrarily chosen coefficients $[b_{-1} = 0.0189, b_0 = 0.0093, b_1 = 0.0016]$. Two test input sequences were generated: one comprised of

white noise; the second a combination of white noise and a sinusoid, designed to represent an autocorrelated input signal. The coefficients were used to filter each of these input sequences to produce a "true" output. At this point each output sequence was corrupted with both white and colored observation noise. These output sequences are shown in the bottom panels as blue x's and green +'s, respectively. The "truth" is represented with red o's.

The Kalman Filter was used to recover the coefficients from the noisy test data using the following parameters: $\theta(0) = \mathbf{0}$; $\mathbf{P}(0) = \mathbf{I}$; $\mathbf{Q}^v = 0$, and $\mathbf{Q}^e = 0.01$. The estimated parameters are shown in the top three panels using the corresponding colored symbols. The estimates made using test data generated with colored inputs generally converge faster than those made using test data generated using white noise inputs. This is sensible because the variance of the colored inputs used to generate these test data was higher than the variance of the white noise inputs. In other words they were more dynamically exciting. Looking solely at the estimates made using white noise test inputs, however, it appears as though convergence rates favor the system corrupted by white noise over the system corrupted by colored noise. It has been hypothesized by the author that this is because the linear estimation theory behind the Kalman Filter actually assumes white observation noise, but there is nothing in the algorithm itself to supports this idea.

The similar convergence rates for the systems driven by colored inputs, regardless of observation noise properties, is equally confusing. A review of the literature provided nothing to suggest that this is a general property of the Kalman Filter when used for system identification. It might be a function of a resonance between the colored inputs and the colored observation noise, but this is impossible to determine from the data presented here. Further study may be warranted, involving the use of different inputs and observation noise sequences with similar statistical properties, as well as a more thorough literature review to determine, either empirically or analytically, whether this

is indeed a general property of the Kalman Filter.

The next set of test-case scenarios was designed to assume time-varying filter coefficients. The model structure used was similar to the one just discussed, but b_{-1}, b_0, and b_1 were systematically varied as scaled sine, cosine, and tangent functions, respectively. The first two coefficients were designed to be linearly independent of one another and vary smoothly over time, while the third coefficient was designed to have a discontinuity in order to test the Kalman Filter's ability to recover from such a dramatic event. The same input and observation noise sequences as those just described were used to generate the test sequences, while the only change to the Kalman Filter itself was to make $\mathbf{Q}^v = 10^{-5}$, thereby allowing the parameters to track the changing dynamics.

The results are shown in Figure 3.5. The estimation generally converges faster when the input driver is a white noise process (left-hand panels), as opposed to the colored noise inputs (right-hand panels), especially when recovering from the discontinuities in the third coefficient variations. The white or colored nature of the added observation noise appears to have little influence on convergence rates regardless of the type of model input. The ad hoc nature of the Kalman Filter's forgetting, however, makes it difficult to make definitive statements along these lines when process noise is allowed. The Kalman Filter appears to have little trouble tracking the slowly varying coefficients with the process noise assumption made here, so it is likely that reducing $\mathbf{Q}^v$ in general will result in less noisy estimates of the filter coefficients. A more sophisticated approach might involve setting the process noise for b_1 to be slightly higher than it is for the other two parameters. This will allow $b_1$to adapt more quickly after encountering the discontinuities at $t = 400$ and 1200, which in turn will minimize the perturbing effect of the discontinuities on the other two parameter estimates (remember, the correction is determined by multiplying the Kalman Gain by the prediction residual, which will be the same for each of the parameters regardless of which one varies most).

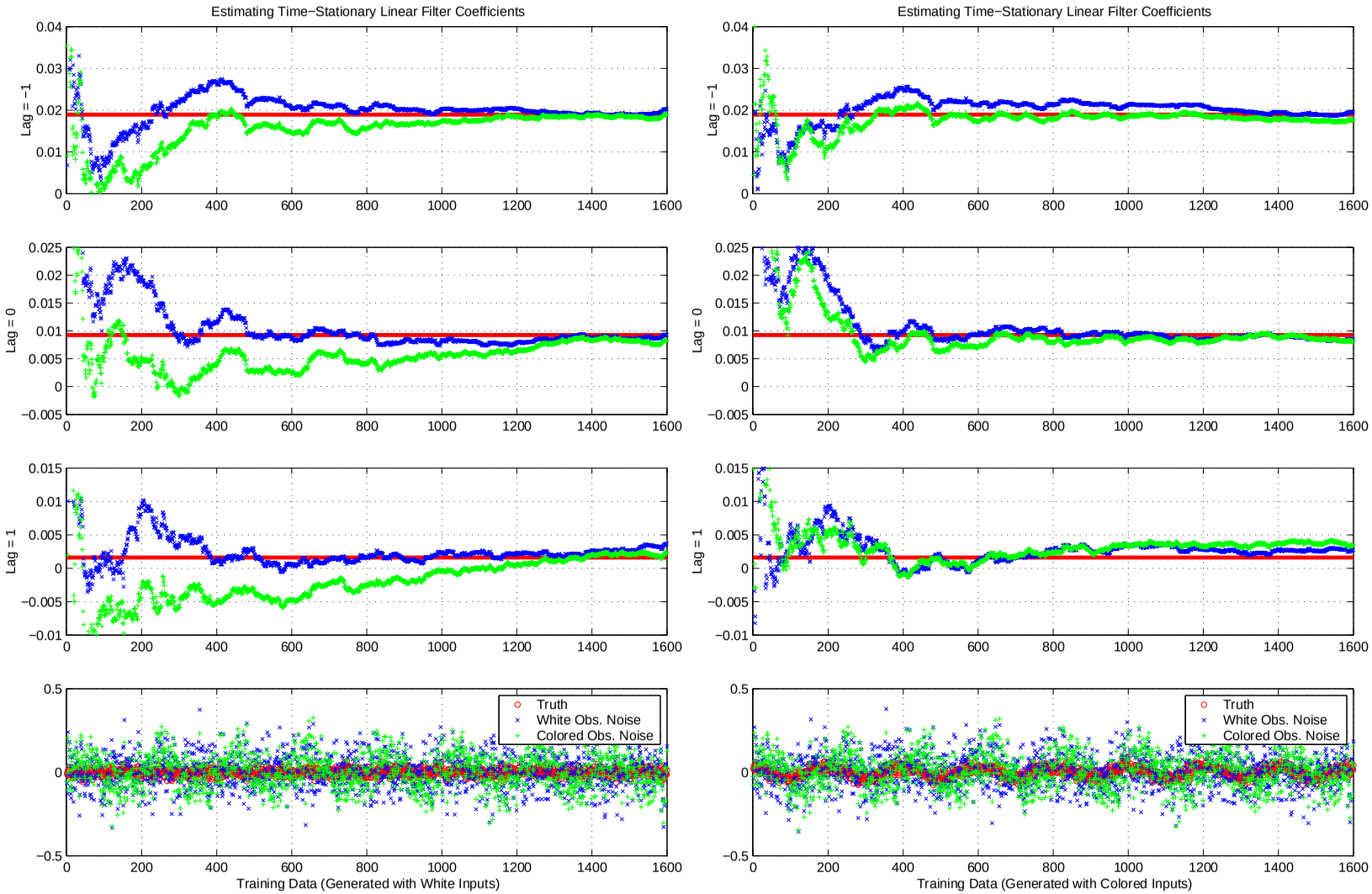


Figure 3.4: Test-case stationary FIR parameter identification using the Kalman Filter.

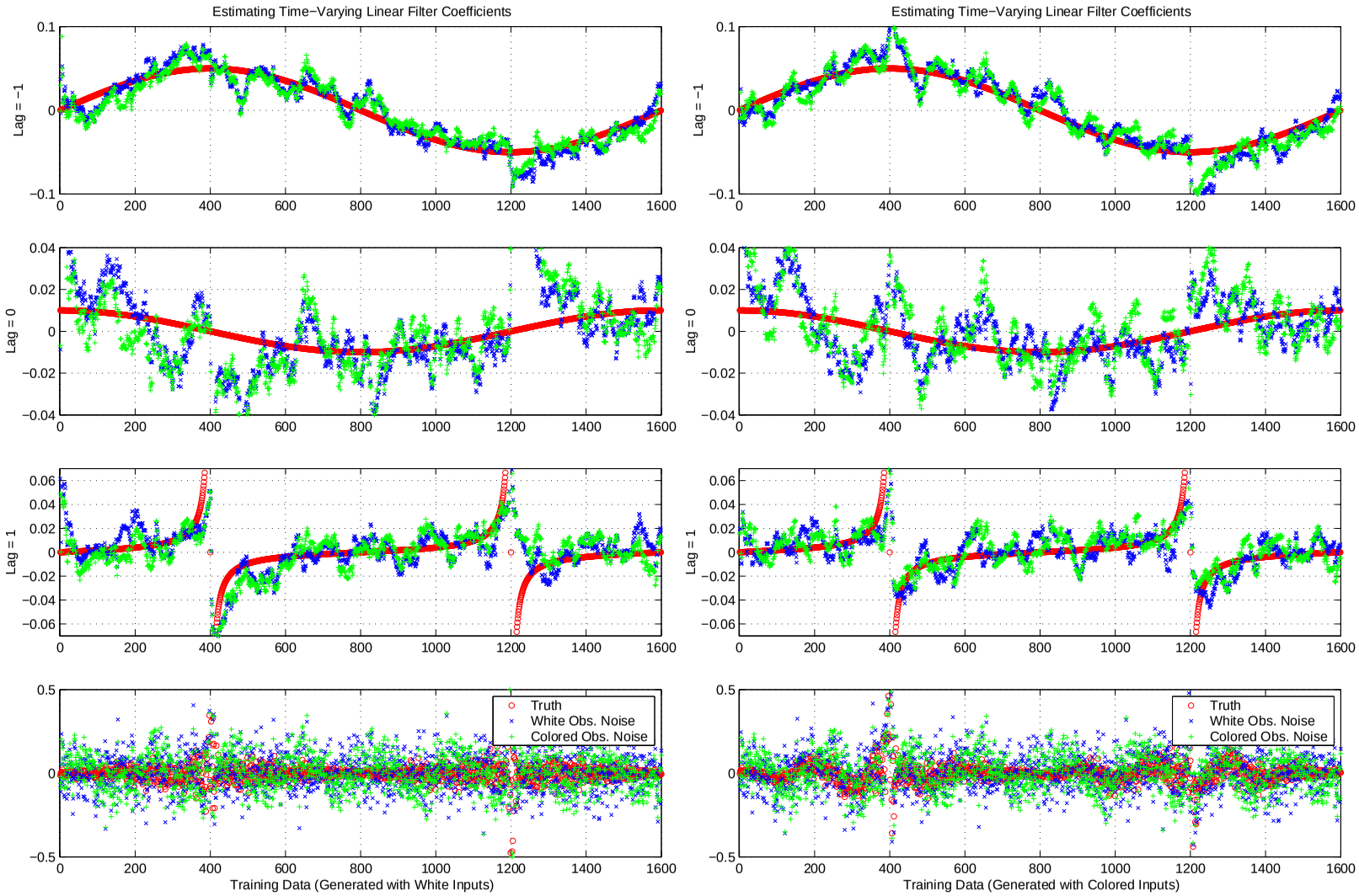


Figure 3.5: Test-case time-varying FIR parameter identification using the Kalman Filter.

3.3 Chapter Summary

Chapter 3 has provided an overview of several different system identification techniques that could be used to determine the coefficients of simple finite impulse response (FIR) linear filters. One-shot batch processors like the traditional least-squares and more flexible weighted least-squares techniques were described. It was shown that the accumulation of a large regression matrix and calculation of its pseudo-inverse could be used to determine the set of coefficients that minimize the least-squares loss function J. Despite the fact that large-scale batch processors can usually be accommodated easily with today's larger computers and optimized linear algebra libraries, a recursive formulation of the least-squares (RLS) algorithm was presented that determines the parameters of time-stationary linear models in order to set the stage for more useful adaptive system identification tools.

This RLS algorithm still has value because it is amenable to the introduction of a forgetting factor λ, causing the identification algorithm to gradually disregard observations made further and further in the past, and adapt to more recent conditions. The particular case presented provided the RLS with a finite memory that weighed past observations exponentially less as time passed. For example, for $\lambda = 0.9$, new data are 3, 8, 24, and 68 times more significant than old data 10, 20, 30 and 40 samples prior. The RLS with forgetting suffers from stability problems when the input is not sufficiently dynamically exciting. If input is zero, the covariance matrix $\mathbf{P}(t)$ used to help determine the correction factor $\gamma(t)$ that adjusts the model parameters at each new time-step will have a tendency to grow exponentially, potentially destabilizing the algorithm beyond recovery. Additionally, if the input is non-zero, but constant, and larger than a certain threshold determined by the current state and λ, $\mathbf{P}(t)$ and $\gamma(t)$ will quickly approach zero, regardless of the forgetting factor. This can force the algorithm into a state that may not be adequately responsive when the input once again becomes exciting.

The Kalman Filter operates in a similar manner, but it is more resilient to the stability issues faced by the RLS. The covariance "blow-up" is limited to a linear function of time, rather than exponential. In addition, $\mathbf{P}(t)$ and $\gamma(t)$ will not necessarily tend toward zero just because there is non-zero but non-exciting input. One final advantage of the Kalman Filter over the RLS is that, even though the length and shape of its effective memory is not as well defined as it is for the RLS, it allows the user to adjust the forgetting for individual parameters, something not possible with the RLS's simple scalar λ. Test case scenarios using both time-stationary and time-varying FIR models indicate that the Kalman Filter tracks non-stationary dynamics well by adjusting the filter coefficients over time in order to best fit the most recent observations.

Chapter 4

Electron Radiation Belt Response

Chapter 3 described how the Kalman Filter can be used to recursively and adaptively determine the parameters of an FIR linear filter model. A simple example was provided demonstrating how the inclusion of process noise in the Kalman Filter algorithm allows a baseline FIR filter's coefficients to vary with time as it tracks the changing dynamic properties of a non-stationary linear system.

This chapter presents the results of a validation study designed to demonstrate the Kalman Filter's ability to ingest real data from a physical system known to exhibit non-linear dynamical behavior, and systematically adjust the coefficients of a higher-order FIR linear filter to account for the non-linearities that manifest themselves in a non-stationary time series. Specifically, we treat the coupling between solar wind inputs and energetic radiation belt electrons. Not only are electron flux predictions significantly improved, but careful analysis of both the time-varying model parameters and the structure of the resulting residuals provide valuable insight into the underlying dynamics governing the observed variability of the system being modeled. Both single and multiple input filters are analyzed and compared in order to partly separate non-stationary behavior that results from failure to include all relevant inputs from non-stationary behavior that is a direct consequence of non-linear dynamics.

4.1 Single-Input Models

The solar wind bulk speed has historically been considered the primary driver of radiation belt electron flux variability (Paulikas and Blake, 1979; Baker et al., 1986), and as such, it has been used in numerous FIR filter-based dynamics studies as well as operational forecast models (see Section 1.3 and references therein). Indeed, Rigler et al. (2004) focused almost exclusively on this single solar wind input in a study using adaptive linear filters to predict radiation belt electrons. Identical analyses using additional solar wind measurements as single-input drivers were performed subsequently and are presented here.

4.1.1 Response Function Profiles

Stationary Responses

To begin, time-stationary causal FIR filter coefficients were determined from the solar wind and radiation belt electron flux data described in Chapter 2, using the standard least-squares techniques described in Chapter 3. These were then ordered by geomagnetic L-shell, and are presented in Figure 4.1 as a 2-dimensional (time-lag vs. L-shell) surface, with contours that represent varying strengths of the ideal linear response to discrete unit impulses in each of the solar wind inputs. To be more precise, the units associated with the filter coefficients are not actually "Log Flux", but rather they are the units required to give "Log Flux" as output when multiplied by the corresponding solar wind input. The first panel is mostly consistent with the prior work of Vassiliadis et al. (2002), who identified at least two spatially distinct regions in terms of dynamic response to changes in the solar wind speed. The first covers a broad range of the higher L-shells ($L \geq 4$), and is characterized by a "dropout" (i.e. negative log-flux response), followed by a 1-2 day delayed peak, and finally an extended decay period

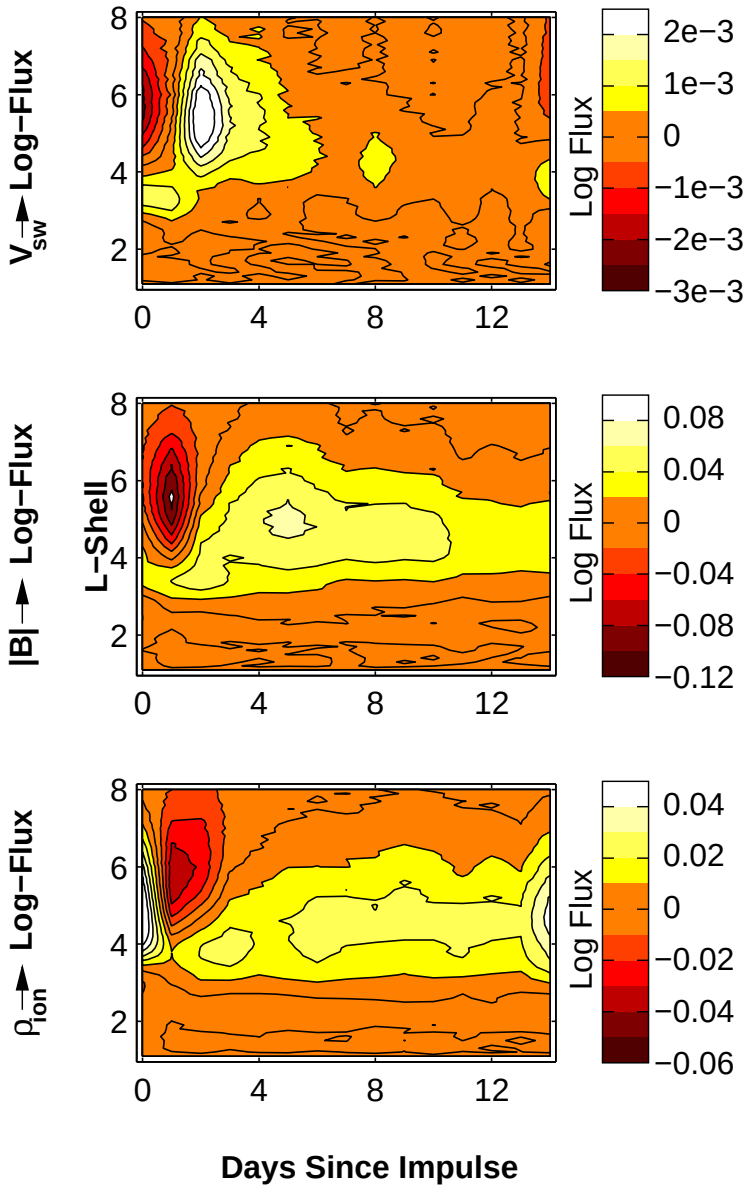


Figure 4.1: Stationary finite impulse response (FIR) functions are plotted vs. L-shell. The abscissa represents the number of days since a discrete impulse in the solar wind bulk speed (V_{sw}), IMF magnitude ($|\mathbf{B}|$), and ion density (ρ_{ion}) respectively.

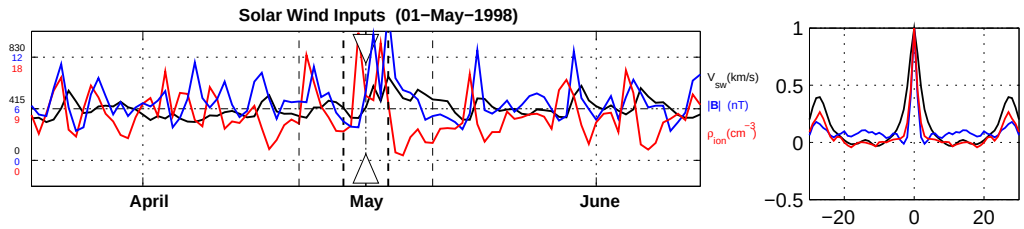


Figure 4.2: A 90 day sample of solar wind input data is shown on the left. While not 100% repeatible, it is clear that enhancements in solar wind ion density are typically followed quickly by enhancements in magnetic field strength. Increases in solar wind speed then follow 1-2 days later. Normalized autocorrelation functions on the right better illustrate the more impulsive nature of enhancements in ion density and magnetic field strenth, as their time-scales are typically less than half the time-scale of solar wind speed enhancements.

that is inversely proportional to L-shell. The second is more spatially and temporally confined. It lies between L=3 and L=4, and exhibits a more-or-less immediate enhancement (the delay is less than a day) that returns to background levels just as quickly.

Panels 2 and 3 of Figure 4.1 offer similar electron flux response profiles for the interplanetary magnetic field magnitude ($|\mathbf{B}|$) and solar wind ion density (ρ_{ion}). They possess almost the same spatial/temporal organization to the $V_{sw} \rightarrow \log(\text{flux})$ response function profile, especially the dropout $\rightarrow$ peak $\rightarrow$ decay response at higher L-shells. The most straight-forward explanation for this high degree of similarity is that the coefficients for each of the independently determined response function profiles are actually calculated using, to a large extent, the same sequences of electron flux data. There are at least two circumstantial arguments in support of this. First, a fundamental difference between the profiles is that all of the peaks and/or troughs tend to occur 1-2 days later in the $|\mathbf{B}| \rightarrow \log(\text{flux})$ and $\rho_{ion} \rightarrow \log(\text{flux})$ profiles than the $V_{sw} \rightarrow \log(\text{flux})$ profile, corresponding to the fact that, on average, changes in $|\mathbf{B}|$ and ρ_{ion} tend to precede changes in V_{sw} by a day or two. Second, the $|\mathbf{B}| \rightarrow \log(\text{flux})$ and $\rho_{ion} \rightarrow \log(\text{flux})$ profiles appear more extended in time, almost as if the $V_{sw} \rightarrow \log(\text{flux})$ profile had been stretched like

a rubber sheet. This can be explained by noting that the largest changes in $|\mathbf{B}|$ and ρ_{ion} tend to be more impulsive (i.e., they exhibit sharper peaks) than the largest changes in V_{sw}. Both of these relationships can be seen clearly in Figure 4.2, and have been noted previously in the literature (e.g., Baker, 2000). If the response functions are indeed derived from the same output sequences, then the non-zero portions of the $|\mathbf{B}| \rightarrow \log(\text{flux})$ and $\rho_{ion} \rightarrow \log(\text{flux})$ response functions must necessarily "stretch out" in time in order to produce the same output as the $V_{sw} \rightarrow \log(\text{flux})$ response function, simply due to the integrated nature of the discrete convolution.

Nothing noted in the response function descriptions thus far helps to understand exactly which of the inputs, V_{sw}, $|\mathbf{B}|$, or ρ_{ion}, is the most important driver of the observed changes in radiation belt electrons, although some insight can be gained by analyzing the relative timing of the peaks, and more importantly, by calculating the simultaneous response functions in order to separate statistically the pertinent dynamics (see Section 4.2). What is implicit to the above explanation, however, is the fact that most of the response that the filters describe is actually the relaxation of the radiation belts after a relatively quick and impulsive enhancement. In other words, the FIR response function profiles are mostly representative of the same set of internal dynamical processes, which can be described roughly as an over-damped harmonic oscillator if the apparent exponential decay is to be believed. The inability of the FIR model to separate the directly driven response from the system relaxation is inherent to its structure (Johansson, 1993; Ljung, 1999; Nelles, 2001), since it possesses no mechanism for feeding back its own output. More sophisticated linear filters designed to explicitly model dynamic feedback are suggested in Chapter 5, and described in detail in Appendix C.

One part of the delayed enhancement seen in the $\rho_{ion} \rightarrow \log(\text{flux})$ response functions that is almost certainly NOT a consequence of shared magnetospheric dynamics is the strong impulsive enhancement seen across a broad range of L-shells at zero-lag. This appears to be isolated to the $\rho_{ion} \rightarrow \log(\text{flux})$ response profile, although the relative

timing of impulses in each of the inputs begs the question of whether this enhancement simply occurs at negative time-lags that are not shown in Figure 4.1. In fact, weak acausal negative responses to increases in solar wind speed have been noted previously for this region (Vassiliadis et al., 2002, 2003b; Rigler et al., in press; the last did not actually discuss this feature, but merely presented the response function profiles). Solar wind speed and ion density are highly anti-correlated, due mostly to well-known processes associated with incompressible fluid flows, so these responses are actually consistent with one another. Since it is very unlikely that there is a causal relationship that allows observed changes in radiation belt electron flux to forecast changes in solar wind speed, it would appear that the $\rho_{ion} \rightarrow \log(\text{flux})$ response is more likely to be physical.

In contrast, there also appears to be the edge of peak near the 14-day boundary of the $\rho_{ion} \rightarrow \log(\text{flux})$ profile. Such a delayed peak in a response function is unlikely to be physical, but it is too structured to be an "edge" effect (i.e. unrealistic values for the first and/or last coefficients in a response function that may be caused by poorly chosen boundary conditions in the regression formulation). In addition, it is relatively persistent through time, as will be shown shortly in an analysis of the adaptive response function profiles. An explanation can be found in the fact that there is a weak but significant positive correlation between V_{sw} and ρ_{ion} at a lag of around 13-14 days. There is no unambiguous physical explanation for this, but it is probably related to a solar wind sector structure that is more complex than portrayed by Figure 1.2. In any case, the long-delayed positive peak in the $\rho_{ion} \rightarrow \log(\text{flux})$ response is most likely due to much earlier solar wind speed increase, and not any enhancement in the ion density. This hypothesis finds strong support in the multi-input FIR profiles presented in Section 4.2.1, where this response peak disappears in models that ingest solar wind speed and ion density data simultaneously.

Non-Stationary Responses

Vassiliadis et al. (2002) also looked into seasonal, semi-annul, and solar cycle-dependent variations in the V_{sw} driven radiation belt electron flux response by analyzing FIR profiles calculated from appropriate sub-sets of the training data. A similar analysis was conducted here using a slightly different approach. The Kalman Filter generates a new set of filter coefficients with each new set of daily-averaged observations, which in this case was every day. Different sub-sets of the coefficients were then averaged to produce response function profiles corresponding to different climatological effects. One advantage to this approach is that any edge effects resulting from poorly chosen initial and/or boundary conditions are eventually "forgotten" in a recursive identification scheme with finite memory. For reference, the observation error covariance of the Kalman Filter used to generate all of the adaptive predictions was set equal to one, so all observations were equally weighted, and the process noise covariance was set to the somewhat arbitrary value of 10^{-6}.[1]

The panels in Figure 4.3 were each generated by averaging three months worth of FIR profiles centered on the relevant northern hemisphere solstice or equinox. The resulting response function profiles are very similar to the profiles presented in the Vassiliadis et al. paper. In particular, a slight weakening of the positive delayed response observed at higher L-shells occurs during the summer. This is consistent with the so-called "summer doldrums" (Baker et al., 1999a), a regular lessening of radiation belt electron flux during the northern hemisphere summer. One significant difference is that the Vassiliadis et al. results indicated that the low-altitude immediate response disappears in the summer, while the results in Figure 4.3 indicate that this localized response

[1] Actually, a substantial amount of theoretical analysis and trial and error was undertaken, using $V_{sw} \rightarrow \log(\text{flux})$ models, before signifi cant prediction improvements were realized using the covariance values listed. The same values were then applied to Kalman Filters for the other single-input models, as well as multi-input model discussed in Section 4.2. It is shown in sequel that one or the other, or both of these covariance values were ill-suited for the non-$V_{sw} \rightarrow \log(\text{flux})$ models and an explanation is put forth as to why this was so. No changes were made to correct this problem for this thesis because it seemed to be a valuable lesson-learned that should be passed on to the reader.

is, if anything, slightly enhanced in the summer. This discrepency cannot be explained via simple visual examination of the response function profiles. Oddly, the text of the Vassiliadis et al. paper suggested a solstice-equinox asymmetry that was not apparent to this author upon examination of either set of seasonally-averaged response function profiles (i.e., the Vassiliadis et al. plots, or those shown here). More detailed analysis of the actual coefficients and/or the methods used to seasonally bin the data is warranted, but has not been conducted as of the publication of this thesis. There is little or no significant variation in the other inputs' response profiles.

As seen in Figure 4.4, much more pronounced variations occur over periods longer than a year, suggesting strong solar cycle dependence in the response functions. The $V_{sw} \rightarrow \log(\text{flux})$ profiles indicate a radiation belt response that is much stronger in 1997 than the other years shown, consistent with similar plots presented by Vassiliadis et al (2002), but otherwise they exhibit a trend toward weaker responses as one progresses from solar minimum (~1996) toward solar maximum. The $|\mathbf{B}| \rightarrow \log(\text{flux})$ electron flux response profiles exhibit a very obvious trend toward weaker responses as one progresses toward solar max. In contrast, the long-delayed positive response in the $\rho_{ion} \rightarrow \log(\text{flux})$ profiles appear to migrate to shorter and shorter time-delays, while the more immediate negative response gets stronger. It begins to increasingly resemble the other two profiles, although it never fully loses the 14-day peak described earlier. Base on earlier arguments, this suggests that, more and more, the adaptive estimation of time-dependent filter coefficients for both the $V_{sw} \rightarrow \log(\text{flux})$ and $\rho_{ion} \rightarrow \log(\text{flux})$ profiles is being dictated by common electron flux observations.

4.1.2 Predictions vs. Observations

The primary objective of this thesis is to demonstrate the power of the Kalman Filter as an adaptive space weather prediction tool capable of tracking the highly variable, and ultimately non-linear, dynamics governing the coupling between the solar

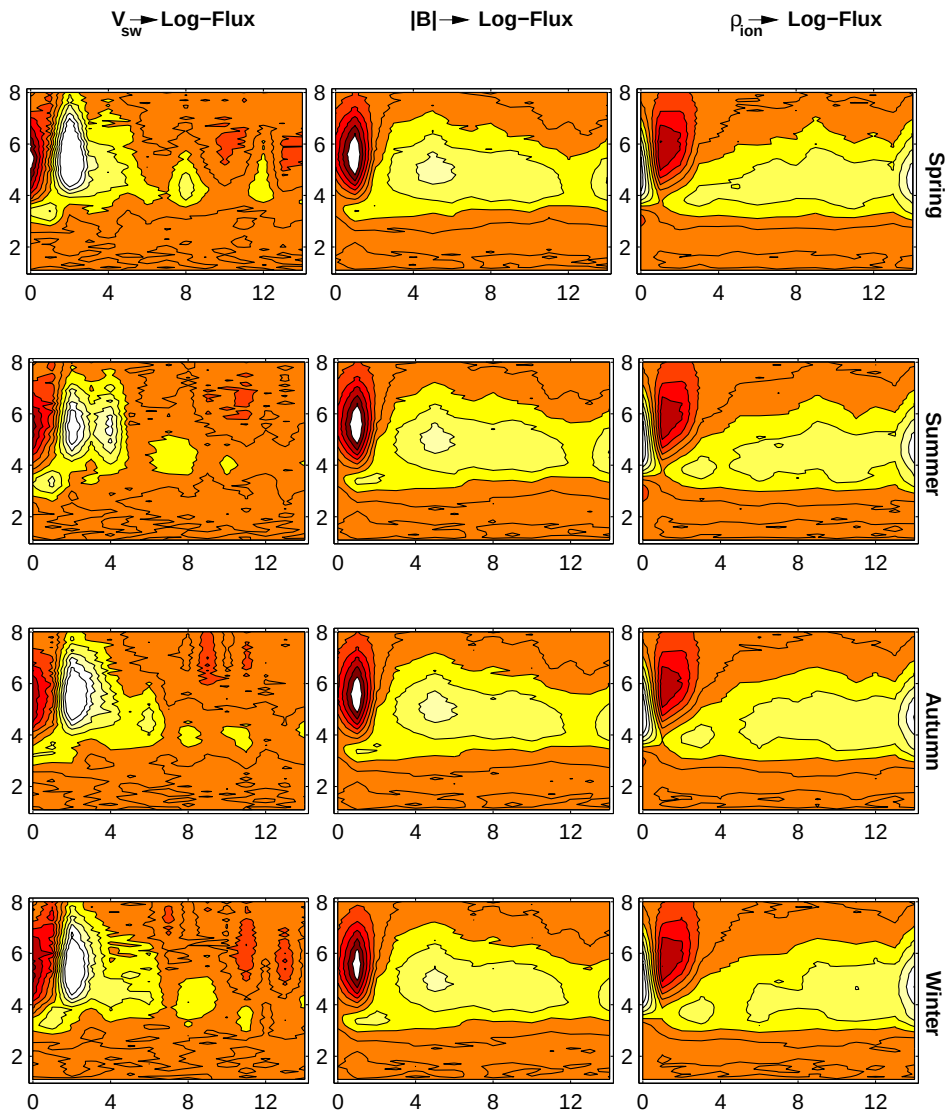


Figure 4.3: Seasonally averaged (centered on the listed solstice/equinox) FIR response functions are plotted vs. L-shell. The ordinates, abscissas, and color scales are identical to those used in Figure 4.1.

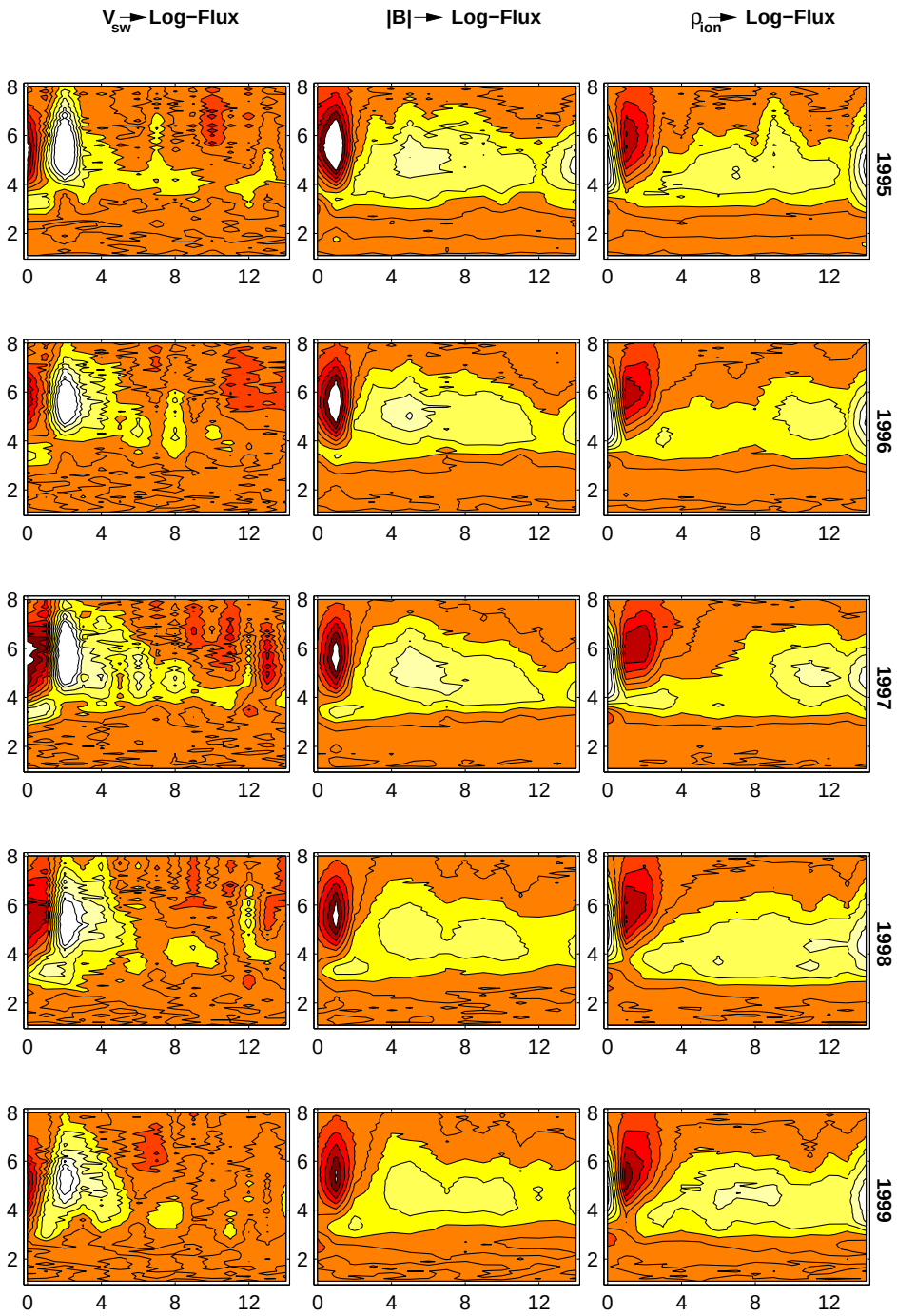


Figure 4.4: Annually averaged FIR response functions are plotted vs. L-shell. The ordinates, abscissas, and color scales are identical to those used in Figure 4.1.

wind and the radiation belt electron fluxes. A detailed analysis of actual predictions and their residuals is therefore necessary before conclusions can be drawn regarding the appropriateness of Kalman Filter-based adaptive FIR models.

Figure 4.5 plots predictions made from both the time-stationary and time-varying response function profiles presented previously. One will note the the lack of predicted dynamics for most of 1994 and subsequent shorter periods. These correspond to data gaps in the OMNIWeb solar wind inputs that were simply replaced by the appropriate historical averages, as described in detail in Chapter 2. They have been included in this, and subsequent plots of predictions and residuals, because, for select instances, they help to illustrate the transient characteristics of the adaptive system identification scheme used (this is not always true, but the data remains plotted, regardless, for the sake of consistency). Data from these periods were not included in the calculation of each model's respective in-sample prediction efficiencies (PE). PE is calculated according to Equation 4.1 (PE is discussed in more detail below), and is plotted to the right of each prediction profile in Figure 4.5.

$$\mathrm{PE} = 1 - \left(\left(\frac{1}{\mathrm{N}} \sum_{\mathrm{i}=1}^{\mathrm{N}} \mathrm{err}_{\mathrm{i}}^{2} \right) / \sigma_{\mathrm{obs.}}^{2} \right) \tag{4.1}$$

From the stationary prediction time series alone, it is immediately clear that the relative importance of each input type changes with time. The V_{sw} based predictions are strongest early on, especially in 1995, with some less extreme enhancements in 1998 and 1999. $|\mathbf{B}| \rightarrow \log(\mathrm{flux})$ predictions are most intense later on, in particular near the end of 1998. The most obvious explanation for this phenomenon is found in the input time series for these periods (see Figure 2.2). Solar wind speed is consistently high, with periodic peaks during 1995. It is also high in 1998 and 1999, if not as consistent. The strength of the interplanetary field, however, is strongest, on average, in 1998, with several extreme enhancements extending beyond the plot's scale later in the year. The largest peaks in both of these prediction time series are readily visible in the actual ob-

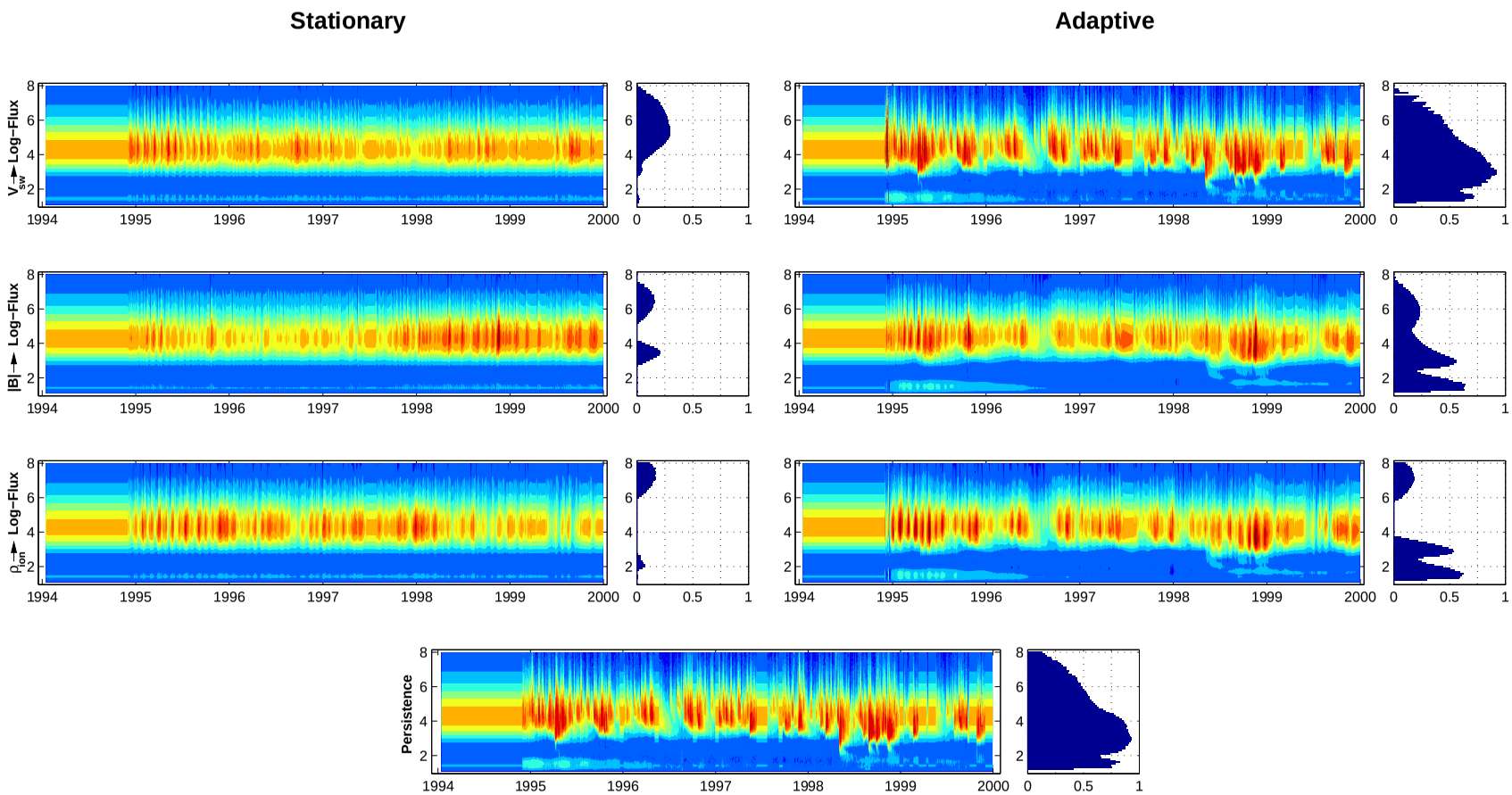


Figure 4.5: Radiation belt electron $\log_{10}(\text{flux})$ predictions from the stationary filters shown in Figure 4.1, and the adaptive filters that were used to produce the profiles in Figures 4.3 and 4.4, are plotted vs. L-shell. L-shell-dependent prediction efficiencies on the right are plotted between zero and one, even though negative values do exist, especially for the ρ_{ion} and $|\mathbf{B}|$ driven predictions. Horizontal contours correspond to data gaps where historical averages were substituted and were not used to calculate prediction efficiencies. Persistence "predictions" are shown for reference.

servations. The peaks seen in the $\rho_{ion} \rightarrow \log(\text{flux})$ predictions, however, exhibit a less regular pattern. In fact, there are several instances when ρ_{ion} based predictions appear to be anticorrelated with observations (mid-1996, late 1998 and 1999, for example).

A cursory glance at the adaptive $V_{sw} \rightarrow \log(\text{flux})$ predictions, and comparison with the observations, suggests that the adaptive scheme worked quite well. The fact that the most active regions of the radiation belts shift in position through time, rather than exhibiting the simple static response to changing inputs that can be seen in the stationary predictions, is one indication that non-linearities are being accounted for. A similar shift in the position of the peaks seen in the averaged response function profiles presented in the previous section is not immediately obvious, but analysis of the filters at shorter time scales shows that they are indeed changing in response to changes in the radiation belts, not just the varying inputs.

Improved prediction efficiencies are another indication that non-linearities are being accounted for. The adaptive model PE profiles are similar to those exhibited by the static models, simply higher. These improved PEs appears to be proportional to the PE for persistence at the similar L-shells (shown in the bottom panel of Figure 4.5). This is readily explained by the fact that in regions of high persistence the Kalman Filter is allowed more time to nudge the model parameters to properly reproduce the observed state. In the case of the $V_{sw} \rightarrow \log(\text{flux})$ predictions, the adaptation speed is fast enough that today's prediction, based on yesterday's correction, provides flux estimates that are very similar to yesterday if solar wind conditions have not changed much.

Neither the magnitudes nor the dynamic range of the other two inputs are as high as for V_{sw}, so the corresponding model parameters do not adapt as quickly from time-step to time-step. This can actually be seen in the predictions as delayed (relative to the observations) dropouts (mid-1996) and enhancements (late-1998), as well as by noting the different rates at which the predictions decay to the average background

levels substituted for real data during periods with large data gaps (note the gap in mid-1997, for example). The result is a PE profile that is generally better than for static models, but not nearly as improved as that for $V_{sw} \rightarrow \log(\text{flux})$ predictions.

Using Equation 4.1, the PE metric was calculated independently for each L-shell by subtracting the ratio of the mean squared error of the residuals over the variance of the observations from one. If the residuals comprise a zero-mean, normal (i.e. Gaussian) distribution, PE describes the fraction of dynamic variance in the data the model was able to reproduce. Even if the residuals do not meet this requirement, PE should be considered a useful "skill score" that compares the model with the sample mean, but it becomes difficult to say anything meaningful about the appropriateness of the model's dynamical structure. It is therefore useful to look more closely at the distribution, and ultimately, the correlated structure of the residuals themselves. Figure 4.6 plots time series of residuals from each of the prediction profiles in Figure 4.5 at L=6.6 (geostationary orbit), along with their corresponding distribution histograms. If the mean of the residuals deviates significantly from zero, or the distribution is skewed to one side or the other, this is an indication that the model is biased (it consistently over or under estimates the system output). It appears as though, with varying degrees of success, the adaptive models not only provide better predictions, but they tend to nudge the obviously biased residuals of the stationary models toward a more symmetric and normal distribution. This implies that the adaptive estimator is working properly (if not perfectly), since, if the local-linear assumptions underlying the algorithm are valid at each time step, residuals should be white. It is interesting to note that the persistence prediction residuals, while possessing a relatively small spread, are indeed slightly skewed. This is not unexpected however, since persistence is nothing more than a 1st order autoregressive filter ($AR_1 = 1$). AR filters are well-known to produce biased output in general (Nelles, 2001).

Occasionally even further analysis of residual structure is required in order to

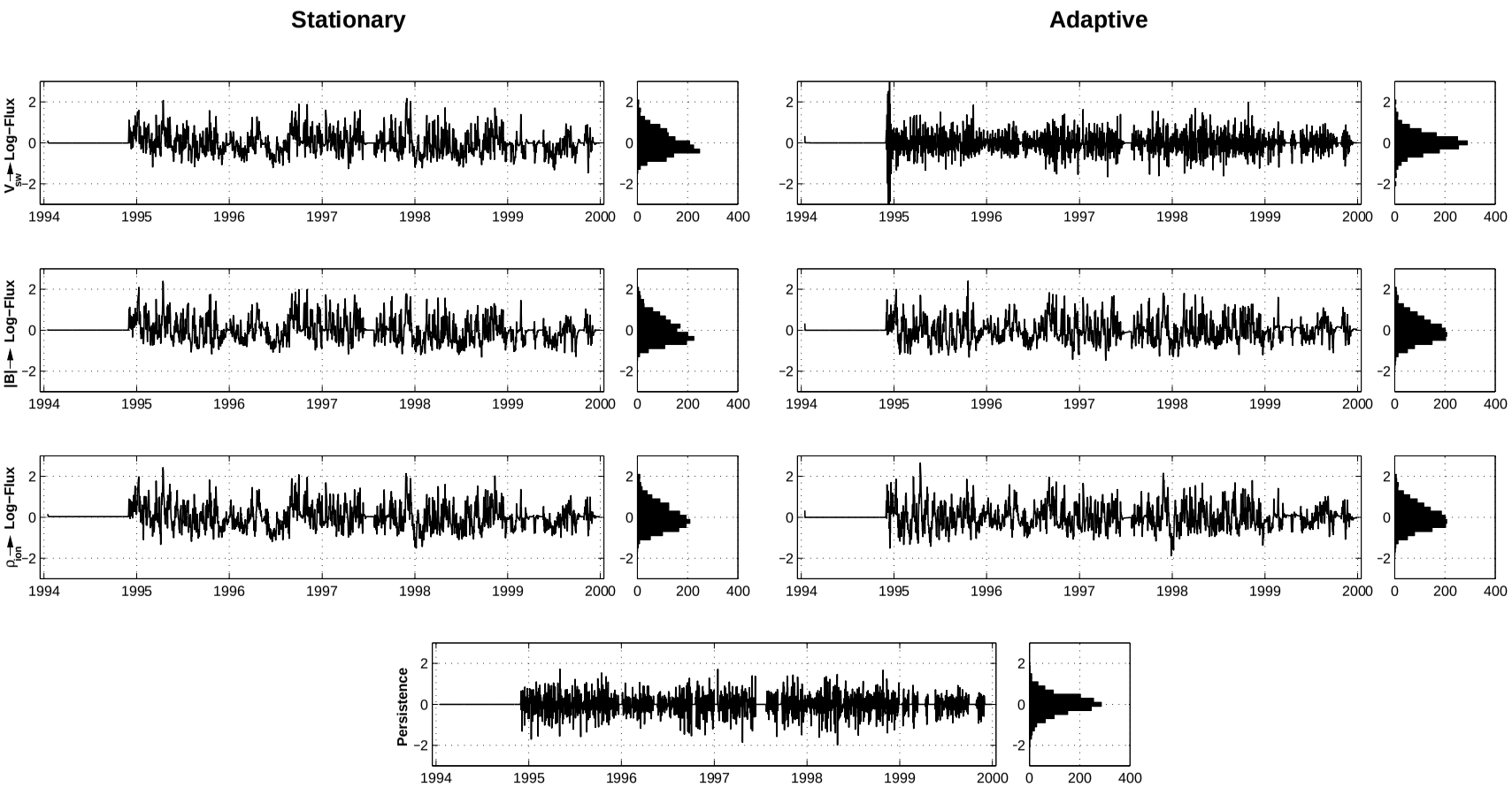


Figure 4.6: Prediction residuals for each single-input model (stationary, adaptive, and persistence) are plotted vs. time for L=6.6. The associated error distribution is displayed to the right of each as a histogram with bins defined between -2 and 2. Horizontal lines correspond to data gaps where historical averages were substituted and were not included in the histograms. Persistence "prediction" residuals are shown for reference.

determine the appropriateness of a particular model. For instance, suppose that the residuals were periodically slightly more than zero for an extended period (i.e., more than a few samples), followed by another similarly extended period where they were comparably less than zero. This would be almost impossible to discern in a histogram with any reasonable bin-size. It may be possible to identify such structure by looking closely at the residual time series, but more often than not such a trivial deviation will go unnoticed by the human eye. This kind of deviation becomes readily apparent, however, if Δt-dependent correlation functions are calculated from the residuals.

Figure 4.7 shows the auto-correlation functions (ACF) for the residuals of each of the three model types. The adaptive $V_{sw} \rightarrow \log(\text{flux})$ predictions appear to be the only model to take good advantage of the dynamic information available in the output signals (i.e., electron log-flux). The significant deviations from whiteness (coefficients that extend beyond the 95% confidence interval) in the rest of the residual ACFs indicate that there was information remaining in the output signals that that might have been used to provide better predictions (Johansson, 1993; Ljung, 1999).

The latter result is expected, since these FIR models are designed to filter solar wind inputs to produce electron flux predictions. They have no explicit autoregressive or persistence terms (persistence is only implicit to the adaptive models, with its influence being completely dependent on the adaptation rate). When the autocorrelations of persistence predictions are studied, the ACF coefficient amplitudes drop significantly (top panel, Figure 4.9).

A more appropriate metric might be the cross-correlation between the prediction residuals and the inputs used to create them (V_{sw}, $|\mathbf{B}|$, and ρ_{ion}). The generally small correlation coefficients seen in both the stationary and adaptive model residual CCFs of Figure 4.8 make sense, since 1) it has been historically established that a significant linear relationship exists between these solar wind inputs and radiation belt electrons, and 2) these models are essentially calculated using the correlations between the solar

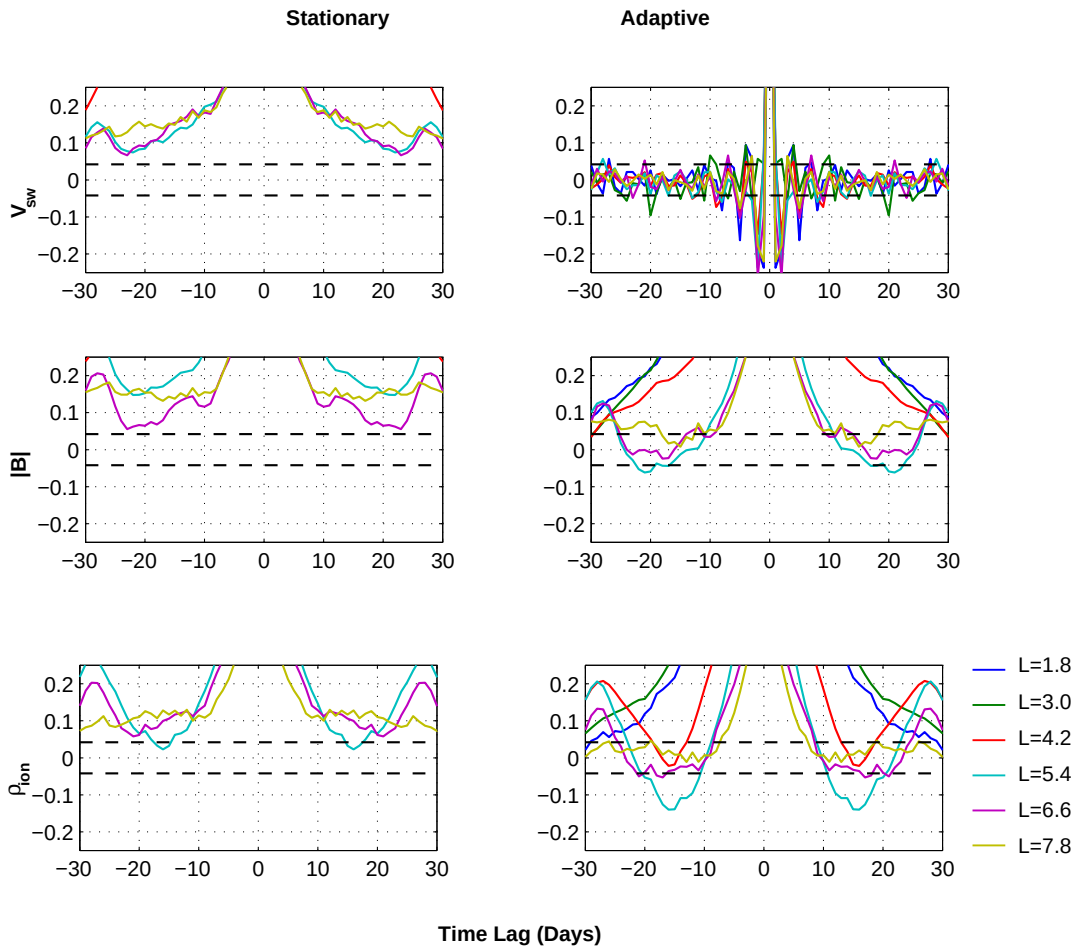


Figure 4.7: Auto-correlation functions (ACFs) of prediction residuals for stationary and adaptive models are plotted for select L-shells. The horizontal dashed lines to either side of zero demarcate the 95% confidence interval wherein correlation coefficients for lags$\neq$0 must fall for the signal to be considered "white".

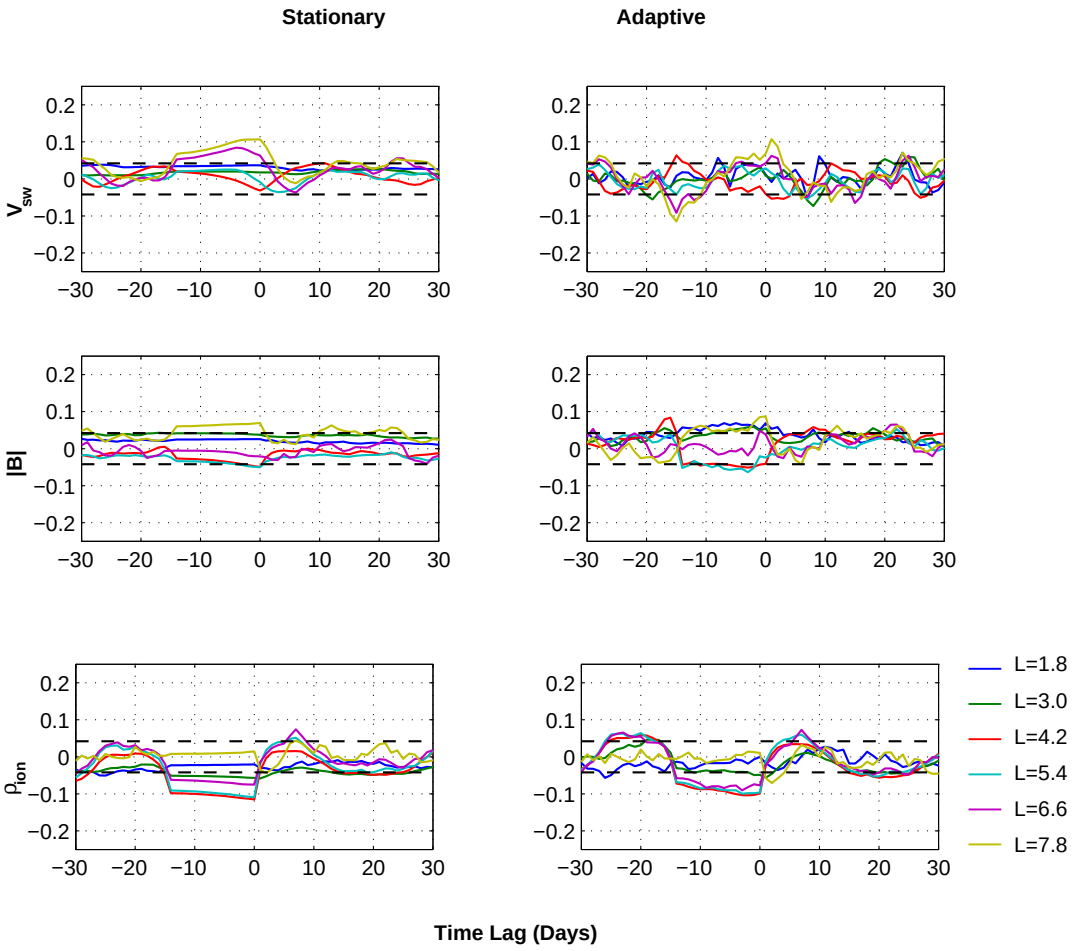


Figure 4.8: Cross-correlation functions (CCFs) of prediction residuals with their corresponding inputs for stationary and adaptive models are plotted for select L-shells. The horizontal dashed lines to either side of zero demarcate the 95% confidence interval wherein correlation coefficients for lags$\neq$0 must fall for the signals to be considered uncorrelated.

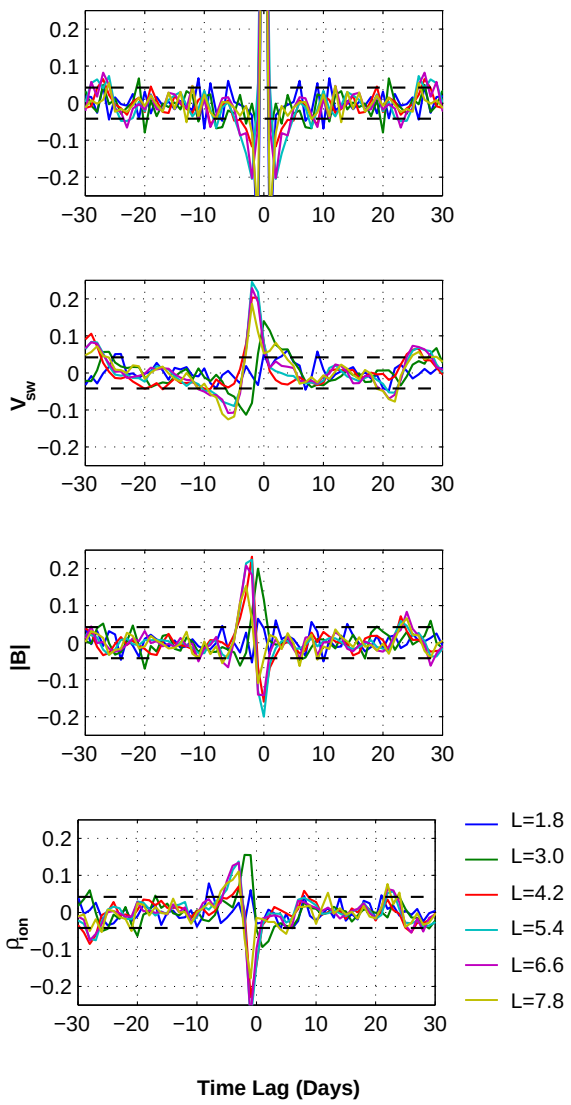


Figure 4.9: ACF (top) and CCFs of persistence "prediction" residuals with each input. Obviously persistence uses no actual input, but the CCFs illustrate the fact that including input should improve prediction quality. As before, the horizontal dashed lines to either side of zero demarcate the 95% confidence interval wherein correlation coefficients for lags$\neq$0 must fall for the signals to be considered uncorrelated.

wind inputs and the radiation belt electron fluxes. To further emphasize this point, note the relatively large correlations in the persistence CCFs. This is expected because persistence does not explicitly incorporate any of these inputs.

When interpreting these results, it is important keep in mind that a correlation at positive lags implies some sort of feedback in the input, violating the implied assumptions of causality, but not necessarily saying anything of consequence about the structure of the dynamical models (Ljung, 1999). Significant correlations at negative lags signify an actual deficiency in the model structure, which might be due to too few coefficients (insufficient model order), to missing inputs, or any number of things, up to and including the possibility that there are significant non-linear dynamics that are not being picked up by the model.

4.2 Multi-Input Models

When one studies the FIR profiles, predictions, PE profiles, and residual correlation functions in the last section more closely, it becomes increasingly clear that different solar wind inputs excite different regions of the radiation belts preferentially. This phenomenon was noted observationally as early as ten years ago (Baker et al., 1994), but it has been studied in detail only recently, using a variety of both solar and solar wind inputs, as well as select geomagnetic indices (Vassiliadis et al., 2003a, 2004). As noted in the latter articles, however, definitive conclusions about physical causality, based on the predictions of separately identified single input filters, should be made with caution since observed changes in the system output might be similarly predicted using different filtered inputs. There are at least three reasons for this lack of uniqueness: 1) poorly resolved input/output data may result in the aliasing of dynamic events into common time-bins; 2) a system's response may be to a multiplicative or otherwise NON-linear combination of different input variables; and 3) a system may respond to a linear combination of different inputs.

Item 1 is a common problem in empirical modeling, and a particular problem for this study due to the highly-averaged nature of the data being used. Poorly resolved data is a largely self-imposed limitation for this study since, as described in Chapter 2, those data used in the current study started out at a higher resolution, and were subsequently smoothed and resampled at the current 24-hour cadence. Some possible remedies for the data problems that motivated this averaging are discussed in Chapter 5. In addition, little can be done about item 2 except to put considerably more thought into the choice of model inputs than was done for this study, another topic discussed in Chapter 5. Even then, the perfect combination of physical parameters to satisfy these more rigorous input/output requirements is often difficult or impossible to measure directly. The modeler must then resort to employing ad-hoc transformations of imperfect observations, with causes ranging from poorly intercalibrated instruments to, in rare instances, imperfect temporal alignment of observations from different sources.

The third possible cause of similar predictions from different inputs listed above can easily be accounted for by arranging the regression matrix and parameter vector in the standard least-squares estimation procedure, or the regression and parameter vectors in the Kalman Filter algorithm, to solve for the response to an arbitrary number of different simultaneous inputs. Assuming the dynamic range of the training data is large enough to separate linear from non-linear combinations (i.e., xy is not often equal in value to x+y), the resulting multi-input filter profiles will be independent of one another and amount to dynamically separable responses to the different inputs.

4.2.1 Response Function Profiles

Stationary Responses

Figure 4.10 presents time-stationary multi-input FIR profiles that were determined using the techniques described in Chapter 3. The first notable difference between

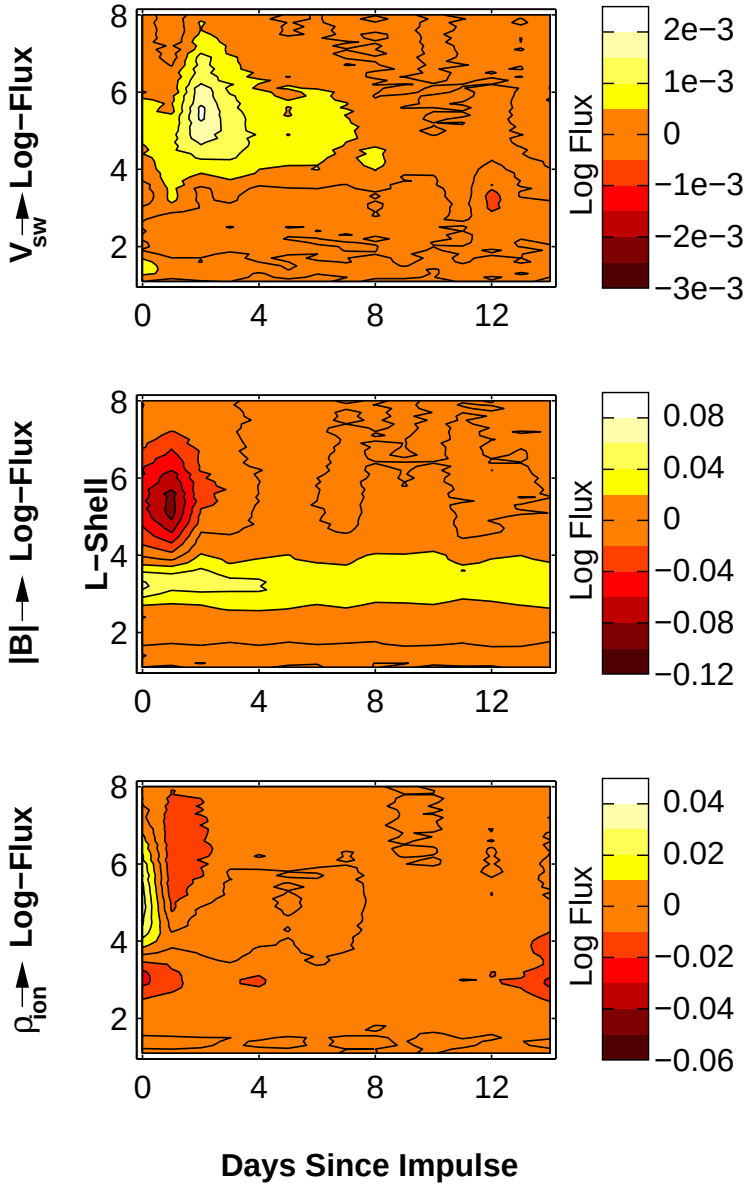


Figure 4.10: Stationary multi-input finite impulse response (FIR) functions are plotted vs. L-shell. The abscissa represents the number of days since a discrete impulse in the solar wind bulk speed (V_{sw}), magnetic field magnitude ($|\mathbf{B}|$), and ion density (ρ_{ion}) respectively.

these response profiles and those derived from single-input filters (see Figure 4.1) is that the extended positive response noted at higher L-shells for all three single-input profiles is only evident in the $V_{sw} \rightarrow \log(\text{flux})$ component of the multi-input response profiles. Since V_{sw} is better correlated with changes in electron flux at these altitudes and time-lags, the corresponding filter coefficients receive most of the weighting. In contrast, the more immediate dropout seen at higher L-shells was apparently better correlated with $|\mathbf{B}|$ than any other input, and was similarly isolated in the $|\mathbf{B}| \rightarrow \log(\text{flux})$ profile. This, of course, all assumes that no non-linear combination of V_{sw} and $|\mathbf{B}|$ exists that would be impossible to separate using linear response functions. The typically low cross-correlation between these two input variables suggests that this is not the case.

Another change from the single-input stationary FIR profiles is what appears to be a migration of the short-lived positive response, previously noted at L=3-4 in the $V_{sw} \rightarrow \log(\text{flux})$ profile, to a much more extended response at similar L-shells in the $|\mathbf{B}| \rightarrow \log(\text{flux})$ profile. One explanation for this stretching was given in the previous section, simply pointing out that the typically more impulsive changes in $|\mathbf{B}|$ relative to changes in V_{sw} require a more extended response in order for the convolution to produce the same observed system output. The extent to which this response has been stretched, however, is surprising. There is a slightly negative response at these altitudes in the $\rho_{ion} \rightarrow \log(\text{flux})$ profile that may serve to temper this response somewhat. This is intriguing because it is consistent with the corresponding stationary PE profile from Figure 4.5, which shows significant predictive ability around L=3 despite the fact that the single-input $|\mathbf{B}| \rightarrow \log(\text{flux})$ profile exhibited no such positive response.

The $\rho_{ion} \rightarrow \log(\text{flux})$ component of the multi-input profile shown in Figure 4.10 is quite flat at higher altitudes. It again appears as though the extended positive response, previously attributed to the internal radiation belt dynamics that govern the exponential decay of electron flux, was reallocated to the $V_{sw} \rightarrow \log(\text{flux})$ FIR profile, just as it was for the $|\mathbf{B}| \rightarrow \log(\text{flux})$ profile. The mysterious 14 day peak in the

$\rho_{ion} \rightarrow \log(\text{flux})$ profile also disappeared. As previously mentioned, it is very unlikely that such a strong peak occurring so late in the response profile could have any reasonable physical cause. The 14 day peak was therefore the likely result of solar wind speed interactions with the magnetosphere 14 days prior when there was no corresponding ρ_{ion} enhancement. What little structure remains in the $\rho_{ion} \rightarrow \log(\text{flux})$ profile is confined to the 0-2 day period immediately following the solar wind impulse, and seems to be a damped version of the same response region in the corresponding single-input profile shown in Figure 4.1.

Non-Stationary Responses

Figures 4.11 and 4.12 were produced in the same manner as their single-input counterparts presented previously in Figures 4.3 and 4.4. In light of the preceding discussion of stationary multi-input FIR profiles it was surprising to note that these seasonally and annually-averaged adaptive FIR profiles were not more like the static multi-input responses in Figure 4.10. In fact, in some ways they have more in common with their single-input counterparts (neglecting the reduction in overall response that comes from its redistribution across three different input drivers). Most notable, the migration of the L=3-4 response cannot be seen at all, and it remains exclusively in the $V_{sw} \rightarrow \log(\text{flux})$ profile. The extended positive response at higher L-shells is mostly gone from the $|\mathbf{B}| \rightarrow \log(\text{flux})$ profile, but it is still very noticeable in the $\rho_{ion} \rightarrow \log(\text{flux})$ profile, even though the 14 day peak is gone.

One explanation for the inconsistent separation of linear responses between the static and adaptive multi-input profiles is related to the different adaptation rates of each model mentioned briefly in Section 4.1.2. Since the $V_{sw} \rightarrow \log(\text{flux})$ model adapts more quickly due to its larger absolute value and the larger dynamic range of the solar wind speed measurements, it is able to modify the response function to predict properly the observed flux changes before either of the other models even realize anything

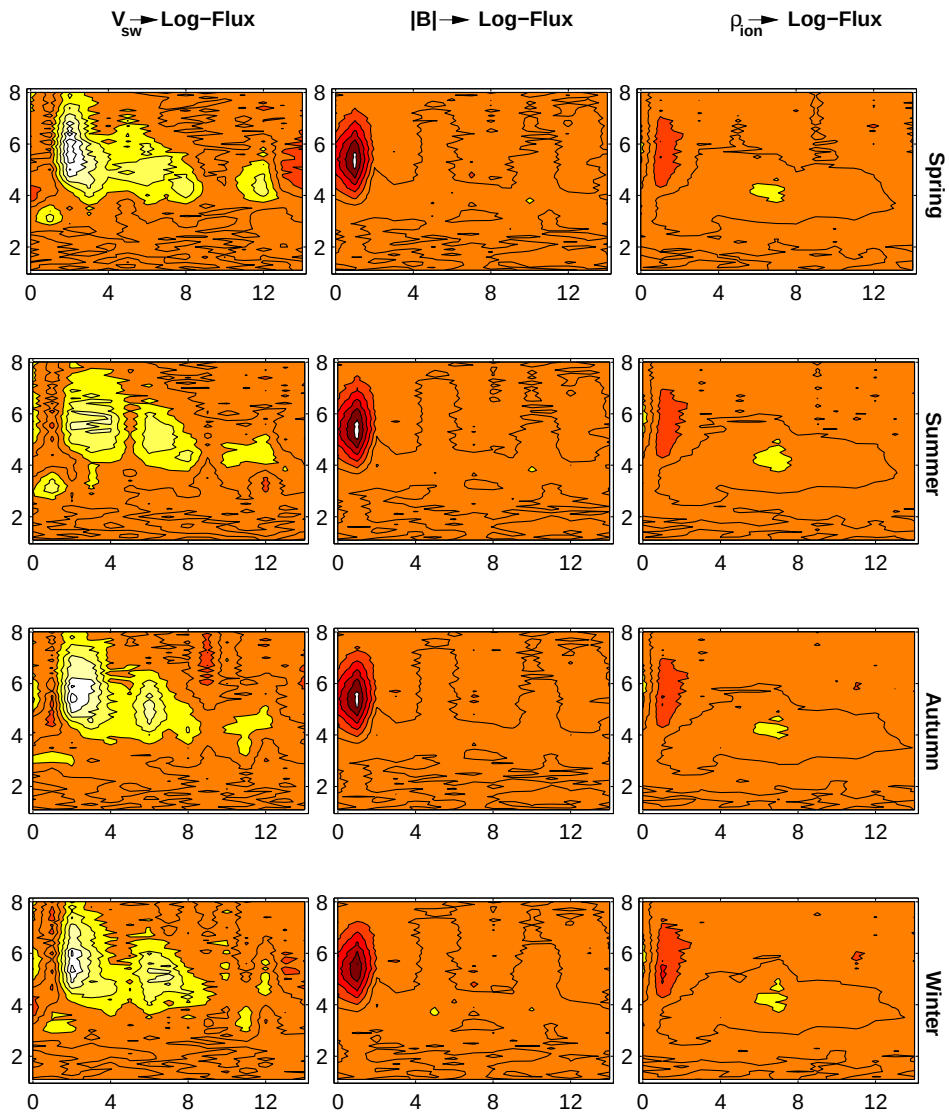


Figure 4.11: Seasonally averaged (centered on the listed solstice/equinox) multi-input FIR response functions are plotted vs. L-shell. The ordinates, abscissas, and color scales are identical to those used in Figure 4.10.

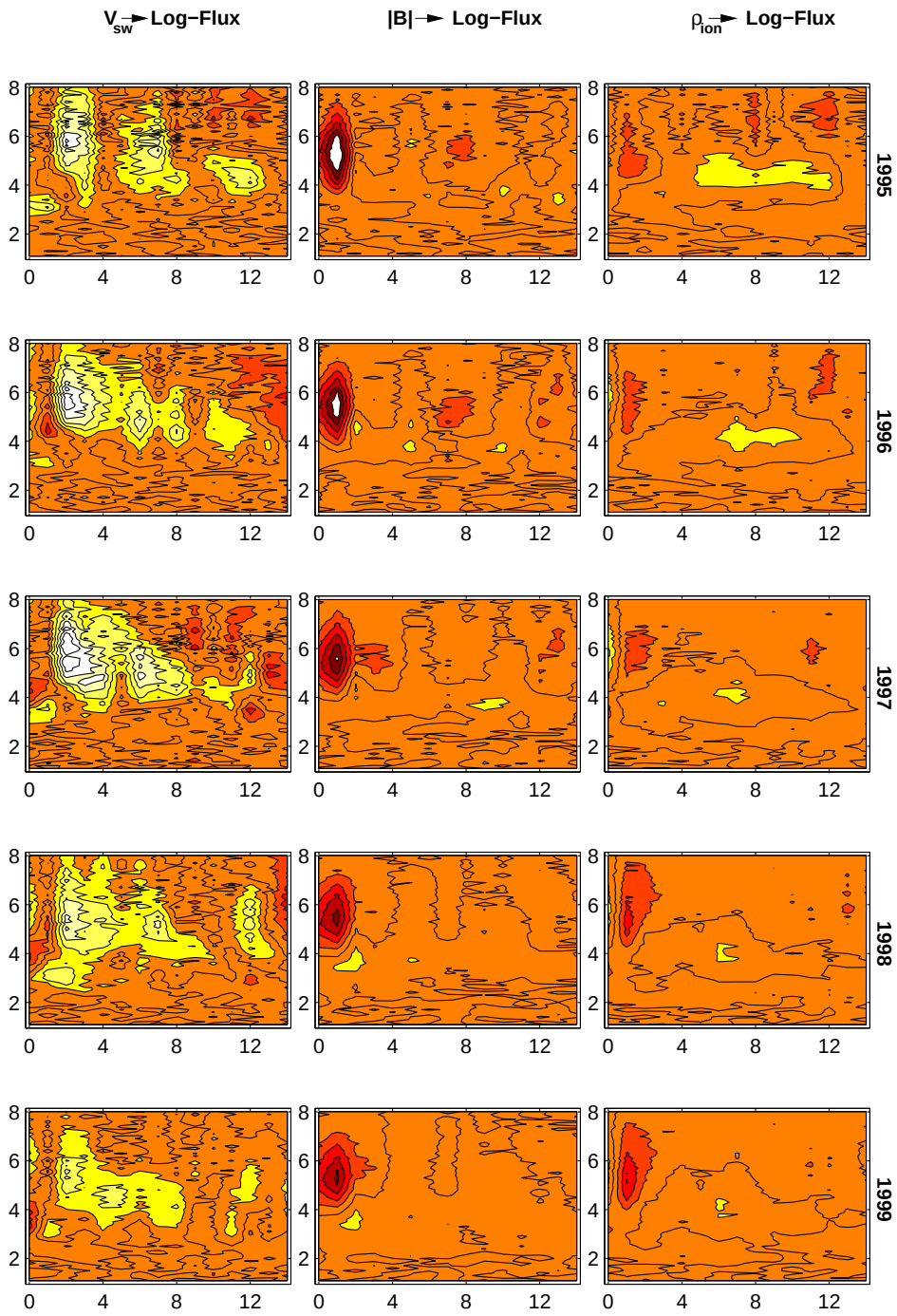


Figure 4.12: Annually averaged multi-input FIR response functions are plotted vs. L-shell. The ordinates, abscissas, and color scales are identical to those used in Figure 4.10.

has changed. This is not meant to attribute undue intelligence (or lack thereof) to the Kalman Filter, but it is simply a fact that there is only so much dynamical information available in the training data capable of exciting a dynamic response. The model that is able to adapt most quickly to non-linear changes in the system will use this up before the others have a chance to exploit it.

The semi-annual and solar-cycle variations are otherwise very similar to those already described with regard to the single-input profiles.

4.2.2 Predictions vs. Observations

In spite of the admitted issues with the adaptive multi-input response function profiles, the preferential excitation of the radiation belts by different solar wind inputs should be obvious after studying the static multi-input response functions. If not, plotting the output of each component as a prediction profile will make it very clear. Figure 4.13 shows the individual components of both stationary and adaptive multi-input predictions in its top three panels, while presenting the sum total of these three components plus the resulting prediction efficiency. It makes no sense to try to calculate PE for any of the component prediction profiles because that would assume that the other two inputs were equal to zero at the same time. However, it is clear that the broad PE peak seen at higher L-shells in the total predicted flux corresponds mostly to the $V_{sw} \rightarrow \log(\text{flux})$ subset of predictions, and that the narrower PE peak centered around L=3 corresponds mostly to the $|\mathbf{B}| \rightarrow \log(\text{flux})$ subset of predictions. The $\rho_{ion} \rightarrow \log(\text{flux})$ subset of predictions exhibits a slight peak somewhere between the other two, and serves to smooth the transition between the other two regions in the PE profile.

Each of the individual time-stationary prediction profiles maintain a constant mean vs. L-shell, and enhancements are directly related to changes in the corresponding inputs (both traits consistent with a time-stationary dynamical process). However,

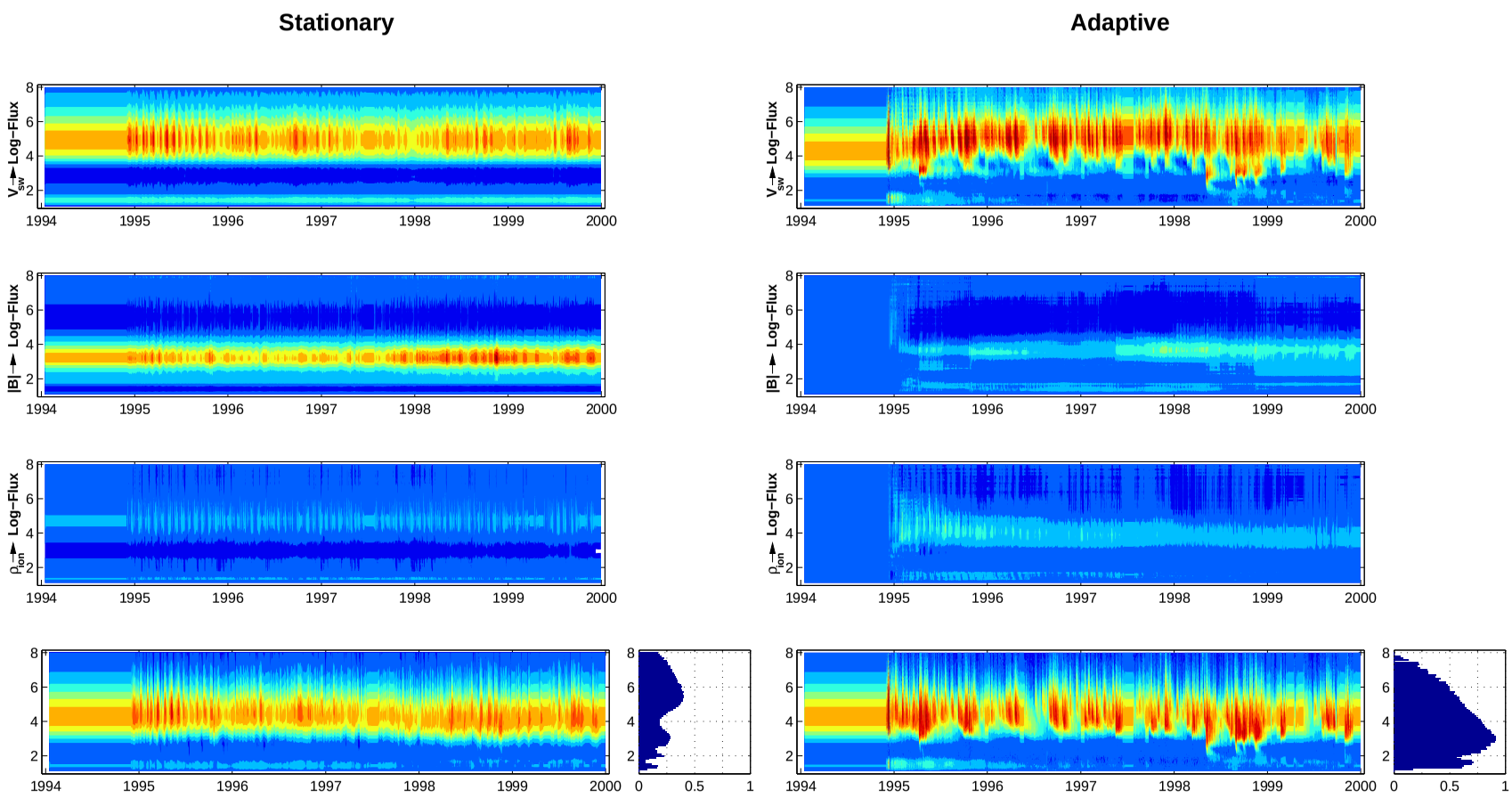


Figure 4.13: Radiation belt electron log(flux) predictions from the stationary (left) and adaptive (right) multi-input filters are plotted vs. L-shell. The sum of the top three panels provides the total predicted flux in the bottom panel, where L-shell-dependent prediction efficiencies between zero and one are provided. Horizontal contours correspond to data gaps where historical averages were substituted.

the three components summed together exhibit mildly non-stationary characteristics, resulting in a mean that varies with time and enhancements that are not so easily correlated with changes in any single input. Never-the-less, it is important to remember that this is still the product of a time-stationary linear model.

With this in mind, it is logical to assume that an adaptive system identification scheme would do even better when applied to a multi-input model structure since it would not be required to track non-stationary dynamics that were simply a consequence of unconsidered inputs. The PE profile for the summed predictions shown on the right in Figure 4.13, however, is almost identical to the PE profile for the adaptive predictions generated with only solar wind speed as an input (see Figure 4.5). The immediate explanation is simply that the V_{sw} component of the multi-input predictions comprises the vast majority of the total electron flux output, and that it makes sense that the multi-input PE profile would resemble the single-input PE profile.

Unfortunately, as alluded to in the previous section, the real explanation is not quite so simple, and is likely related to different adaptation rates. The first hint that this is the case can be found in the near-zero predictions for most of 1994 for the adaptive multi-input $|\mathbf{B}| \rightarrow \log(\text{flux})$ and $\rho_{ion} \rightarrow \log(\text{flux})$ prediction subsets. Since there are no dynamic change in the inputs for this period, there is no way to separate statistically the individual dynamics. Since the $V_{sw} \rightarrow \log(\text{flux})$ model adapts most quickly due to its larger magnitude, it will quickly converge to a solution that predicts the average electron flux values that were used to substitute for real measurements during this extended data gap. Once it converges, there is nothing left for the other components to predict, and both they and their predictions remain near zero.

Once real data, exhibiting natural variations, is incorporated into the Kalman Filter, the FIR parameters adapt to minimize prediction residuals. However, due to the slower adaptation of the $|\mathbf{B}| \rightarrow \log(\text{flux})$ and $\rho_{ion} \rightarrow \log(\text{flux})$ model parameters, the adaptive $V_{sw} \rightarrow \log(\text{flux})$ model continues to account for the lion's share of the dynamic

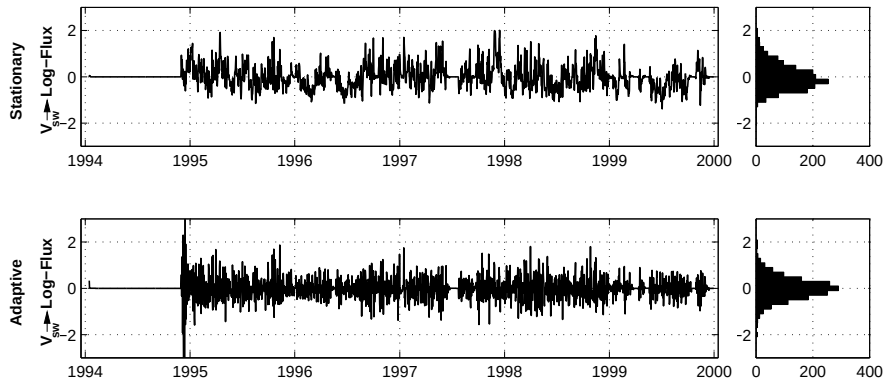


Figure 4.14: Prediction residuals for each multi-input model (stationary and adaptive) are plotted vs. time for L=6.6. The associated error distribution is portrayed as a histogram to the right, with bins defined between -2 and 2. Horizontal lines correspond to data gaps where historical averages were substituted and were not included in the histograms.

variability, even when (and where) it is not truly the most effective dynamic driver. As mentioned previously, these results were included to serve as a lesson to the reader (see Footnote 1 on page 68).

One way to overcome this adaptation rate problem is to manually adjust the covariance matrices so that the adaptation speed for each of the individual inputs results in a roughly equal share of known electron flux response during the data gap period. This is not a very elegant approach, but in truth, it falls in line with typical Kalman Filter usage in engineering applications, whether as a data assimilation or system identification tool. It is not very elegant, and will prove difficult or impossible for any system with more than two or three inputs. In addition, an approach so prone to user bias may very well lead to a biased physical interpretation. Nevertheless, it is a straight-forward and logical next step toward developing these models into something with more scientific and/or engineering utility.

Another possible remedy to the adaptation rate problem that has been suggested involves simply normalizing the inputs. This would result in parameter adjustments

that were only a function of the residual amplitude, giving similar adaptation rates for each component of the multi-input model. However, it is possible that the absolute magnitudes of the different inputs really do influence the proportion of the predicted output that should be attributed to a particular input. This is the case for time-stationary multi-input models, and there is no reason to believe that it wouldn't also hold true for adaptive models. The extended Kalman Filter, described in more detail in Chapter 5, offers an alternative to the recursive linear regression described thus far. It does not directly incorporate the system inputs into the calculation of the Kalman gain, and should therefore adjust all parameters in a consistent manner according to the most recent prediction error.

Figures 4.14, 4.15, and 4.16 are simply provided for comparison with those similar plots presented for single-input models. The only notable differences exist in the residual distributions and structure of the stationary multi-input model residuals. The error distribution is somewhat closer to normal, but still significantly skewed at L=6.6. The autocorrelated structure of the residuals is also somewhat improved over the single-input models, but still falls well outside of the 95% confidence intervals. The crosscorrelated structure of the residuals with each of the different inputs improves somewhat, although, unsurprisingly, this is simply a function of the improvement in PE at a particular L-shell.

4.3 Chapter Summary

Chapter 4 demonstrated the ability of the Kalman Filter to work with real-world data and track non-linear changes in the electron radiation belts that can be modeled as non-time-stationary linear dynamics (i.e. optimal linear FIR filter coefficients that vary with time). Impressive improvements in prediction efficiencies and other more sophisticated residual-based metrics were achieved for single-input models driven by solar wind speed (V_{sw}). Less impressive, but still significant, improvements were seen

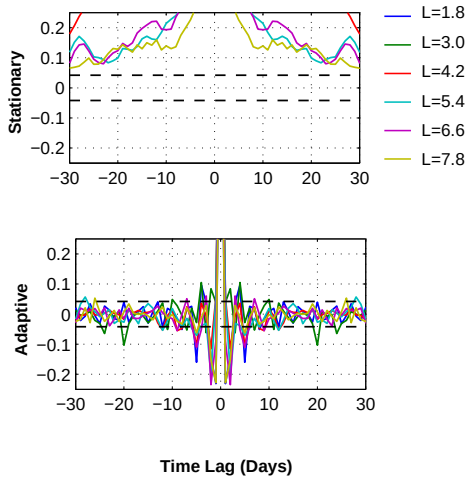


Figure 4.15: Autocorrelation functions (ACFs) of prediction residuals for stationary and adaptive multi-input models are plotted for select L-shells. The horizontal dashed lines to either side of zero demarcate the 95% confidence interval wherein correlation coefficients for lags$\neq$0 must fall for the signal to be considered "white".

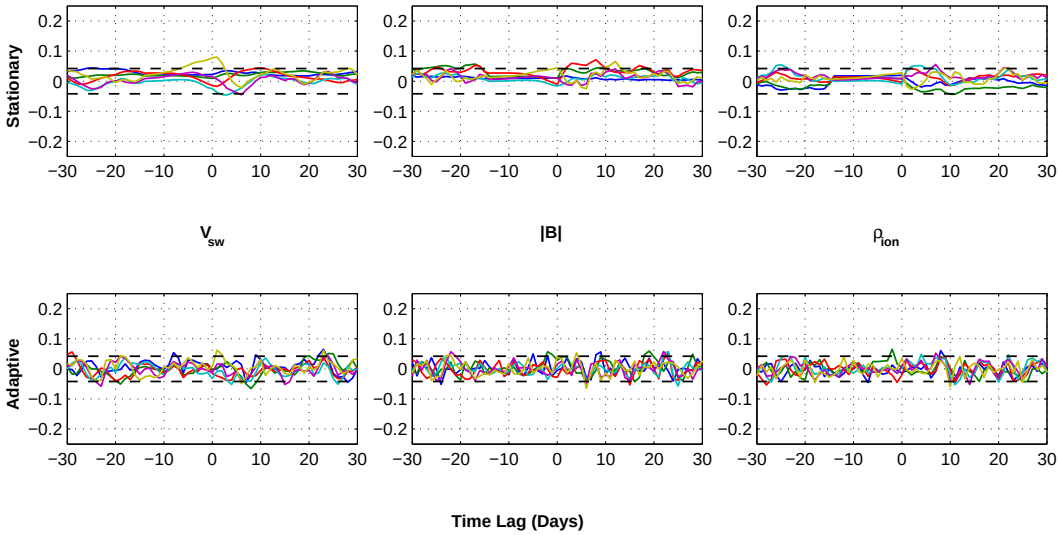


Figure 4.16: Cross-correlation functions (CCFs) of prediction residuals with different solar wind inputs for stationary and adaptive multi-input models are plotted for select L-shells. The horizontal dashed lines to either side of zero demarcate the 95% confidence interval wherein correlation coefficients for lags$\neq$0 must fall for the signals to be considered uncorrelated.

for models driven by the IMF magnitude ($|\mathbf{B}|$), and solar wind ion density (ρ_{ion}). These "less impressive" improvements are attributed to the fact that these models could not adapt as quickly as the $V_{sw} \rightarrow \log(\text{flux})$ model within the Kalman Filter framework due to their smaller dynamic variability.

An attempt was made to separate true non-linear dynamics from non-stationary changes in the electron flux that were due to the fact that the radiation belts are driven by more than a single solar wind input. Stationary multi-input FIR response functions were calculated, and plotted vs. L-shell. An unintended positive consequence was that it was shown definitively that different regions of the radiation belt are excited by different inputs preferentially, confirming recent studies that were based on single-input FIR models alone (Vassiliadis et al., 2004). Specifically, solar wind speed tends to dominate at higher altitudes ($4<L<8$), IMF magnitude tends to dominate over a narrower range of lower altitudes ($3<L<4$), and ion density dominates somewhere in between, and acts to smooth the transition between them. It was subsequently shown that small but noticeable variations in the distribution of electron log-flux vs. L-shell could indeed be reproduced with time-stationary multi-input models.

When these multi-input models were incorporated into the adaptive Kalman Filter, however, what should have been a predictable flaw in the proposed technique became apparent. The adaptation rate of the Kalman Filter for a particular model parameter is a function of the absolute magnitude and dynamic range of the inputs, in addition to the parameter covariance matrices defined a priori. Since the covariances were set to a single value and kept constant with time, the input with the largest values and, in general, exhibiting the largest dynamic range, namely the solar wind bulk speed, accounted for a disproportionate amount of the electron flux variability. Possible remedies to this problem will be discussed in more detail in Chapter 5, but the results were left as-is to serve as a lesson for the future.

Appendix B presents higher resolution versions of all the residual analysis figures

presented in this chapter, including the predictions themselves, should the reader wish to analyze the results more closely.

Chapter 5

Future Directions

The preceding chapters demonstrated the value of the Kalman Filter as a powerful adaptive system identification and space weather prediction tool when combined with FIR linear filter models. There remains much room for improvement, however. For example, the tools developed for this validation study have not been applied to optimal combinations of input and output data that might be more appropriate for an operational forecast model, or perhaps more importantly, improved physical understanding. In addition, the data that were used had a very course time-resolution. While current operational radiation belt models do no better, the ability to predict changes in radiation belt electrons at a higher (i.e. sub-daily) cadence, as well as a more extensive spatial distribution, will represent a tremendous step forward in space weather forecasting. Finally, it is clear from the residual analyses presented in Chapter 4 that, at least for certain regions of the radiation belts, some of the dynamics still remain to be accounted for. A number of variations on the general linear filter exist that should be explored before proceeding to use more complicated non-linear model structures.

5.1 Data

The most important considerations for any empirical radiation belt model must relate to data. How trustworthy is the data? Are the different data dimensionally consistent? Is the data appropriate for on-line operational forecast models or physical in-

terpretation? Admittedly, the prototypical nature of the studies presented in this thesis were focused primarily on validating the numerical techniques, and did not consider any of these questions adequately. Here, based in part on lessons learned during the studies discussed previously, we suggest several alternative approaches to data selection that will provide additional engineering and scientific utility to the models already presented.

5.1.1 Solar Wind

The OMNI data used in this study will not suffice for an operational model because they are not available in real-time. In addition, the OMNI data set fails to provide statistical characterizations for the different parameters that are required for any reliable optimization algorithms. We now consider possible alternative sources currently available, or expected in the near future.

The types of solar wind data that would be most useful to an operational forecast model for radiation belt electrons are not necessarily the same types of data used for this study (i.e. bulk speed, magnetic field intensity, and ion density). One important consideration is that, if higher cadence predictions were desired, the variations in the vector components of the interplanetary magnetic field (IMF) would likely become more important, since the assumption that oscillations average out to zero at daily timescales would no longer be valid. Another unstated assumption was that simple ions comprised the bulk of mass density, but helium ions constitute nearly 15% of the mass in the solar wind and can play a significant role in determining the solar wind dynamic pressure. A better separation of these two values would certainly prove useful at higher time resolutions.

If the intercalibration of various measurements can be trusted, it may prove beneficial to use more physical combinations of the solar wind parameters. These might be determined by expert analysis (i.e., experts determine which combinations are likely

to be most relevant, as in Li et al. (2001b). A more formal dimensional analysis may also be conducted to determine all of the most relevant combinations of parameters (e.g., Perreault and Akasofu, 1978; Vasyliunas et al., 1982; Gonzalez, 1990). Finally, a systematic survey using all ”reasonable” combinations of parameters, linear and non-linear, would not be beyond consideration if the total number of inputs were kept to a minimum (i.e. ≤ 3), an approach currently being explored by Vassiliadis and Fung at the Goddard Space Flight Center.

One of the primary goals of the OMNI project is to combine all solar wind data into a single consistent set amenable to scientific examination and interpretation. One might question, then, why the parameters included in the OMNI data set are often dimensionally inconsistent. Indeed, just the limited set of inputs taken from the OMNI data set for the studies presented in this thesis have units of km/s (solar wind speed), nT (nano-Tesla, magnetic field measurements), and cm^{-3} (solar wind ion density). Scientific interpretation of the solar wind input data, and more importantly, the input data's dynamic relationship to magnetospheric parameters, including relativistic electron flux, requires dimensional consistency. While it should be straight-forward for a user to convert the given data to a more consistent set of units, the fact that many of these parameters are actually derived quantities based on other parameters with unknown uncertainties makes this a somewhat dubious endeavor.

Ideally the statistical characterization of these input parameters would involve combining pre- and post-launch instrument calibration studies and rigorously derived confidence intervals for detector performance, thus allowing a robust statistical normalization between the different data sets. In reality, the statistical uncertainties for many solar and/or space plasma detectors are not well established at all, either due to inadequate ground-based calibrations, or changes in detector efficiencies after launch. Future missions should work harder to remedy this shortcoming, and provide the relevant

A mission currently providing real or near-real-time solar wind measurements

is the Advanced Composition Explorer (ACE) satellite. ACE is stationed well outside the terrestrial magnetosphere, in the solar wind, approximately 1.5 million kilometers ($\sim$220 RE) upstream of the Earth. ACE was primarily designed for scientific purposes, but it provides nearly continuous real-time measurements of relevant solar wind parameters, many of which have been incorporated in existing space weather models (Li et al., 2001b; SEC, 2003). While the end of its likely operational life is approaching, both the scientific and engineering communities have come to acknowledge its vital role in scientific magnetospheric studies and space weather applications. It is probable that a similar mission will follow.

5.1.2 Radiation Belts

SAMPEX electron flux data comes from a single source, and is immune to many of the calibration problems just described. The issue of dimensional consistency is also moot, since differences in units are automatically caught by the linear filter coefficients. On the other hand, the electron flux data used in the studies presented in this thesis were processed significantly so that observations could be binned and ordered by geomagnetic L-shell. There is substantial uncertainty in this process, especially for measurements made at relatively high latitudes where the static magnetic field model used for L-shell mapping is less likely to be valid at any particular time. This uncertainty needs to be better understood, if not actually corrected, and provided alongside the actual data if this data is to be incorporated into a true operational space weather forecast model.

All of this may be a moot point, however, since SAMPEX, and indeed most satellites with radiation belt sensors, do not make their data publicly or otherwise easily available in a real-time manner. In fact the only easily accessed real-time radiation belt observations currently come from the National Oceanic and Atmospheric Administration (NOAA) geostationary GOES satellites, which cover a scientifically interesting,

but operationally limited region of the near-Earth radiation environment. A number of other operational satellites actually make in situ energetic electron measurements and provide, or can be easily modified to provide, their data in real or near-real-time. These include the Los Alamos National Laboratory (LANL) geostationary satellites, which provide extensive local-time coverage of the radiation belts at geostationary altitudes; a growing number of the U.S. Department of Defense Global Positioning System (GPS) satellites, which also offer extensive local-time coverage in addition to passing through a range of L-shells from 4.2 at the equator, out beyond 7.0 for the most highly-inclined orbits; and finally the NOAA Polar-orbiting Operational Environmental Satellite (POES) satellite, which carries a set of Medium Energy Proton and Electron Detector (MEPED) instruments that measure electrons in three different energy channels (30, 100, and 300 keV, respectively[1]) across the same range of geomagnetic L-shells as SAMPEX. Various efforts are currently under way to make much of this data publicly available in real time (J. Green, personal communication), but cost constraints and a significant amount of institutional inertia and antiquated proprietary thinking about experimental data must be overcome first.

Finally, it is worth mentioning the National Polar-orbiting Operational Environmental Satellite System (NPOESS). First proposed in 1994, NPOESS is a plan for a multi-purpose, multi-agency, mutli-satellite system whose POES-like radiation belt sensors will fly through all relevant L-shells in a manner similar to SAMPEX and POES. It is unlikely to see its first satellite launch before 2009, but it will undoubtedly prove itself invaluable in the burgeoning field space weather forecasting (NPOESS, 2001).

[1] These are not quite the "killer electrons" that are of most concern from a space weather perspective, but they can help determine the shape of the electron energy spectrum, which in turn can be used to extrapolate measured fluxes to higher energies.

5.2 Improved Model Structures

The FIR linear filter that was used in this study is an extremely, perhaps overly, simple model structure for describing what is ultimately a very complicated dynamical interaction between the solar wind and the electron radiation belts. It was chosen so that the application of the Kalman Filter for adaptive linear system identification could be carefully studied without focusing unnecessarily on the actual underlying prediction model. There are any number of more sophisticated dynamical model structures that should provide not only better predictions, but allow better temporal and spatial resolution, as well as help separate the directly solar wind-driven dynamics from the internal magnetospheric dynamics that all contribute to the acceleration and loss of electrons in the Van Allen radiation belts.

5.2.1 Dynamic Feedback and Noise Filters

One significant problem encountered in this study was the tremendous amount of data that was necessarily dropped, and replaced with time-averages, simply due to the fact that each predicted output required 15 neighboring inputs for the filter to operate on. As already explained in Chapter 2, the solar wind input for 1994 was not nearly as bad as the training data used to produce the models and predictions in Chapter 4 would suggest. It was simply that it was rare for more than 15 days to pass without a data gap, thereby requiring that all the data was dropped for the identification procedure.

The use of so-called output error (OE) models that actually feedback on their own output (see Appendix C), requires fewer lagged coefficients in order to produce an infinite impulse response (IIR) that resembles the FIR functions used in the current study. This would require some sort of recursive non-linear parameter estimation algorithm based on a gradient descent or Newton-type minimization. It will be shown shortly how the Kalman Filter can be modified to serve as a non-linear estimator for

linear models with dynamical feedback

A second problem encountered in the current study was the low time resolution of the data that was used for training the FIR models and making predictions. As explained in greater detail in Chapter 2, 24-hour averages were used in order to avoid complications that arose from non-input-driven auto-correlations in the training data, a result of the well-known longitudinal and local time variations often observed in electron fluxes, especially in the outer radiation belt. These variations are not dynamically related to the solar wind (although, in some cases, they result from the relatively constant compression of the Earth's magnetic field by the solar wind), and tend to corrupt the parameter estimates of FIR (and ostensibly OE) models by introducing non-physical oscillations and other correlated structure to the response functions.

Persistent correlated signals can be reproduced by filtering white noise. The output can be thought of as colored noise, but this is no reason to assume that it is unusable. On the contrary, even though models that possess both a deterministic and a stochastic component (see Appendix C) may or may not provide better predictions than the purely deterministic models used thus far, by accounting for persistent correlated signals, they allow better separation of the externally driven and recurrent internal dynamics that both dictate radiation belt variability at sub-daily time resolution. This approach has tremendous scientific potential, since even after decades of study, the fundamental dynamics of the radiation belts is still poorly understood at all but the longest time-scales.

Some initial work along these lines, using so-called autoregressive filters with eXogenous (ARX) filters, demonstrates the utility of this approach. More detail is provided in Appendix C, but this type of model essentially combines a deterministic FIR filter to model the driven dynamics, and a stochastic autoregressive (AR) filter that reproduces the persistent recurrent structure in the signal. If the model is optimized as a 1-step predictor, Equation 5.1 describes how the best estimate of the output $\hat{y}(t)$ is generated by filtering both input and output up to and including the most recently

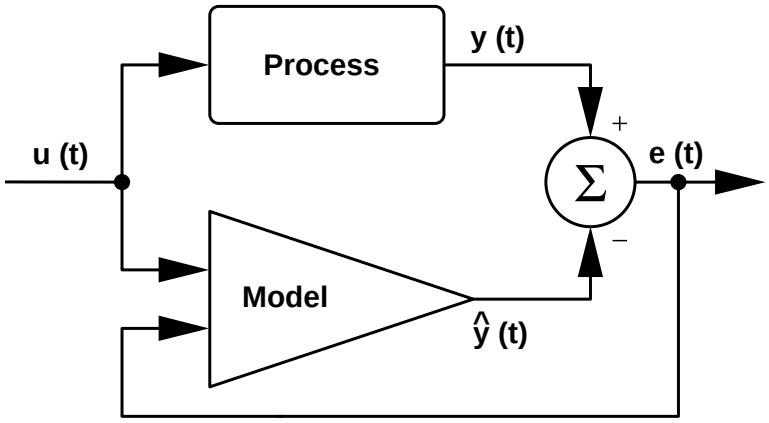


Figure 5.1: A generalized single input, single output system identification configuration for a time-stationary model with non-deterministic feedback. Note that residuals are explicitly incorporated into the model, requiring system output up to and including the most recent observation.

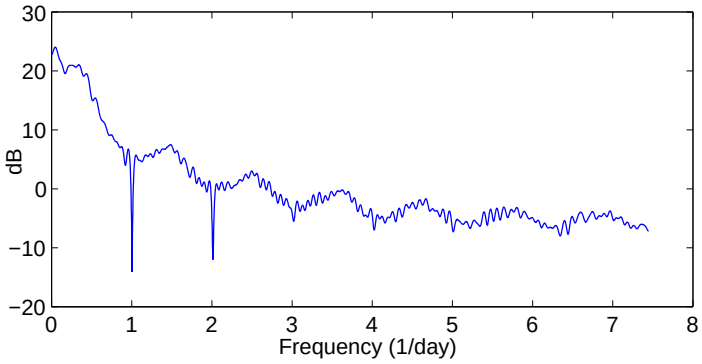


Figure 5.2: The frequency response of a 1-step AR prediction filter for SAMPEX electrons at L=6.5. This weak low-pass filter has a strong band-stop at 1 per day and associated harmonics.

observed output.

$$\hat{y}(t|t-1) = \sum_{\tau=0}^{n} b_\tau u(t-\tau) - \sum_{\tau=1}^{n} a_\tau y(t-\tau) \tag{5.1}$$

If an AR filter determined using sub-daily resolution geostationary electron flux were identified and applied to a broad-band white-noise sequence, only the power shown in Figure 5.2 would remain. A diurnal signal often noted in SAMPEX measurements, and usually attributed to the passage of the South Atlantic Anomaly (SAA) under the satellite's orbital plane, is obvious at a frequency of once per day. The "band-stop" seen at two per day is most likely a harmonic due to the fact that the diurnal variation is typically not very sinusoidal, but more of a spike. Frequency responses at lower L-shells exhibit more pronounced harmonics at higher integer frequencies, probably because these are closer to the heart of the SAA.

The other term in Equation 5.1 model is an FIR filter similar to those discussed in Chapter 4. Figure 5.3 compares the daily-averaged response functions with orbit-averaged response functions for both individual single-input single-output (SISO) filters and each of the components of a multi-input single-output (MISO) filter. It is interesting to note that there is little difference between the sub-daily and daily resolution response functions. The one exception is in the $V_{sw} \rightarrow \log(\text{flux})$ "dropout" for the SISO model. It appears as though the loss of $\log(\text{flux})$ is mostly subsequent to the solar wind impulse, indicating causality. While this is not surprising, this seems to be the first time it has been illustrated so quantitatively.

5.2.2 3-D Cartesian Coordinates

One assumption made repeatedly throughout this study is that a one-dimensional "spatial" coordinate system (L-shell) based on uncertain coordinate transformations is somehow best for our modeling objectives. While such an approach has obvious scientific merit, it is of little use to most satellite designers and mission operators, who

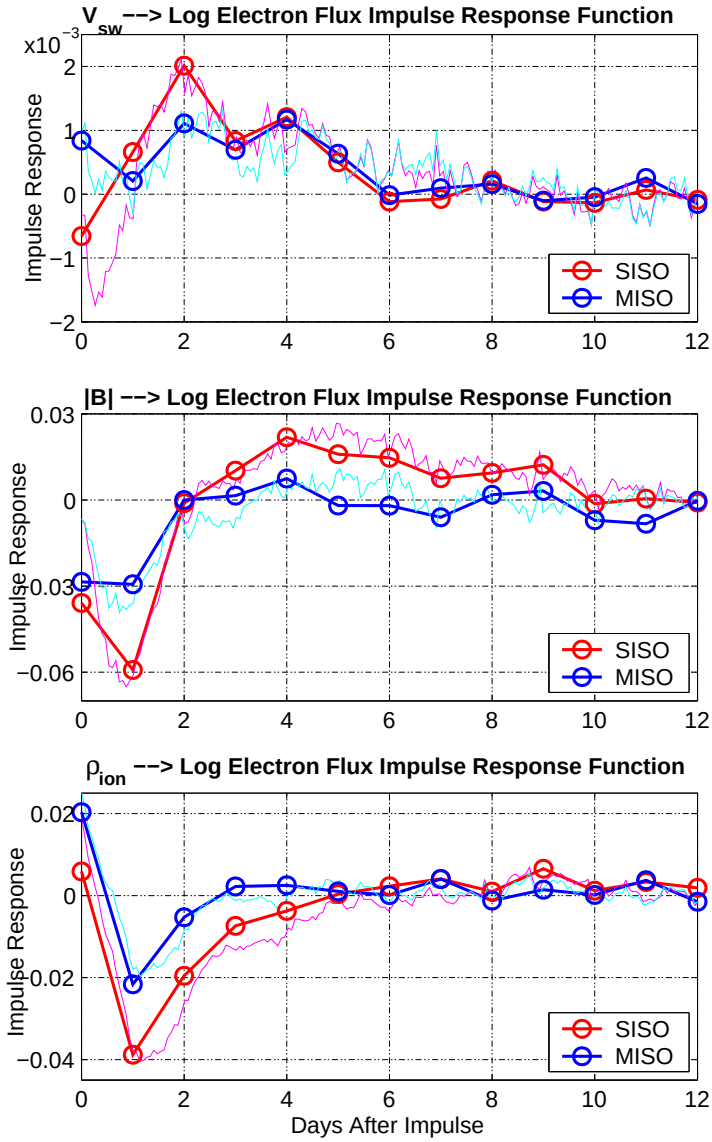


Figure 5.3: A comparison of single input, single output (SISO) impulse responses vs. multiple input, single output (MISO) impulse responses at daily (red/blue) and SAMPEX orbit-averaged (magenta/cyan) time resolutions. These are for 2-6MeV electrons at L=6.5.

will typically require data in more standard 3-D Cartesian coordinates so that virtual satellites can be inserted to explore the local radiation environment.

From an operational perspective, it may be better to forgo a certain level of physical understanding of the system dynamics, and start over with a more geographically (as opposed to geomagnetically) consistent coordinate system that does not require the expertise of a magnetospheric physicist to understand. Finite volumes would be defined, each with its own linear prediction filter that would be trained recursively as various radiation sensors flew through the volume. There would probably be a number of different problems stemming from oscillations that are a result of the physically inappropriate coordinate system, but it was just shown that this can be largely accounted for by using more sophisticated model structures with colored noise filters in addition to deterministic input-output models.

Finally, both deterministic and stochastic feedback coupling between each of the finite volumes should be built into the framework by designing true multi-input mutli-output (MIMO) models. Not only will this help maintain reasonable prediction quality in regions with no radiation sensors at that moment (since they would be coupled to at least some volumes that did contain a radiation sensor), but it now becomes possible to actually analyze the internal dynamical coupling between different regions of the radiation belts. The ability to separate those regions with a directly driven response to the solar wind changes, and those whose changes result from to their proximity to a responsive region, will undoubtedly have a significant impact on scientific understanding of the terrestrial magnetosphere.

Experience gained in the development of grids for sophisticated MHD and/or particle models will prove invaluable to this endeavor, and of course the accurate intercalibration of different radiation belt sensors is essential. Despite these difficulties, however, this kind of empirical global model is extremely scalable and parallelizable, allowing one to start with fairly large finite volumes for preliminary development and

validation, then quickly scale up as computer speed and memory continue to improve .

5.2.3 State Space Models and the Extended Kalman Filter

The improvements suggested thus far are far from trivial, and would in fact be virtually impossible if they were implemented in the same Kalman Filter framework that was presented in previous chapters. To start with, there is nothing in the framework capable of keeping track of the actual state of the system (remember, the FIR coefficient vector θ is not a true state, simply a vector of filter coefficients). There is no need for this with the ARX structure because, like the FIR model, it is linear in its parameters. However, the ability to keep track of a system state is critical for determining the feedback coefficients used in even more advanced linear model structures like those described in Appendix C.

It is possible to post-filter the most recent errors and regression vector with the latest $\hat{\theta}$ in an effort to follow the steepest gradient toward the minimum in a non-quadratic loss function. This is more or less the recursive equivalent to what is done in non-recursive iterative optimization techniques like Levenberg-Marquardt non-linear regression. One major drawback to this approach for our purposes is that the regression matrix for multi-output models with feedback can become extremely burdensome as all possible cross-couplings must be parameterized. It turns out that a more flexible linear model structure exists that is based on a true state-space representation of the system dynamics.

$$\begin{aligned} x(t+1) &= \mathbf{A}x(t) + \mathbf{B}u(t) + v(t) \\ y(t) &= \mathbf{C}x(t) + \mathbf{D}u(t) + e(t) \end{aligned} \tag{5.2}$$

It is important to note that matrices **A**, **B**, **C**, and **D** are not the same variables as the filters used previously. **A** is a square matrix with dimension (n,n) representing the nth-order internal dynamics of the system. **B** is a matrix with dimension (n,m), corresponding to the dimension of **A** and the number of inputs. **C** is a matrix with

dimensions (p,n), corresponding to the number of outputs and the dimension of $\mathbf{A}$, and is used to map the state to an actual observable for calculating prediction errors. Finally $\mathbf{D}$ is a matrix of dimension (p,m), and allows direct filtering of input without it being affected by the system dynamics. The random variables $v(t)$ and $e(t)$ are the same as before, except that now they represent actual uncertainty in the modeled state and observations, respectively, not just the model parameters θ or regression vector ϕ components. Note that n only specifies the order of the state, not necessarily the number of time-lags used in the model. The two are equivalent only for single-output systems (i.e. dimension of $\mathbf{C}$ is $(1,n)$).

The connection between this and the previous state-space representation used with the Kalman Filter as a simple recursive linear least-squares estimator is obvious if $x = \theta$, $\mathbf{A} = \mathbf{I}$, $\mathbf{B} = \mathbf{0}$, $\mathbf{D} = \mathbf{0}$, and $\mathbf{C} = \phi$ (Ljung and Söderström, 1983; Ljung and Gunnarsson, 1990). This new representation is constrained to work with causal dynamical systems, but the additional flexibility offered by allowing these state-space model parameters to vary and provide optimal predictions is clear. For the moment, assume $\mathbf{A}$, $\mathbf{B}$, $\mathbf{C}$, and $\mathbf{D}$ are known and constant. Equation 5.3 then represents a form of data assimilation that optimizes the state $x(t)$ using linear models as a baseline.

$$\begin{aligned}
\hat{x}(t+1) &= \mathbf{A}x(t) + \mathbf{B}u(t) + \mathbf{K}(t)\left(y(t) - \mathbf{C}x(t)\right) \\
\mathbf{K}(t) &= \left(\mathbf{C}\mathbf{P}(t-1)\mathbf{C}^{\mathsf{T}} + \mathbf{Q}^{e}\right)^{-1}\mathbf{A}\mathbf{P}(t)\mathbf{C}^{\mathsf{T}} \\
\mathbf{P}(t+1) &= \left(\mathbf{I} - \mathbf{K}(t)\mathbf{C}(t)\right)\mathbf{P}(t) + \mathbf{Q}^{v}
\end{aligned} \tag{5.3}$$

Any number of techniques might be used to determine these state-space matrices, but as we demonstrated in previous chapters, there is little difference between estimating a true state, and estimating linear model parameters. The state vector x can be designed to include unknown model parameters. Redefining θ to include all adjustable parameters of the state-space matrices gives:

$$\theta = [a_1 \cdots a_n b_1 \cdots b_n c_1 \cdots c_n d_1 \cdots d_n]^{\mathsf{T}} \tag{5.4}$$

The extended state vector is then given by:

$$z(t) = \begin{bmatrix} x(t) \\ \theta(t) \end{bmatrix} \tag{5.5}$$

Rewriting Equation 5.2 to use this new state vector gives:

$$\begin{aligned} z(t+1) &= \begin{bmatrix} \mathbf{A}(\theta)x(t) + \mathbf{B}(\theta)u(t) \\ \theta(t) \end{bmatrix} + \begin{bmatrix} v_x(t) \\ v_\theta(t) \end{bmatrix} \\ y(t) &= \mathbf{C}(\theta)x(t) + \mathbf{D}(\theta)u(t) + e(t) \end{aligned} \tag{5.6}$$

When incorporated into the Kalman Filter this constitutes a non-linear estimation algorithm whose corresponding update equations are significantly more complex than Equations 5.3. Ljung (1979) describes in detail this extended Kalman Filter (EKF) identification algorithm, including a very thorough analysis of the algorithm's convergence properties. He confirmed a number of contemporary "rumors" about the convergence properties of the EKF, including 1) in general, the EKF is not guaranteed to converge on a fixed set of parameters; and 2) even when it appears to converge, these estimates are likely biased. Finally, Ljung offers an alternate formulation of the state-space update equations, based on the innovations form state-space model, that does guarantee convergence.

$$\begin{aligned} x(t+1) &= \mathbf{A}x(t) + \mathbf{B}u(t) + \mathbf{K}\varepsilon(t) \\ y(t) &= \mathbf{C}x(t) + \mathbf{D}u(t) + \varepsilon(t) \end{aligned} \tag{5.7}$$

Here, $\varepsilon(t)$ represents the model residuals, or innovations, and has an associated covariance matrix Λ. The one drawback to this approach is that it does not allow separate characterization of process and measurement noise via $\mathbf{Q}^v$ or $\mathbf{Q}^e$, however the only case when this is desirable is if reliable a priori information about the noise structures is available, a rare event in real-world applications. $\mathbf{K}$ is required to be independent of θ (proven in excruciating detail in Ljung (1979)). This criterion is met when $\mathbf{K}$ is parameterized in a similar manner to the other state-space matrices, and is assumed

to be constant in the time-stationary sense (i.e. $\mathbf{K} \rightarrow$ constant for $t \rightarrow \infty$). It can be shown, for single-output systems, that this very general model structure can be made equivalent to all the transfer function-based models described in Appendix C (see, for example: Guidorzi, 1975, 1981; Ljung, 1999), although it is unclear why such unwieldy dynamic model structures would be desirable once the more flexible innovations-based EKF framework has been established.

5.3 Chapter Summary

Section 5.1 discussed possible real, or near-real time alternatives to the OMNI solar wind inputs and SAMPEX electron flux observations described in Chapter 2 that might be used in an online, operational space weather forecasting tool. Oddly, it is the upstream solar wind monitor ACE that is currently set up for real time data transmission, while the only publicly available radiation belt data is from NOAA and LANL geostationary satellites. Despite its limited energy range, current efforts to make available the POES radiation belt measurements in real time promise to go far in helping fill in the spatial data gaps. It is expected that NPOESS, a proposed constellation of low-altitude, polar orbiting environemental satellites and the follow-on to the POES program, will be a primary source of radiation belt observations within the next five years.

Section 5.2 goes on to discuss a number of possible enhancements to the adaptive linear prediction filters presented in earlier chapters that can take advantage of the (hopefully) soon-to-be-made-available real time data. The first suggested enhancement involves the use of more sophisticated and flexible linear model structures that account for both deterministic and stochastic aspects of the radiation belt dynamical system. One of the most straight-forward implementations of such a system is already in place at the NOAA Space Environment Center, and was described briefly in Chapter 1's discussion of data assimilation in current electron flux prediction models.

In anticipation of a substantially increased amount of radiation belt data in the near future, it was also proposed that a more intuitive (for non magnetospheric specialists) coordinate system be set up, and a three dimensional grid be established with a coupled linear filter predictor associated with each finite volume. Establishing this kind of spatially-local linear filter model requires the more sophisticated model structures just described if the predictions (and possibly their residuals) are to be coupled in a mathematically consistent manner to provide the best possible radiation belt forecasts.

Finally, neither of the forementioned modifications to model structure will be easy, but as one considers the identification problem, it is clear that the computational complexity will begin to approach, if not surpass, that of modern physics-based models if a better framework is not established up front. The innovations form of the standard state-space model can be used to emulate FIR, ARX, and indeed any of the standard polynomial-based models described in Appendix C in a more compact structure. In addition, it is perfectly suited for integration with the Extended Kalman Filter (EKF), providing a mechanism for both adaptive system identification and real time state estimation, while allowing the user easy access to the relevant state variables for validation and monitoring purposes.

Chapter 6

Summary and Conclusions

Chapter 1 provided the background material necessary to motivate the work described in later chapters. This included an overview of the solar and geophysical systems most pertinent to space weather forecasting, with an obvious emphasis on the electron radiation belts. A brief survey of modern radiation belt models was also provided. The overly-brief section on data assimilation efforts could have been extended into a thesis in its own right, where every one of the models discussed previously could have benefited by being integrated into a Kalman Filter framework similar to the one used for the current thesis.

Chapter 2 gave a short description of the data used to train and validate the empirical FIR prediction filters presented later. NASA's OMNI database provided six years, 1994-1999, of high-quality solar wind measurements; specifically the solar wind speed (V_{sw}), the IMF intensity ($|\mathbf{B}|$), and the ion density (ρ_{ion}). Relativistic electron flux observations were taken from the SAMPEX satellite for the same period. The techniques used for synchronizing the two data sets at a daily cadence, as well as replacing missing data in a manner that was statistically benign (at least for use with the simple linear filters used in this study), were also explained.

The remaining three chapters each have their own more detailed summaries. In brief, Chapter 3 described various methods used to determine optimal coefficients for FIR response functions based on sets of input and output data. Particular attention was

paid to adaptive system identification algorithms like the recursive least-squares (RLS) with exponential forgetting, and ultimately, the Kalman Filter with process noise. A validation study using very simple time-stationary and non-stationary dynamic systems demonstrated conclusively the ability of the Kalman Filter to adjust FIR filter coefficients to track noise-corrupted non-stationary signals through time.

Chapter 4 presented spatial (L-shell) profiles of radiation belt electron responses to changes in the different solar wind parameters described above. Time-stationary response function profiles were presented for comparison with non-stationary profiles derived from the adaptive Kalman Filter-based identification algorithm. Connections are drawn between the work presented here, and previous/continuing work by Vassiliadis and colleagues concerning the time-variability of linear prediction filters used to measure effectiveness of various solar wind and geomagnetic parameters at accelerating electrons in the radiation belts. Simultaneous multi-input filters were also calculated, and their time-variations, as determined via the Kalman Filter adaptive identification algorithm, were analyzed. What should have been a predictable problem concerning the tuning of the Kalman Filter's effective memory became readily apparent while determining the multi-input response function profiles. Specifically, using process noise to define the memory in a somewhat ad hoc manner results in adaptation rates that differ depending on the absolute magnitude and dynamic variability of the respective inputs. This resulted in most of the dynamic response of the radiation belts being attributed to solar wind speed effects, even though it is clear from the time-stationary linear filters and predictions that the other inputs play very significant roles in different regions of the magnetosphere. We chose to present these results as a lesson to be remembered anyway, but are confident that process noise estimates that are specifically tuned to individual inputs will result in proper adaptation rates, and lead to interesting new insights that cannot be gained from single-input filters alone.

Finally, Chapter 5 contemplates what changes to the methods and algorithms

used throughout this thesis might be necessary in order to design and implement an online operational radiation belt forecast model. It turns out that access to real time solar wind data generated by the ACE satellite is far easier than acquiring real time radiation belt data from anywhere but geostationary platforms like NOAA's GOES and LANL's geostationary satellites. SAMPEX data is unlikely to be available in real time anytime soon, and even if efforts to make NOAA's polar orbiting POES radiation belt data available in real time are successful, the energy range of the instruments may not be sufficient for some of the most pertinent space weather applications. Enhanced linear model structures are suggested that take advantage of model feedback, as well as correlated "noise" in the observations, accounting for both directly driven radiation belt dynamics and (mathematically) stochastic dynamical processes. The use of a more intuitive (from an operational perspective) three-dimensional coordinate system that can be divided into finite volumes, each with their own coupled linear model, is proposed to take advantage of anticipated increases in the amount and quality of radiation belt data. Finally, a modified state-space model is described that can take advantage of innovations (i.e., uncorrected prediction residuals) to emulate any of the classical polynomial-based linear model structures in a more compact form. This innovations state-space model can be easily integrated into the Extended Kalman Filter (EKF) for both parameter identification and state estimation purposes.

Even with the use of the EKF, the culmination of the improvements in model structures described in Chapter 5 will prove to be significantly more computationally complex than the FIR model structures used in the current study. The question might fairly be asked why would one put this sort of effort into developing an empirical model if the initial motivation was to reduce the computational expense associated with physics-based models. The first and most obvious response is that, for now and for some time to come, empirical models offer the best forecasts, at least in the short-term, even if their scientific utility is not universally recognized. This however leads

to a second, and perhaps more controversial answer. That is, as suggested in Chapter 3, models based on only a partial understanding of the physics governing the system will never be able to simulate anything else, resulting in what has been referred to previously in this thesis as a “philosophical bias”. While they can never directly prove existing scientific theory, well-understood and well-designed empirical models will illustrate shortcomings in current theory and inspire further investigation into the more fundamental physics governing the radiation belt system dynamics.

Bibliography

S.-I. Akasofu. Interplanetary energy flux associated with magnetospheric substorms. Planet. Space. Sci., 27:425–431, 1979.

S.-I. Akasofu. Energy coupling between the solar wind and the magnetosphere. Space Science Reviews, 28:121–190, 1981.

J.H. Allen and D.C. Wilkinson. Solar-terrestrial activity affecting systems in space and on earth. In Solar-Terrestrial Predictions - IV, pages 75–107, 1992.

R. Arnoldy. Signature in the interplanetary medium for substorms. Journal of Geophysical Research, 76(22):5189–5201, 1971.

D.N. Baker. Effects of the sun on the Earth's environment. Journal of Atmospheric and Solar-Terrestrial Physics, 62:1669–1681, 2000.

D.N. Baker. Satellite anomalies due to space storms: The effects of space weather on spacecraft systems and subsystems. In I.A. Daglis, editor, Space Storms and Space Weather Hazards, pages 285–311, Netherlands, 2001. Kluwer Academic Publishers.

D.N. Baker. How to cope with space weather. Science, 297:1486–1487, 2002.

D.N. Baker, J.B. Blake, L.B. Callis, J.R. Cummings, D. Hovestadt, S.G. Kanekal, B. Klecker, R.A. Mewaldt, and R.D. Zwicki. Relativistic electron acceleration and decay time scales in the inner and outer radiation belts: SAMPEX. Geophysical Research Letters, 21(6):409–412, 1994.

D.N. Baker, J.B. Blake, R.W. Klebesadel, and P.R. Higbie. Highly relativistic electrons in the earth's outer magnetosphere 1. lifetimes and temporal history 1979-1984. Journal of Geophysical Research, 91(A4):4265–4276, 1986.

D.N. Baker, S.G. Kanekal, J.B. Blake, and T.I. Pulkkinen. The global efficiency of relativistic electron production in the earth's magnetosphere. Journal of Geophysical Research, 106(A9):19,169–19,178, 2001.

D.N. Baker, S.G. Kanekal, T.I. Pulkkinen, and J.B. Blake. Equinoctial and solstitial averages of magnetospheric relativistic electrons: A strong semiannual modulation. Geophysical Research Letters, 26(20):3193–3196, 1999a.

D.N. Baker, A.J. Klimas, R.L. McPherron, and J. Buchner. The evolution from weak to strong geomagnetic activity: An interpretation in terms of deterministic chaos. Geophysical Research Letters, 17(1):41–44, 1990a.

D.N. Baker, A.J. Klimas, D. Vassiliadis, T.I. Pulkkinen, and R.L. McPherron. Reexamination of driven and unloading aspects of magnetospheric substorms. Journal of Geophysical Research, 102(A4):7169–7177, 1997.

D.N. Baker, G.M. Mason, O. Figueroa, G. Colon, J.G. Watzin, and R.M. Aleman. An overview of the solar, anomalous, and magnetospheric particle explorer (SAMPEX) mission. IEEE Transactions on Geoscience and Remote Sensing, 31(3):531–541, 1993.

D.N. Baker, R.L. McPherron, T.E. Cayton, and R.W. Kebesadel. Linear prediction filter analysis of relativistic electron properties at 6.6 RE. Journal of Geophysical Research, 95(A9):15133–15140, 1990b.

D.N. Baker, T.I. Pulkkinen, J. Buchner, and A.J. Klimas. Substorms: A global instability of the magnetosphere-ionosphere system. Journal of Geophysical Research, 105: 14601–14612, 1999b.

D.N. Baker, T.I. Pulkkinen, X. Li, S.G. Kanekal, K.W. Obilvie, R.P. Lepping, J.B. Blake, L.B. Callis, G. Rostoker, H.J. Singer, and G.D. Reeves. A strong CME-related magnetic cloud interaction with the earth's magnetosphere: Istp observations of rapid relativistic electron acceleration on May 15, 1997. Geophysical Research Letters, 25(15):2975–2978, 1998.

L.F. Bargatze, D.N. Baker, R.L. McPherron, and E.W. Hones Jr. Magnetospheric impulse response for many levels of geomagnetic activity. Journal of Geophysical Research, 90:6387, 1985.

J.B. Blake, R.S. Selesnik, D.N. Baker, and S. Kanekal. Studies of relativistic electron injection events in 1997 and 1998. Journal of Geophysical Research, 106(A9): 19,157–19,168, 2001.

B.R. Boller and H.L. Stolov. Kelvin-helmholtz instability and the semi-annual variation of geomagnetic activity. Journal of Geophysical Research, 75(31):6073–6084, 1970.

D. Boscher, S. Bourdarie, and T. Beutier. Dynamic modeling of trapped particles. IEEE Transactions on Nuclear Science, 43(2):416–425, 1996.

S. Bourdarie, D. Boscher, T. Beutier, J.-A. Sauvaud, and M. Blanc. Electron and proton radiation belt dynamic simulations during storm periods: A new asymmetric convections-diffusion model. Journal of Geophysical Research, 102(A8):17,541–17,552, 1997a.

S. Bourdarie, D. Boscher, T. Beutier, J.-A. Sauvaud, and M. Blanc. Physical modeling of radiation belt response to geomagnetic storm: 1. on a 10 minute-10 hour time scale. Advances in Space Research, 20(3):377–380, 1997b.

G.E.P. Box and G.M. Jenkins. Time Series Analysis Forecasting and Control – Revised Edition. Holden-Day, Inc., Oakland, CA, 1976.

D.H. Brautigam and J.M. Albert. Radial diffusion analysis of outer radiation belt electrons during the october 9, 1990, magnetic storm. Journal of Geophysical Research, 105(A1):291–309, 2000.

D.H. Brautigam and J.T. Bell. CRRESELE Documentation, PL-TR-95-2128. Environmental Research Papers, 1178, 1995.

D.H. Brautigam, M.S. Gussenhoven, and E.G. Mullen. Quasi-static model of outer zone electrons. IEEE Transactions on Nuclear Science, 32:1797, 1992.

J.G. Charney, M. Halem, and R. Jastrow. Use of incomplete historical data to infer the present state of the atmosphere. Journal of Atmospheric Science, 26:1160–1163, 1969.

C.R. Clauer. The technique of linear prediction filters applied to studies of solar wind-magnetosphere coupling. In Y. Kamide and J.A. Slavin, editors, Solar Wind-Magnetosphere Coupling, pages 39–57. Terra Scientific Publishing Company, 1986.

C.R. Clauer, R.L. McPherron, and C. Searls. Solar wind control of the low-latitude asymmetric magnetic disturbance field. Journal of Geophysical Research, 88(A4): 2123, 1983.

C.R. Clauer, R.L. McPherron, C. Searls, and M.B. Kivelson. Solar wind control of auroral zone geomagnetic activity. Geophysical Research Letters, 8:915, 1981.

E.W. Cliver, Y. Kamide, and A.G. Ling. Mountains versus valleys: Semiannual variation of geomagnetic activity. Journal of Geophysical Research, 105(A2):2413–2424, 2000.

M. Codrescu, J.M. Forbes, and R.G. Roble. Thermospheric dynamics: A system theory approach. Radio Science, 25(4):299–308, 1990.

W.R. Cook, A.C. Cummings, J.R. Cummings, T.L Garrard, B. Kecman, R.A. Mewaldt, R.S. Selesnick, E.C. Stone, D.N. Baker, T.T. von Rosenvinge, J.B. Blake, and L.B. Callis. A proton electron telescope for studies of magnetospheric, solar, and galactic particles. IEEE Transactions on Geoscience and Remote Sensing, 31(3):656–571, 1993.

S.R. Elkington, D.N. Baker, and M.J. Wiltberger. Injection of energetic ions during the 31 march 0630 substorm. In Physics and Modeling of the Inner Magnetosphere. AGU Geophysical Monographs, 2004a. in press.

S.R. Elkington, M.K. Hudson, and A.A. Chan. Acceleration of relativistic electrons via drift-resonant interaction with toroidal-mode pc-5 ULF oscillations. Geophysical Research Letters, 26:3273–3276, 1999.

S.R. Elkington, M.K. Hudson, and A.A. Chan. Resonant acceleration and diffusion of outer zone electrons in an asymmetric geomagnetic field. Journal of Geophysical Research, 108(A3):doi:10.1029/2001JA009202, 2003.

S.R. Elkington, M.K. Hudson, M.J. Wiltberger, and J.G. Lyon. Mhd/particle simulations of radiation belt dynamics. Journal of Atmospheric and Solar-Terrestrial Physics, 64:607–615, 2002.

S.R. Elkington, M.J. Wiltberger, A.A. Chan, and D.N. Baker. Physical models of the geospace radiation environment, and the center for integrated space weather modeling. Journal of Atmospheric and Solar-Terrestrial Physics, 2004b. in press.

C.-G. Falthammer. Effects of time-dependent electric fields on geomagnetically trapped radiation. Journal of Geophysical Research, 70(11):2503–2516, 1965.

C.-G. Falthammer. On the transport of trapped particles in the outer magnetosphere. Journal of Geophysical Research, 71(5):1487–1491, 1966.

S.F. Fung. Recent development in the nasa trapped radiation models. In J.F. Lemaire, D. Heynderickx, and D.N. Baker, editors, Radiation Belts: Models and Standards, Geophysical Monograph 97, pages 79–91, Washington D.C., 1996. AGU.

A. Gauthier and I.D. Landau. On the recursive identification of multi-input, multi-output systems. Automatica, 14:609–614, 1978.

V.V. Gavrishchaka and S.B. Ganguli. Optimization of the neural network geomagnetic model for forecasting large-amplitude substorm events. Journal of Geophysical Research, 106(A4):6247–6257, 2001.

W.D. Gonzalez. A unified view of solar wind-magnetosphere coupling functions. Planetary and Space Science, 38(5):627–632, 1990.

J.C. Green and M.G. Kivelson. Relativistic electrons in the outer radiation belt: Differentiating between acceleration mechanisms. Journal of Geophysical Research, 109: doi:10.1029/2003JA010153, 2004.

R.P. Guidorzi. Canonical structures in the identification of multivariable systems. Automatica, 11:361–374, 1975.

R.P. Guidorzi. Invariants and canonical forms for systems structural and parametric identification. Automatica, 17:117–133, 1981.

T. Iyemori, H. Maeda, and T. Kamei. Impulse response of geomagnetic indices to interplanetary magnetic field. Journal of Geomagnetism and Geoelectricity, 31:1–9, 1979.

R. Johansson. System Modeling and Identification. Prentice Hall, Englewood Cliffs, NJ, 1993.

Y. Kamide and J.A. Slavin, editors. Solar Wind-Magnetosphere Coupling, Astrophysics and Space Science Library, Tokyo, 1986. Terra Scientific Publishing Company.

S.G. Kanekal, D.N. Baker, and J.B. Blake. Multisatellite measurements of relativistic electrons: Global coherence. Journal of Geophysical Research, 106(A12):29,721–29,732, 2001.

S.G. Kanekal, D.N. Baker, J.B. Blake, B. Klecker, R.A. Mewaldt, and G.M. Mason. Magnetospheric response to magnetic cloud (CME) events: Energetic particle observations from sampex and polar. Journal of Geophysical Research, 104, 1999.

H.J. Kim and A.A. Chan. Fully adiabatic changes in storm time relativistic electron fluxes. Journal of Geophysical Research, 102(A10), 1997.

H.J. Kim, G. Rostoker, and Y. Kamide. Radial dependence of relativistic electron fluxes for storm main phase development. Journal of Geophysical Research, 107(A11): doi:10.1029/2001JA007513, 2002.

M.G. Kivelson and C.T. Russell, editors. Introduction to Space Physics. Cambridge University Press, 1995.

A.J. Klimas, D. Vassiliadis, and D.N. Baker. Data-derived analogues of the magnetospheric dynamics. Journal of Geophysical Research, 102(A12):26,993–27,009, 1997.

A.J. Klimas, D. Vassiliadis, D.N. Baker, and D.A. Roberts. The organized nonlinear dynamics of the magnetosphere. Journal of Geophysical Research, 101(A6):13,089–13,113, 1996.

H.C. Koons and D.J. Gorney. A neural network model of the relativistic electron flux at geosynchronous orbit. Journal of Geophysical Research, 96:5549–5556, 1991.

C.T.W. Kundstadter. Were we crying wolf? insurance implications of the leonids and other space phenomena. In AIAA Leonids Conference, 1999.

X. Li, D.N. Baker, S.G. Kanekal, M. Looper, and M. Temerin. Long term measurements of radiation belts by sampex and their variations. Geophysical Research Letters, 28 (20):3827–3830, 2001a.

X. Li, D.N. Baker, M. Temerin, T.E. Cayton, G.D. Reeves, R.A. Christensen, J.B. Blake, M.D. Looper, R. Nakamura, and S.G. Kanekal. Multisatellite observations of the outer zone electron variation during the November 3-4, 1993, magnetic storm. Journal of Geophysical Research, 102(A7):14,123–14,140, 1997.

X. Li, I. Roth, M. Temerin, J.R. Wygant, M.K. Hudson, and J.B. Blake. Simulation of the prompt energization and transport of radiation belt particles during the march 24, 1991 ssc. Geophysical Research Letters, 20(22):2423–2426, 1993.

X. Li and M.A. Temerin. The electron radiation belt. Space Science Reviews, 95: 569–580, 2001.

X. Li, M.A. Temerin, D.N. Baker, G.D. Reeves, and D. Larson. Quantitative prediction of radiation belt electrons at geostationary orbit based on solar wind measurements. Geophysical Research Letters, 28(9):1887–1890, 2001b.

L. Ljung. Asymptotic behavior of the extended kalman filter as a paramter estimator for linear systems. IEEE Transactions on Automatic Control, ac-24(1):36–50, 1979.

L. Ljung. System Identification: Theory for the User. Prentice Hall PTF, Upper Saddle River, NJ, 2nd edition, 1999.

L. Ljung and S. Gunnarsson. Adaptation and tracking in system identification – a survey. Automatica, 26:7–21, 1990.

L. Ljung and T. Söderström. Theory and Practice of Recursive Identification. MIT Press, Cambridge, Massachusetts, 1983.

D.J. McComas, S.J. Barne, B.L. Barraclough, W.C. Feldman, H.O. Funsten, J.T. Gosling, P. Riley, R. Skoug, A. Balogh, R. Forsyth, B.E. Goldstein, and M. Neugebauer. Ulysses' return to the slow solar wind. Geophysical Research Letters, 25(1): 1–4, 1998.

D.J. McComas, B.L. Barraclough, H.O. Funsten, J.T. Gosling, E. Santiago-Muñoz, R.M. Skoug, B.E. Goldstein, M. Neugebauer, P. Riley, and A. Balogh. Solar wind observations over ulysses first full polar orbit. Journal of Geophysical Research, 105 (A5):10,419–10,433, 2000.

R.L. McPherron, R.A. Fay, C.R. Garrity, L.F. Bargatze, D.N. Baker, C.R. Clauer, and C. Searls. Coupling of the solar wind to measures of magnetic activity. In B. Battrick and E. Rolfe, editors, Achievements of the International Magnetospheric Study (IMS), Proceedings of an International Symposium, Graz, Austria 26-28 June, number 217 in ESA SP series, page 161. European Space Agency, 1984.

R.A. Mewaldt, A.C. Cummings, and E.C. Stone. The anomalous cosmic rays - interstellar interlopers in the heliosphere and the magnetosphere. Eos Transactions AGU, 85:185, 1994.

D.F. Moorer. Modeling High Energy Electron Flux in the Outer Radiation Belt. PhD thesis, University of Colorado, 2000.

D.F. Moorer and D.N. Baker. Specification of energetic magnetospheric electrons. In P. Song, H.J. Singer, and G.L. Siscoe, editors, Space Weather, 2001.

T. Nagai. Space weather forecast: Prediction of relativistic electron intensity at synchronous orbit. Geophysical Research Letters, 15(5):425–428, 1988.

National Space Science Data Center. OMNIWeb - Near Earth Heliosphere Data, National Space Science Data Center, NASA. http://nssdc.gsfc.nasa.gov/omniweb/, 2003.

O. Nelles. Nonlinear System Identification. Springer-Verlag, Berlin, Heidelberg, Germany, 2001.

NPOESS. IORD II – National Polar-Orbiting Operational Environmental Satellite System (NPOESS). http://npoesslib.ipo.noaa.gov/, 2001.

NSWP. National space weather program strategic plan. Report FCM-P30-1995, Office of the Federal Coordinator for Meteorological Services and Supporting Research, Silver Spring, MD, 1996.

W.P. Olson and K.A. Pfitzer. Magnetospheric magnetic field modeling. Annual scientific report, McDonnel Douglas Astronautics Company, Huntington Beach, CA, 1977.

G.A. Paulikas and J.B. Blake. Effects of the solar wind on magnetospheric dynamics: Energetic electrons at the synchronous orbit. In Qualitative Modeling of Magnetospheric Processes, Geophysical Monograph Series, pages 180–202. AGU, Washington, D.C., USA, 1979.

P. Perreault and S.-I. Akasofu. A study of geomagnetic storms. Geophysical Journal of the Royal Astronomical Society, 54:547–573, 1978.

W.H. Press, S.A. Teukolsky, W.T. Vetterling, and B.P. Flannery. Numerical Recipes in C: The Art of Scientific Computing, chapter 13.6. Cambridge University Press, 2nd edition, 1992.

E.J. Rigler, D.N. Baker, R.S. Weigel, and D. Vassiliadis. Solar wind-driven electron radiation belt response functions at 100-min time scales. Advances in Space Research, in press.

E.J. Rigler, D.N. Baker, R.S. Weigel, D. Vassiliadis, and A.J. Klimas. Adaptive linear prediction of radiation belt electrons using the kalman filter. Space Weather, 2: doi:10.1029/2003SW000036, 2004.

E.A. Robinson. Multichannel Time Series Analysis With Digital Computer Programs. Goose Pond Press, 2nd edition, 1983.

D.J. Rodgers. Empirical radiation belt models. In J.F. Lemaire, D. Heynderickx, and D.N. Baker, editors, Radiation Belts: Models and Standards, Geophysical Monograph 97, pages 147–148, Washington D.C., 1996. AGU.

J.G. Roederer. Dynamics of Geomagnetically Trapped Radiation. Springer-Verlag, Berlin, Germany, 1970.

G. Rostoker, S. Skone, and D.N. Baker. On the origin of relativistic electrons in the magnetosphere associated with some geomagnetic storms. Geophysical Research Letters, 25(19):3701–3704, 1998.

G. Rottman. The solar irradiance and its variations. In The Evolving Sun and its Influence on Planetary Environments, volume 269, pages 25–36. Astronomical Society of the Pacific, San Francisco, California, USA, 2002.

G. Rottman, T. Woods, M. Snow, and G. DeToma. The solar cycle variation in ultraviolet irradiance. Advances in Space Research, 27(12):1927–1932, 2001.

C.T. Russell and R.L. McPherron. Semiannual variation of geomagnetic activity. Journal of Geophysical Research, 78(1):92–107, 1973.

M. Schulz and L.J. Lanzerotti. Particle Diffusion in the Radiation Belts. Springer-Verlag, New York, NY, 1974.

SEC. REFM - Relativistic Electron Forecast Model, Space Environment Center, NOAA. http://www.sel.noaa.gov/refm/, 2003.

Y.Y. Shprits and R.M. Thorne. Time dependent radial diffusion modeling of relativistic electrons with realistic loss rates. Geophysical Research Letters, 31: doi:10.1029/2004GL019591, 2004.

J.J. Shynk. Adaptive IIR filtering. IEEE AASP Magazine, pages 4–29, 1989.

D.H.M. Spector. Building Linux Clusters: Scaling Linux for Scientific and Enterprise Applications. O'Reilly & Associates, Inc., Sebastopol, CA, 1st edition, 2000.

W.N. Spjeldvik and P.L. Rothwell. Handbook of Geophysics and the Space Environment, chapter 5. U.S. Air Force Geophysics Laboratory, 1985.

A.P. Stern and N.F. Ness. Planetary magnetospheres. NASA Technical Memorandum 83841, Goddard Space Flight Center, Greenbelt, MD, 1981.

T.F. Tascione. Introduction to the Space Environment. Krieger Publishing Company, Malabar, Florida, USA, 2 edition, 1994.

D. Vassiliadis. System identification, modeling, and prediction for space weather environments. IEEE Transactions on Plasma Science, 28(6):1944–1955, 2000.

D. Vassiliadis, S.F. Fung, and A.J. Klimas. Interplanetary and magnetospheric state parameters for the radiation belt electron flux. Journal of Geophysical Research, 2004. under review.

D. Vassiliadis and A.J. Klimas. On the uniqueness of linear moving-average filters for the solar wind-auroral geomagnetic activity coupling. Journal of Geophysical Research, 100(A4):5637–5641, 1995.

D. Vassiliadis, A.J. Klimas, D.N. Baker, and D.A. Roberts. A description of the solar wind-magnetosphere coupling based on nonlinear filters. Journal of Geophysical Research, 100(A3):3495–3512, 1995.

D. Vassiliadis, A.J. Klimas, S.G. Kanekal, D.N. Baker, and R.S. Weigel. Long-term-average, solar cycle, and seasonal response of magnetospheric energetic electrons to the solar wind speed. Journal of Geophysical Research, 107(A11): doi:10.1029/2001JA000506, 2002.

D. Vassiliadis, A.J. Klimas, J.A. Valdivia, and D.N. Baker. The D_{st} geomagnetic response as a function of storm phase and amplitude and the solar wind electric field. Journal of Geophysical Research, 104(A11):24,957–24,976, 1999.

D. Vassiliadis, A.J. Klimas, J.A. Valdivia, and D.N. Baker. The nonlinear dynamics of space weather. Advances in Space Research, 26(1):197–207, 2000.

D. Vassiliadis, A.J. Klimas, R.S. Weigel, D.N. Baker, E.J. Rigler, S.G. Kanekal, T. Nagai, S.F. Fung, R.W.H. Friedel, and T.E. Cayton. Structure of earth's outer radiation belt inferred from long-term electron flux dynamics. Geophysical Research Letters, 30(19):doi:10.1029/2003GL017328, 2003a.

D. Vassiliadis, R.S. Weigel, A.J. Klimas, S.G. Kanekal, and R.A Mewaldt. Modes of energy transfer from the solar wind to the inner magnetosphere. Physics of Plasmas, 10(2):463–473, 2003b.

V.M. Vasyliunas, J.R. Kan, G.L. Siscoe, and S.-I. Akasofu. Scaling relations governing magnetospheric energy transfer. Planetary and Space Science, 30:359–365, 1982.

J.I. Vette. Models of the trapped radiation environment, volume I: Inner zone protons and electron. NASA Technical Paper SP-3024, NASA, Washington, D.C., 1966.

J.I. Vette. The AE-8 trapped electron model environment. NASA Technical Paper 91-24, NASA, Greenbelt, MD, 1991.

R.S. Weigel, A.J. Klimas, and D. Vassiliadis. Solar wind coupling to and predictability of ground magnetic fields and their time derivatives. Journal of Geophysical Research, 108:1298, 2003.

R.S. Weigel, D. Vassiliadis, and A.J. Klimas. Coupling of the solar wind to temporal fluctuations in ground magnetic fields. Geophysical Research Letters, 29(19): doi:10.1029/2002GL014740, 2002.

O.R. White and G.J. Rottman. Solar variability observations. Space Science Reviews, 94:93–97, 2000.

N. Wiener. Extrapolation, Interpolation, and Smoothing of Stationary Time Series with Engineering Applications. The Technology Press and John Wiley and Sons, Inc., New York, NY, 1949.

Appendix A

Training/Validation Data

The folowing pages contain high resolution (relative to the rather compressed figures in the main body of the text) time series of the OMNI solar wind speed data (V_{sw}), IMF magnitude ($|\mathbf{B}|$), and ion number density (ρ_{ion}), as well as corresponding 2-D (L-shell vs. log(flux)) time series of the SAMPEX 2-6 MeV electron observations. The data presented is exactly the data that was used to produce all the models and all the predictions presented in the thesis as well as Appendix B, including periods that the time-average of the measurements was substituted for missing data as described in Chapter 2. This is done for posterity's sake (i.e., the author anticipates a need to recall **exactly** what he did to finagle his doctorate 20 years hence since he will have, in all likelihood, lost all digital versions of his work by then)

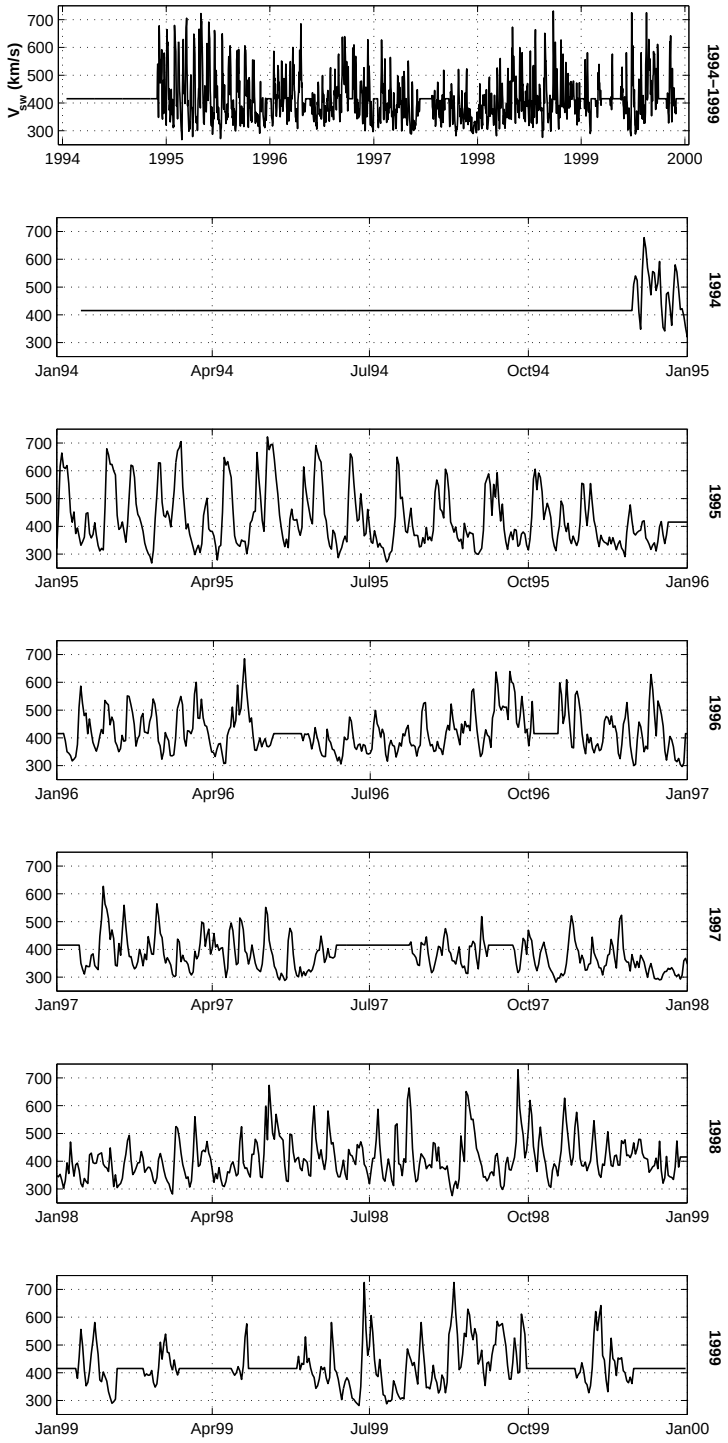


Figure A.1: Daily-averaged solar wind speed measurements are plotted. Data gaps are filled as described in Chapter 2, and easily identified by the horizontal line throughout most of 1994 and subsequent shorter periods.

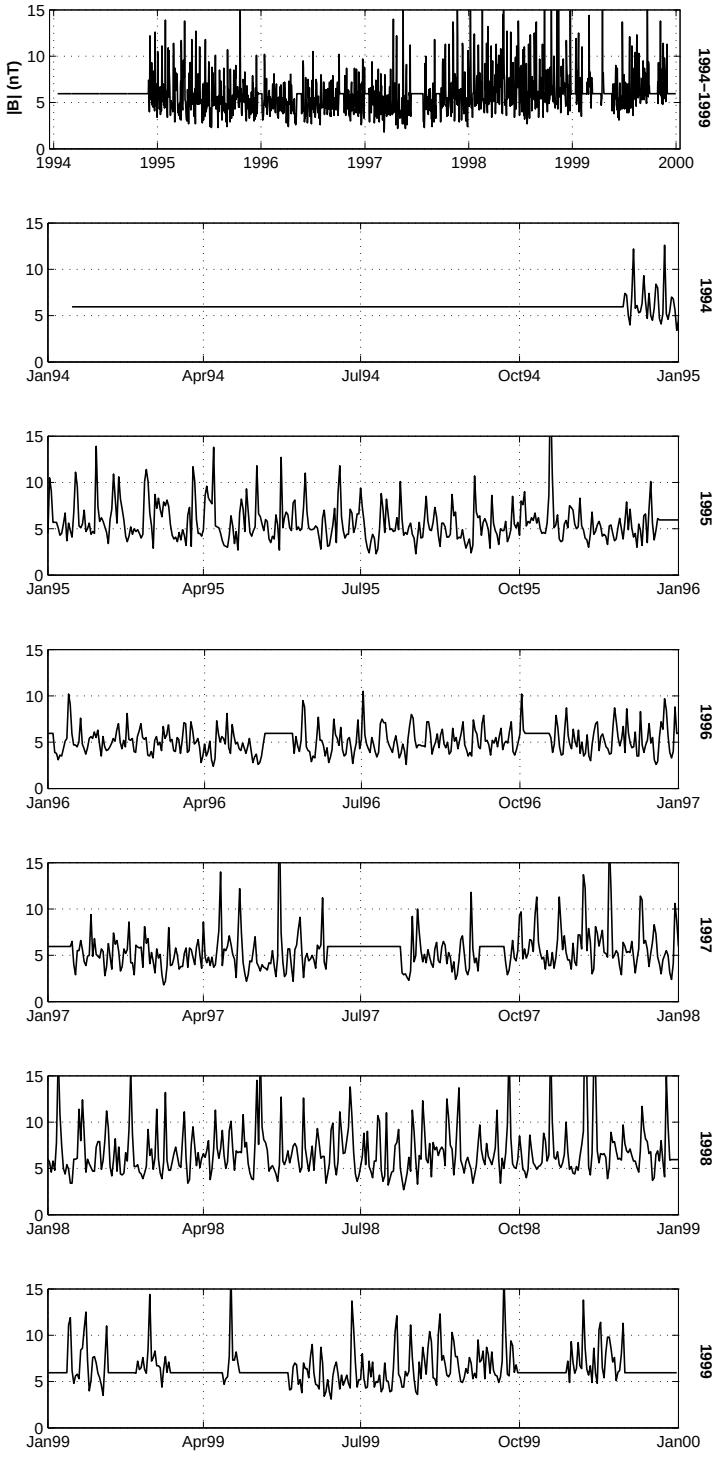


Figure A.2: Daily-averaged inter-planetary magnetic field magnitude measurements are plotted. Data gaps are filled as described in Chapter 2, and easily identified by the horizontal line throughout most of 1994 and subsequent shorter periods.

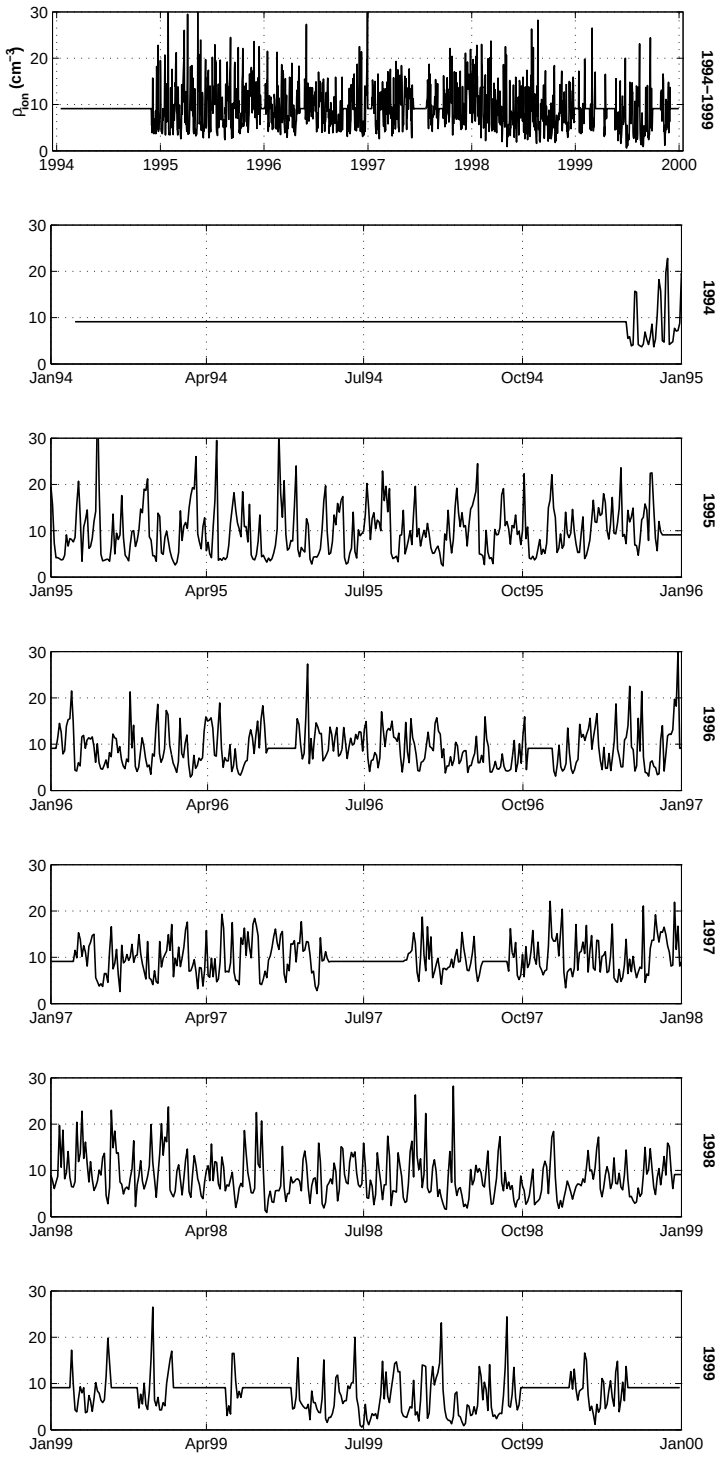


Figure A.3: Daily-averaged solar wind ion density measurements are plotted. Data gaps are filled as described in Chapter 2, and easily identified by the horizontal line throughout most of 1994 and subsequent shorter periods.

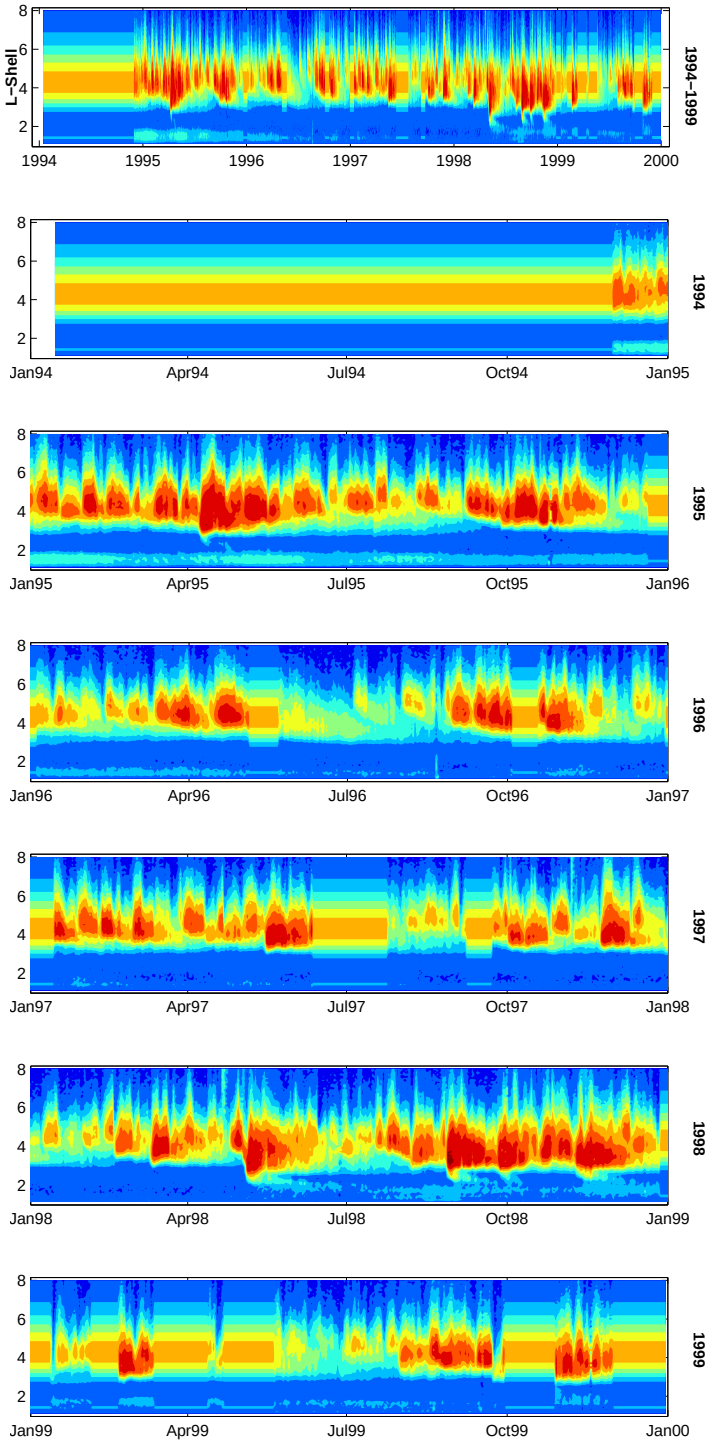


Figure A.4: Daily-Averaged 2-6 MeV electron $\log_{10}(\text{flux})$ measurements are plotted vs. L-shell. Data gaps are filled as described in Chapter 2, and easily identified by the horizontal contours throughout most of 1994 and subsequent shorter periods.

Appendix B

Predictions and Residuals

The folowing pages contain high resolution (relative to the rather compressed figures in the main body of the text) time series of the various predictions, residuals, and correlation functions presented in the thesis. Periods corresponding to when time-averages of the training data were substituted for missing data, as described in Chapter 2, are left in place for the sake of posterity (i.e., the author anticipates a need to recall **exactly** what he did to finagle his doctorate 20 years hence, since he will have, in all likelihood, lost all digital versions of his work by then).

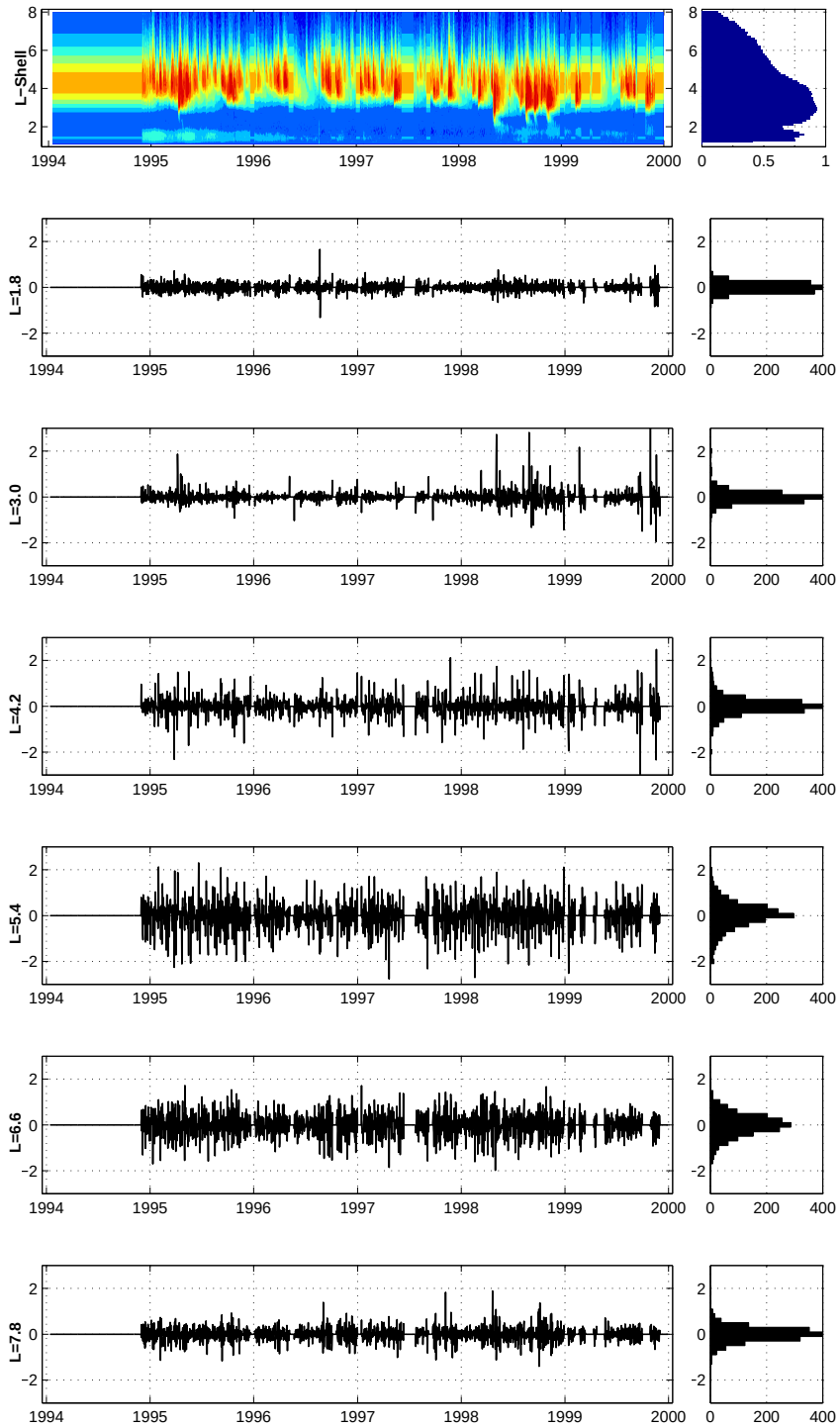


Figure B.1: One-day persistence log(flux) "predictions" are plotted vs. L-shell, along with the resulting residuals and their histograms for L-shells representative of the inner electron radiation belt (L=1.8), the slot region (L=3.0), GPS orbits (L=4.2), the heart of the outer-belt (L=5.4), geostationary orbit (L=6.6), and the outer edge of the stable trapping region (L=7.8).

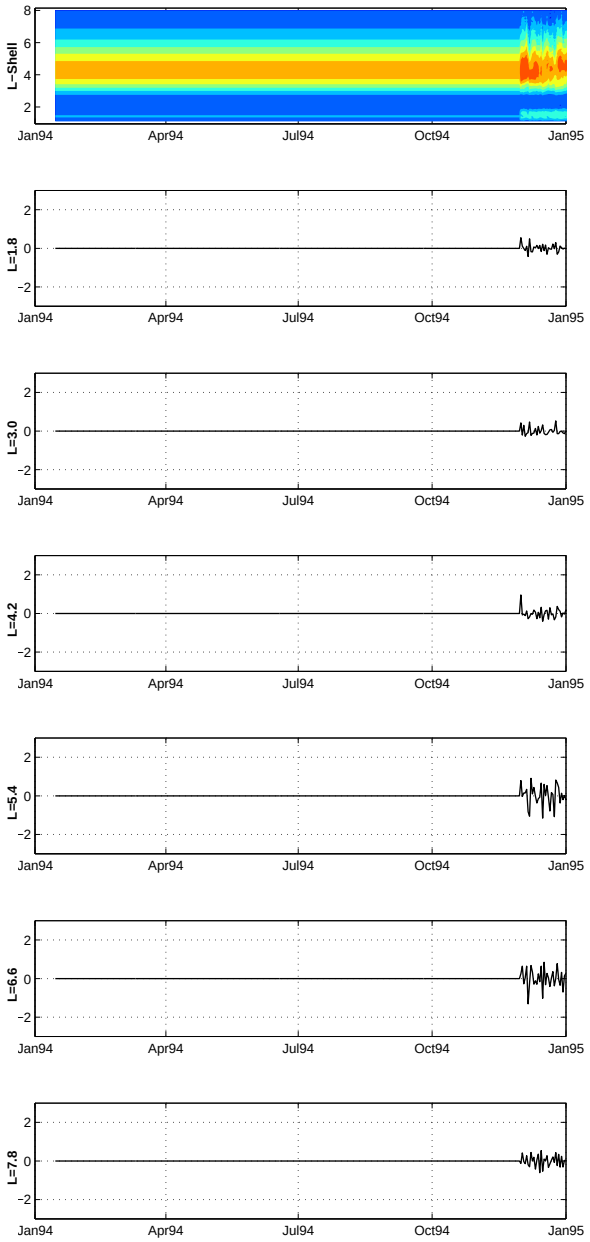


Figure B.2: One-day persistence log(flux) "predictions" are plotted vs. L-shell, along with the resulting residuals for representative L-shells in 1994.

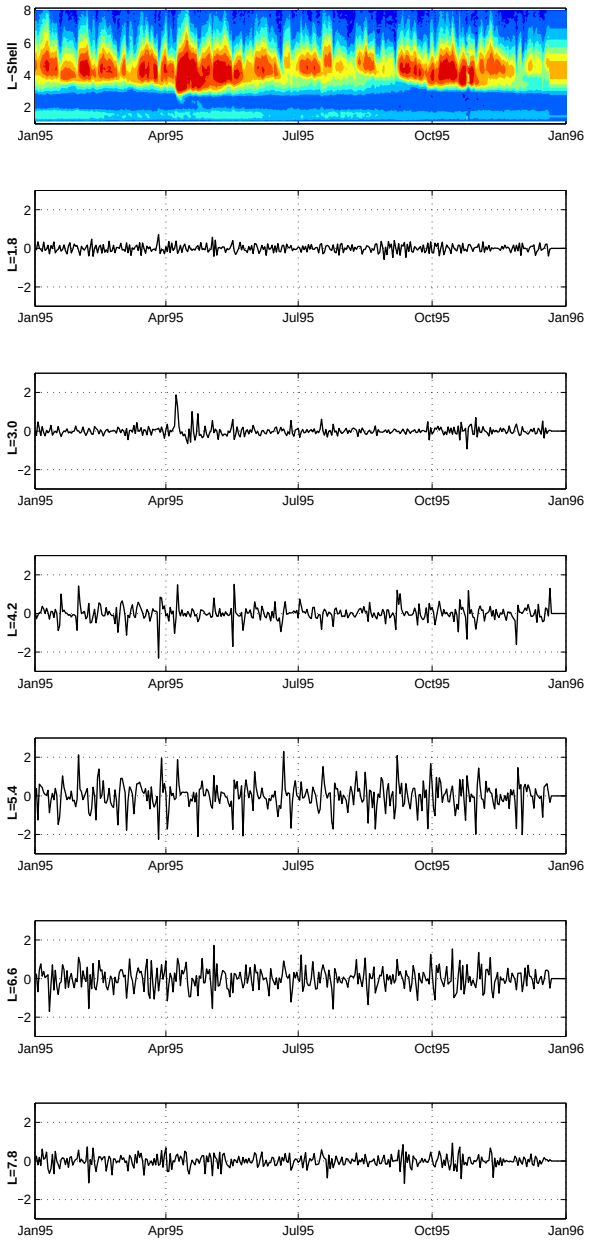


Figure B.3: One-day persistence log(flux) "predictions" are plotted vs. L-shell, along with the resulting residuals for representative L-shells in 1995.

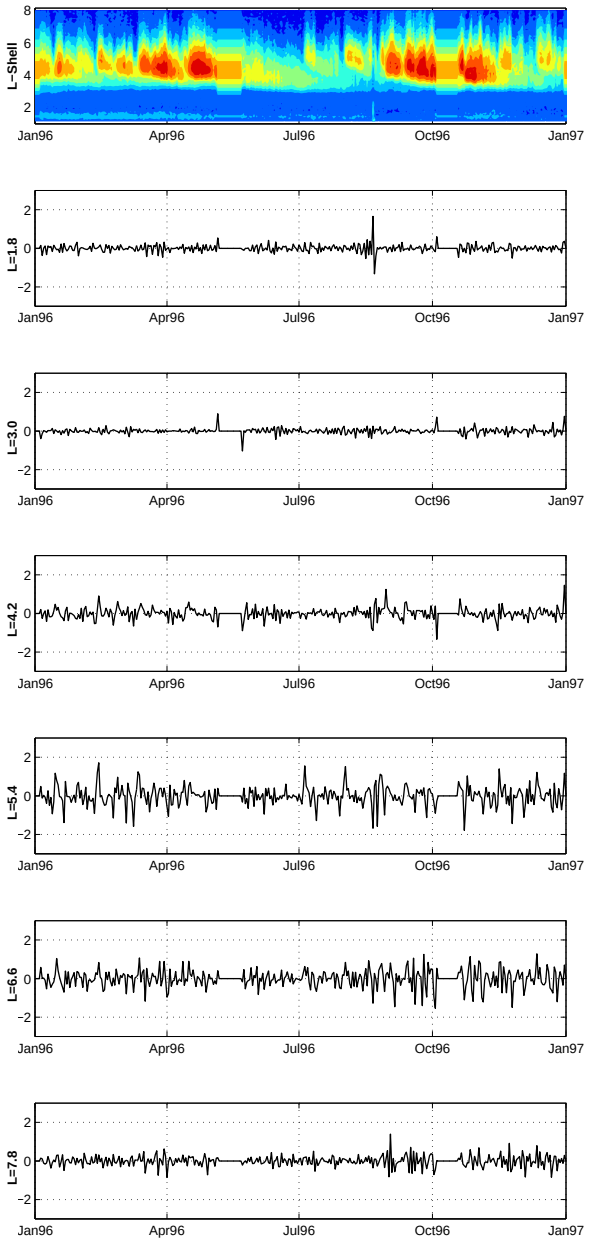


Figure B.4: One-day persistence log(flux) "predictions" are plotted vs. L-shell, along with the resulting residuals for representative L-shells in 1996.

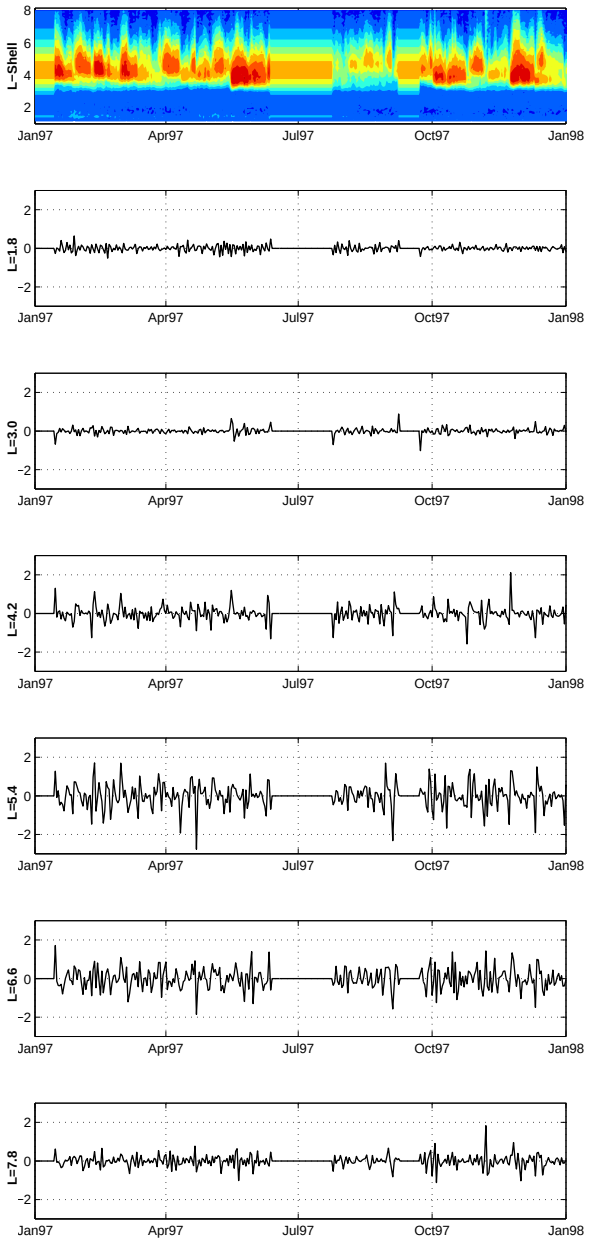


Figure B.5: One-day persistence log(flux) "predictions" are plotted vs. L-shell, along with the resulting residuals for representative L-shells in 1997.

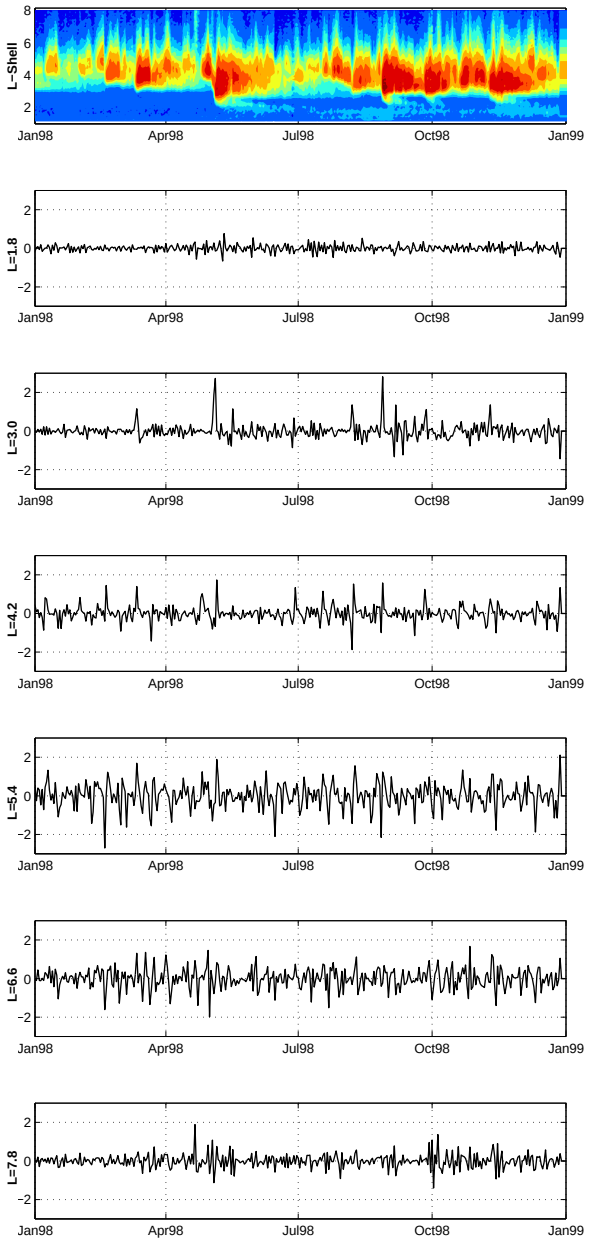


Figure B.6: One-day persistence log(flux) "predictions" are plotted vs. L-shell, along with the resulting residuals for representative L-shells in 1998.

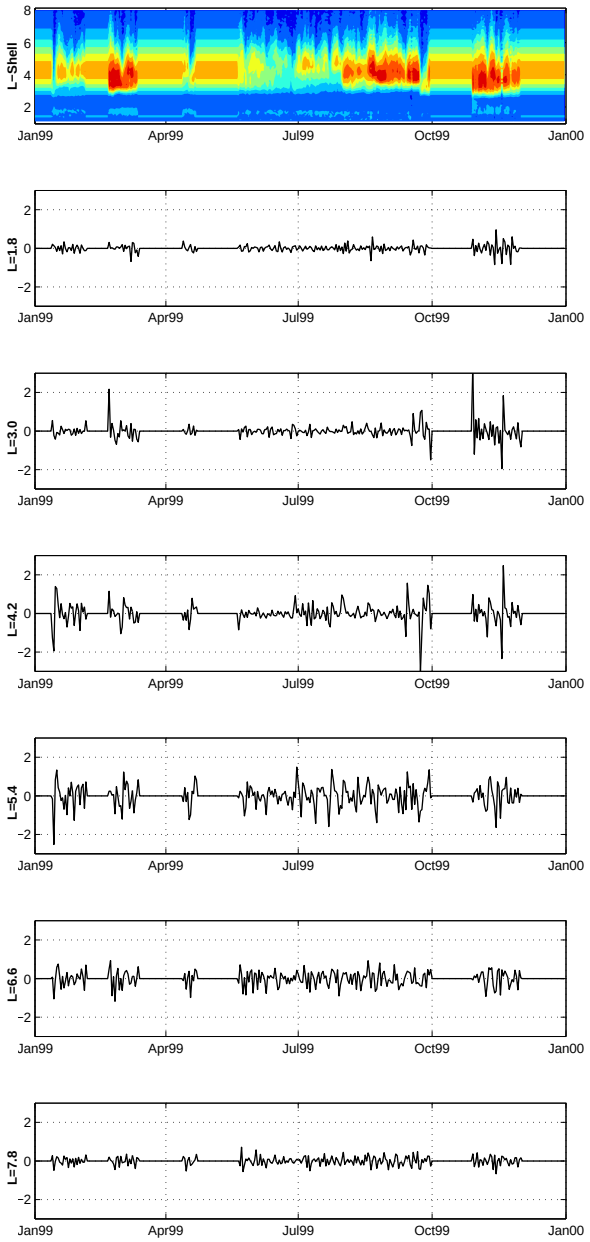


Figure B.7: One-day persistence log(flux) "predictions" are plotted vs. L-shell, along with the resulting residuals for representative L-shells in 1999.

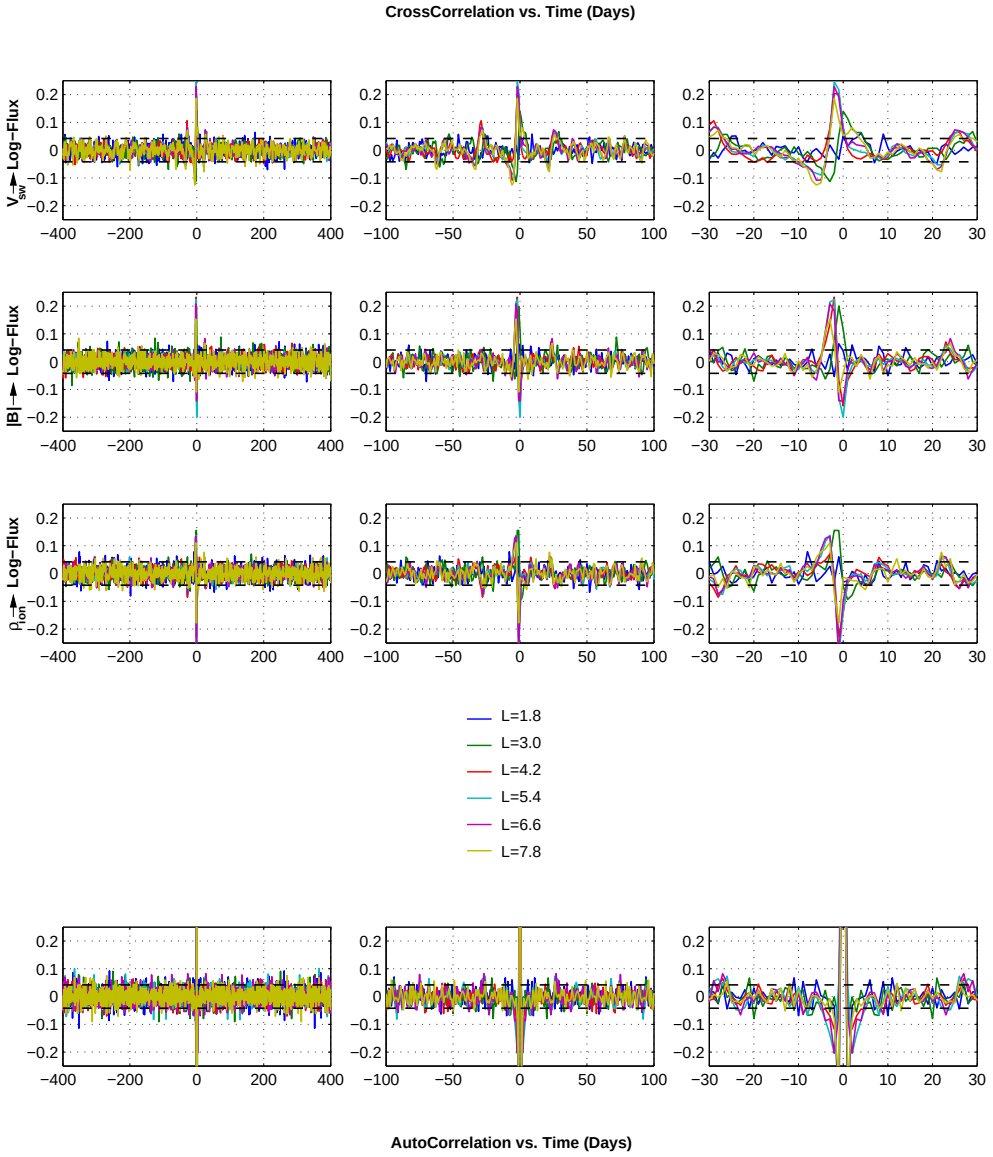


Figure B.8: Cross- and Auto-correlation functions of one-day persistence "predictions". These are essentially the same as the plots presented in Chapter 4, and simply provide additional "zoomed" views of the time-dependent correlation coefficients.

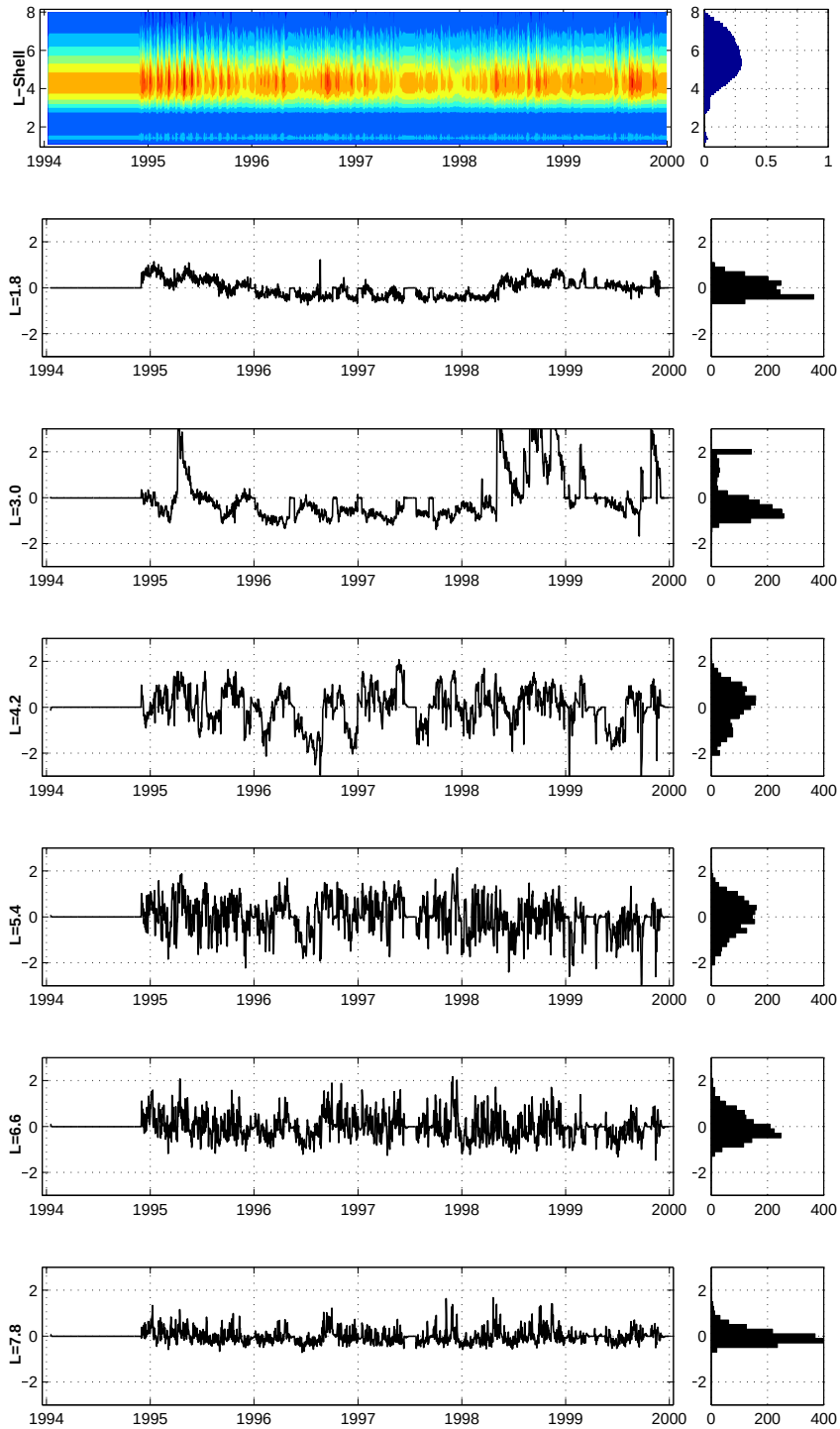


Figure B.9: Stationary $V_{sw} \rightarrow \log(\text{flux})$ predictions are plotted vs. L-shell, along with the resulting residuals and their histograms for L-shells representative of the inner electron radiation belt (L=1.8), the slot region (L=3.0), GPS orbits (L=4.2), the heart of the outer-belt (L=5.4), geostationary orbit (L=6.6), and the outer edge of the stable trapping region (L=7.8).

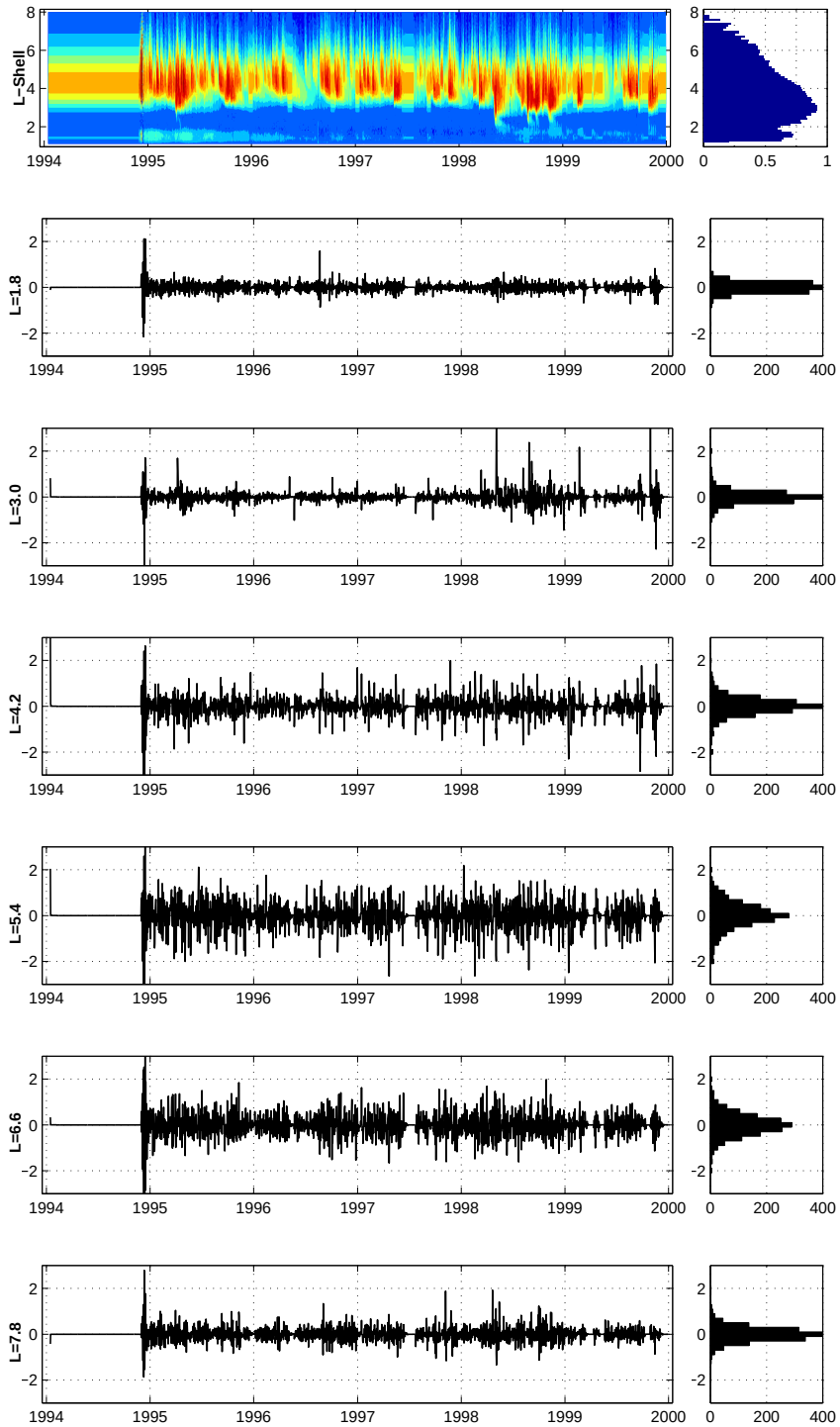


Figure B.10: Adaptive $V_{sw} \rightarrow \log(\text{flux})$ predictions are plotted vs. L-shell, along with the resulting residuals and their histograms for L-shells representative of the inner electron radiation belt (L=1.8), the slot region (L=3.0), GPS orbits (L=4.2), the heart of the outer-belt (L=5.4), geostationary orbit (L=6.6), and the outer edge of the stable trapping region (L=7.8).

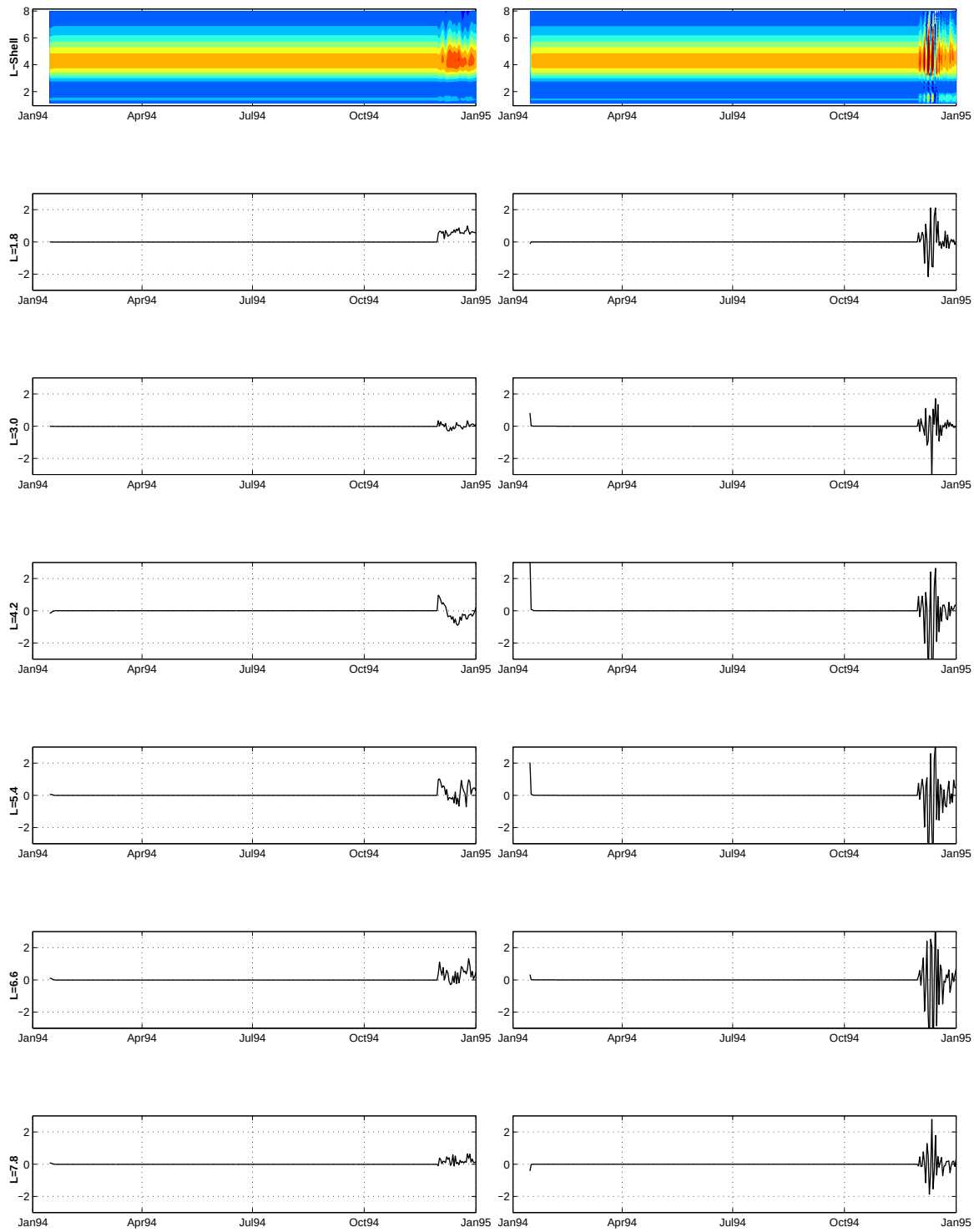


Figure B.11: Stationary (left) and Adaptive (right) $V_{sw} \rightarrow \log(\text{flux})$ predictions are plotted vs. L-shell, along with the resulting residuals for representative L-shells in 1994.

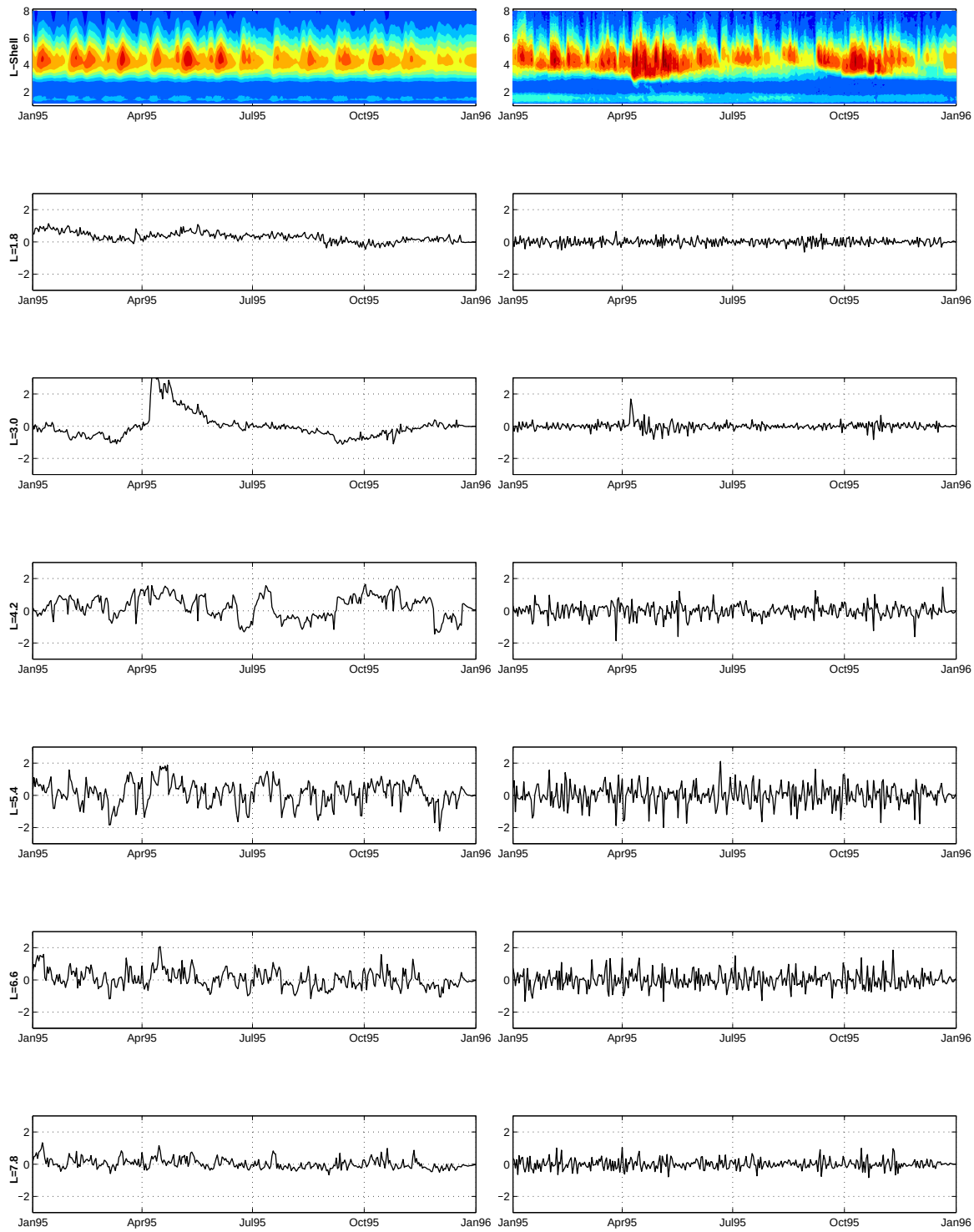


Figure B.12: Stationary (left) and Adaptive (right) $V_{sw} \rightarrow \log(\text{flux})$ predictions are plotted vs. L-shell, along with the resulting residuals for representative L-shells in 1995.

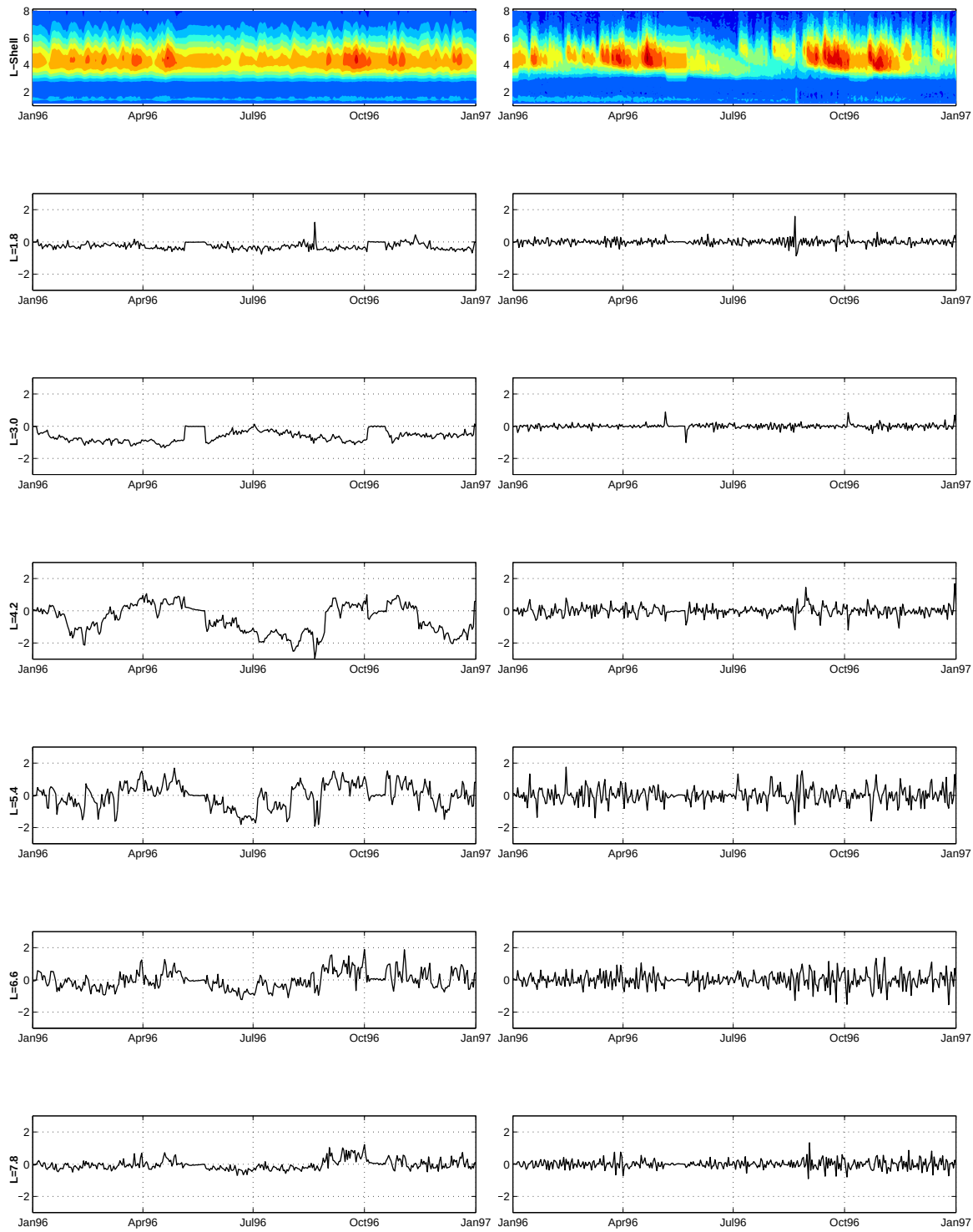


Figure B.13: Stationary (left) and Adaptive (right) $V_{sw} \rightarrow \log(\text{flux})$ predictions are plotted vs. L-shell, along with the resulting residuals for representative L-shells in 1996.

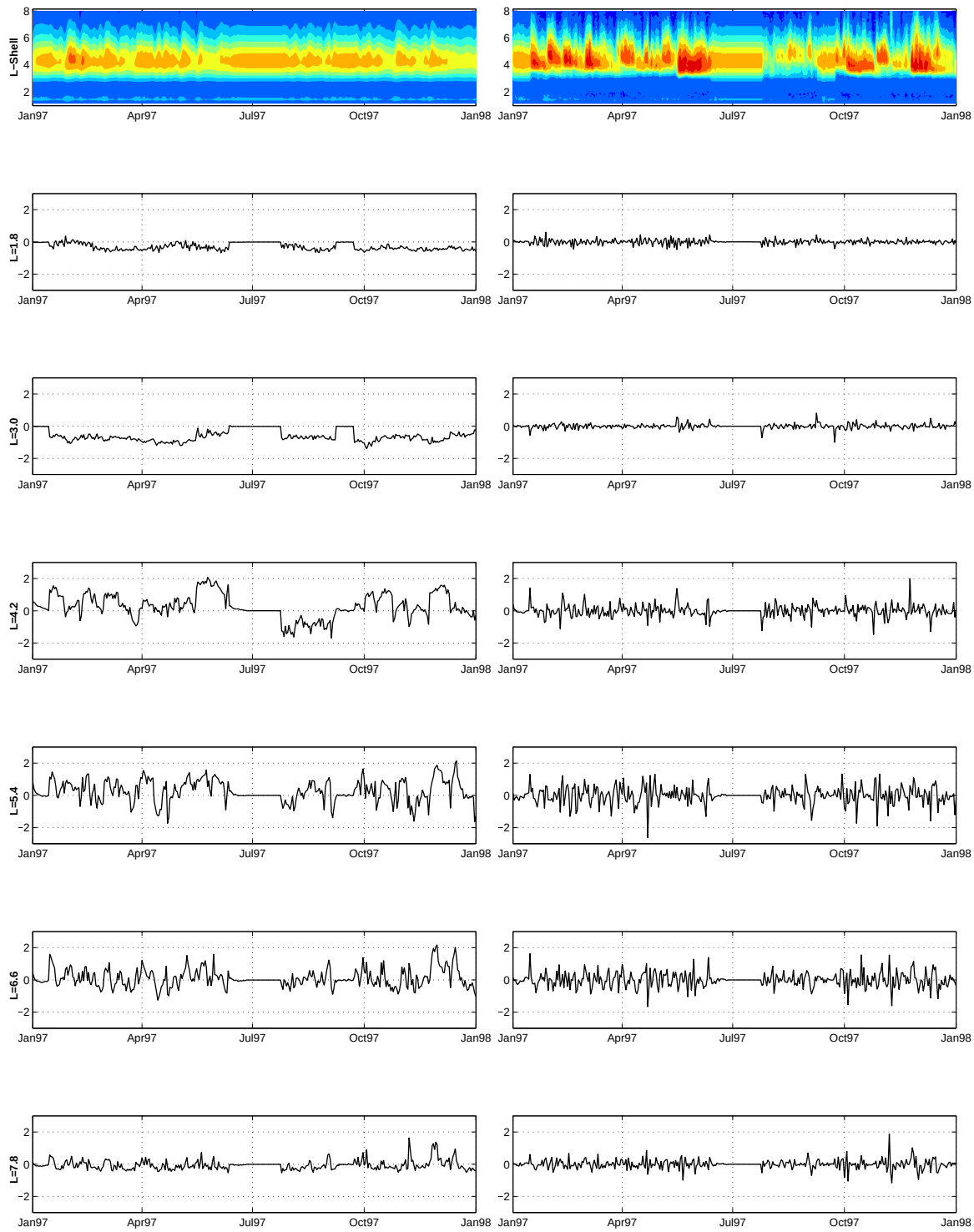


Figure B.14: Stationary (left) and Adaptive (right) $V_{sw} \rightarrow \log(\text{flux})$ predictions are plotted vs. L-shell, along with the resulting residuals for representative L-shells in 1997.

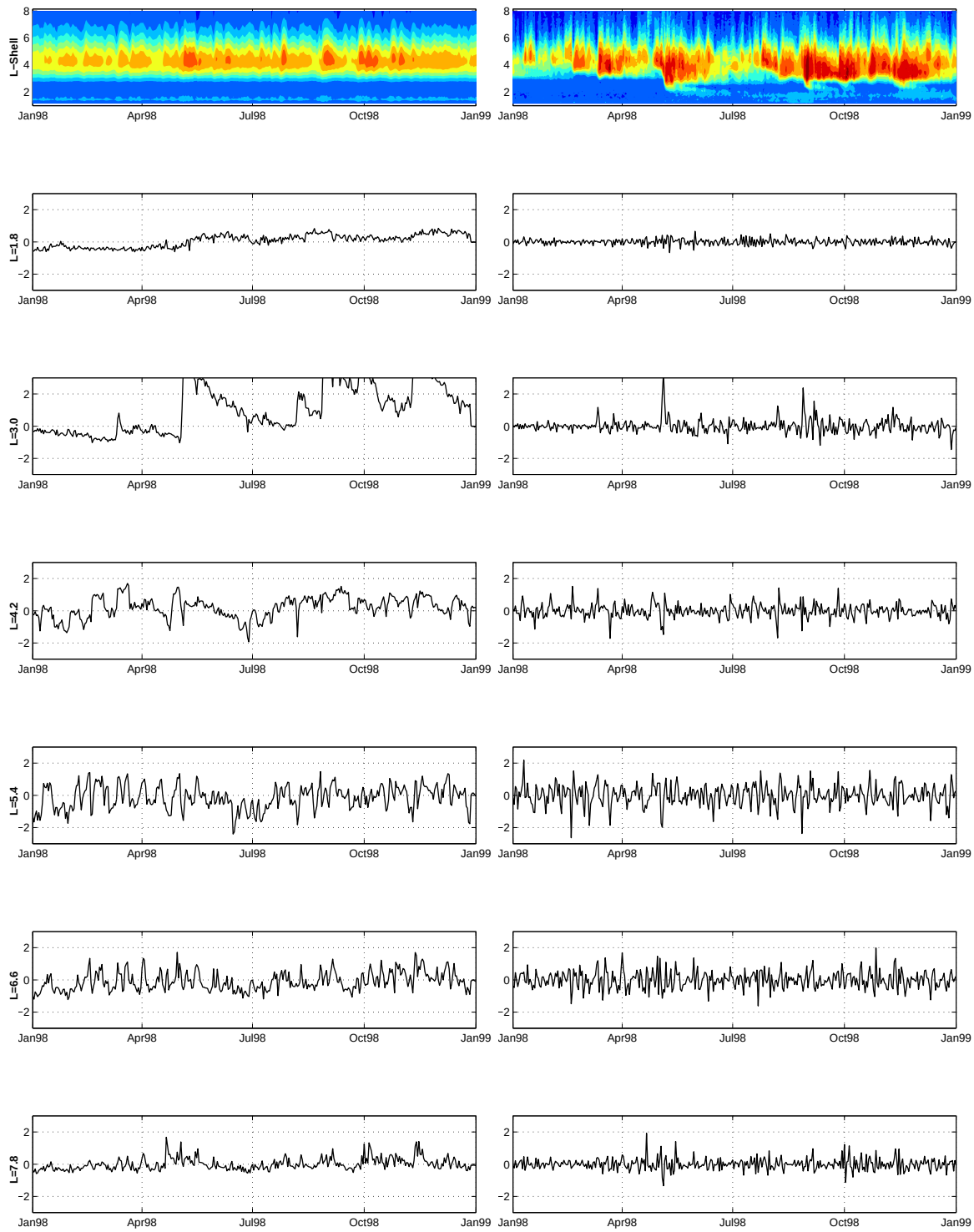


Figure B.15: Stationary (left) and Adaptive (right) $V_{sw} \rightarrow \log(\text{flux})$ predictions are plotted vs. L-shell, along with the resulting residuals for representative L-shells in 1998.

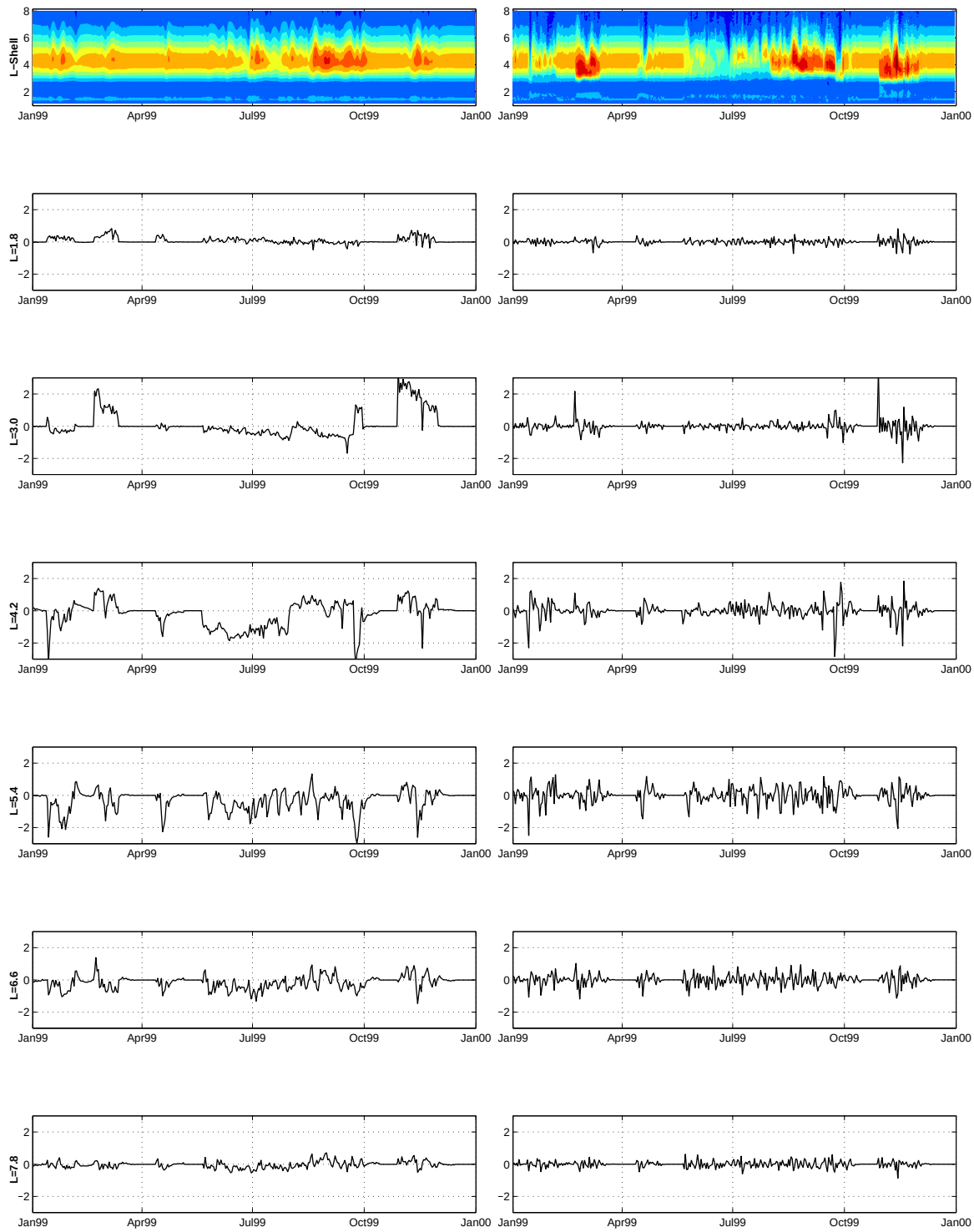


Figure B.16: Stationary (left) and Adaptive (right) $V_{sw} \rightarrow \log(\text{flux})$ predictions are plotted vs. L-shell, along with the resulting residuals for representative L-shells in 1999.

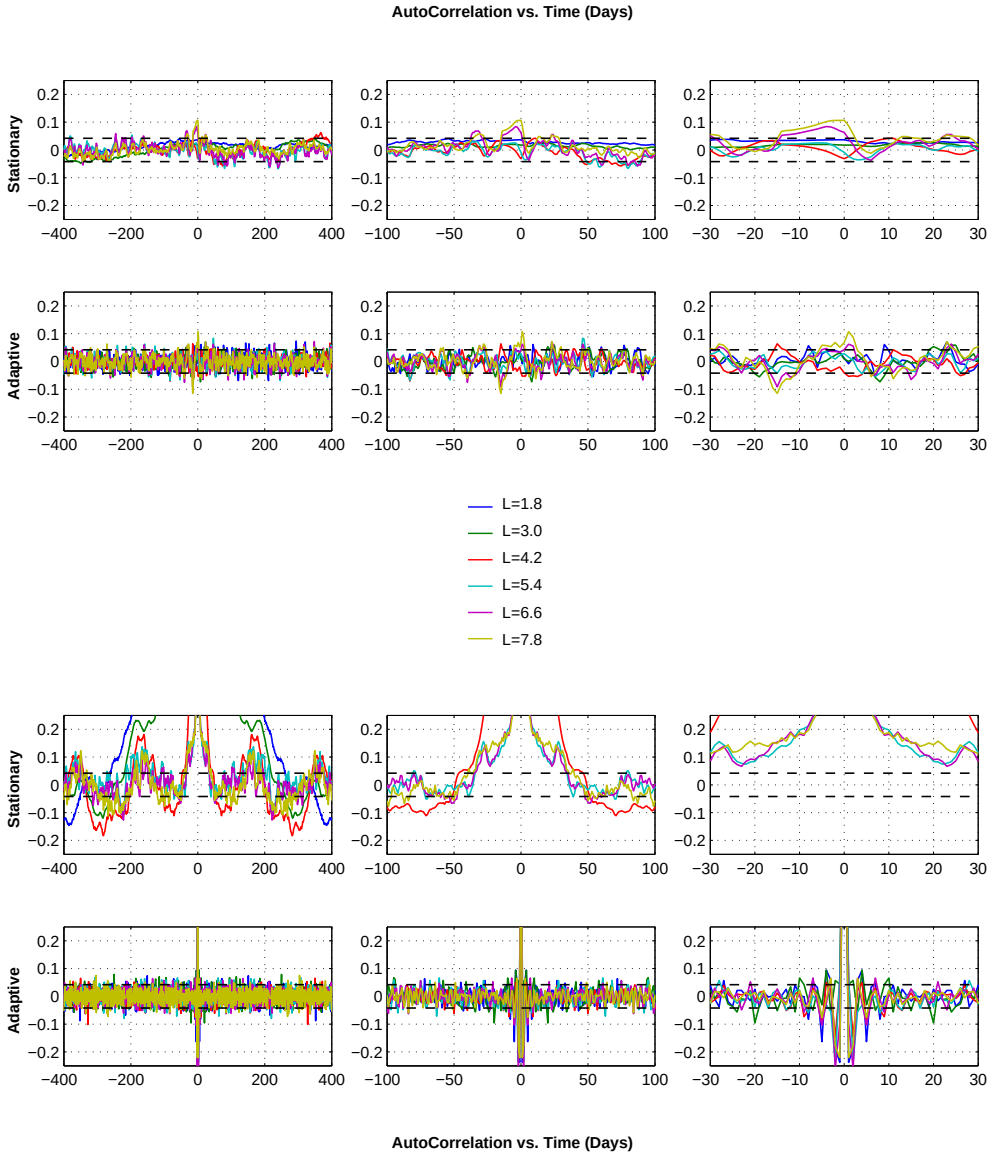


Figure B.17: Cross- and Auto-correlation functions of the $V_{sw} \rightarrow \log(\text{flux})$ residuals for stationary and adaptive predictions. These are essentially the same as the plots presented in Chapter 4, and simply provide additional "zoomed" views of the time-dependent correlation coefficients.

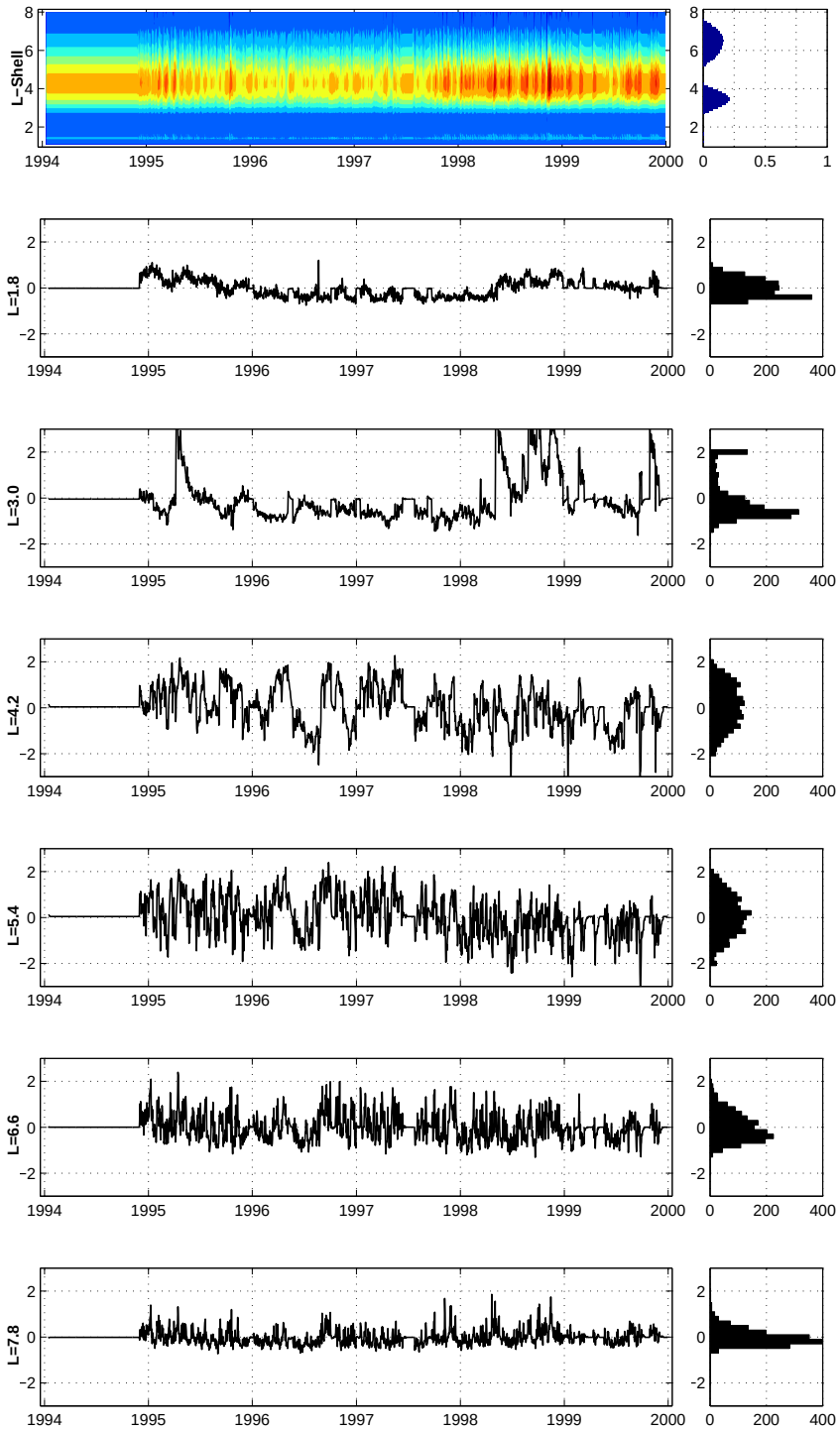


Figure B.18: Stationary $|\mathbf{B}| \rightarrow \log(\text{flux})$ predictions are plotted vs. L-shell, along with the resulting residuals and their histograms for L-shells representative of the inner electron radiation belt (L=1.8), the slot region (L=3.0), GPS orbits (L=4.2), the heart of the outer-belt (L=5.4), geostationary orbit (L=6.6), and the outer edge of the stable trapping region (L=7.8).

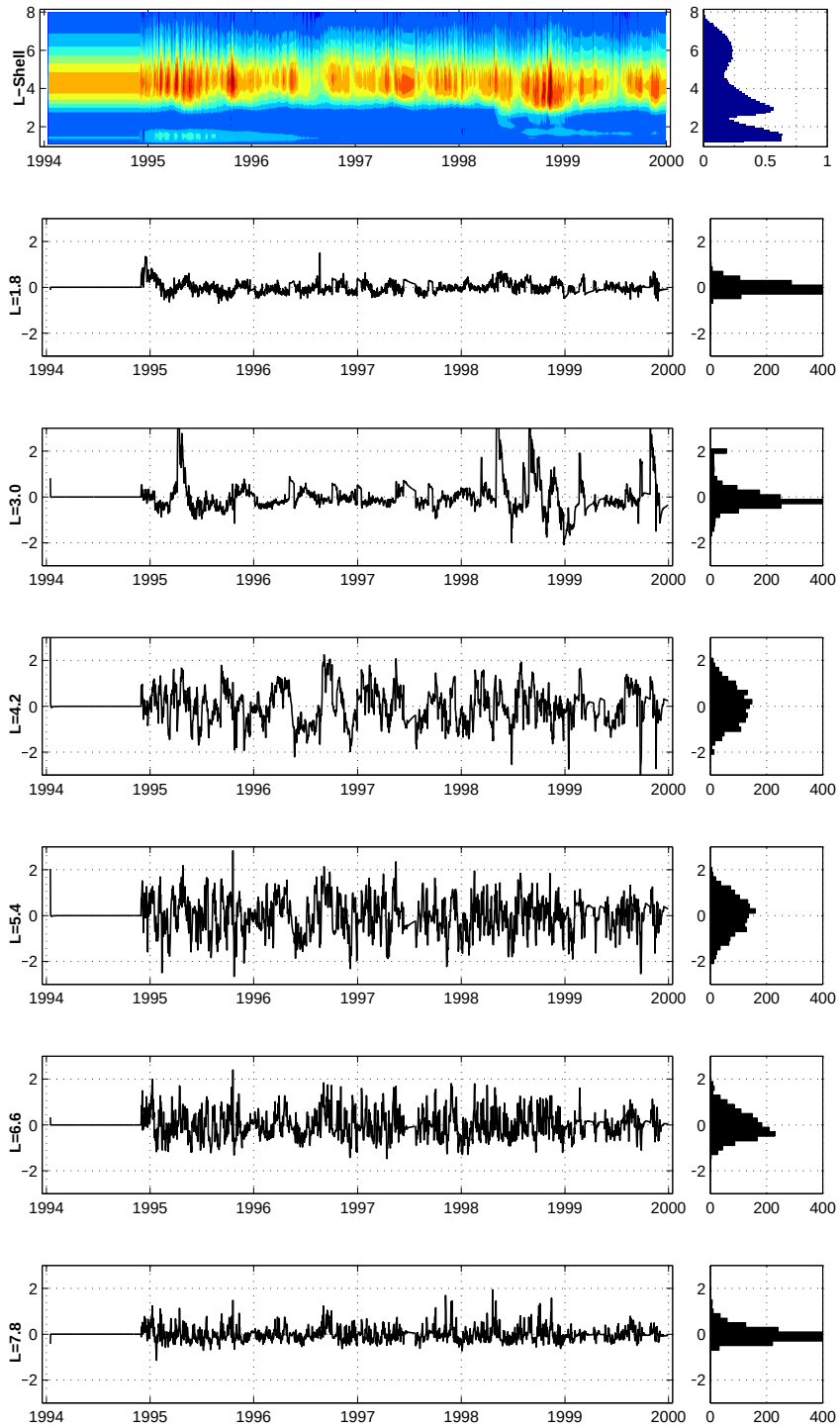


Figure B.19: Adaptive $|\mathbf{B}| \rightarrow \log(\text{flux})$ predictions are plotted vs. L-shell, along with the resulting residuals and their histograms for L-shells representative of the inner electron radiation belt (L=1.8), the slot region (L=3.0), GPS orbits (L=4.2), the heart of the outer-belt (L=5.4), geostationary orbit (L=6.6), and the outer edge of the stable trapping region (L=7.8).

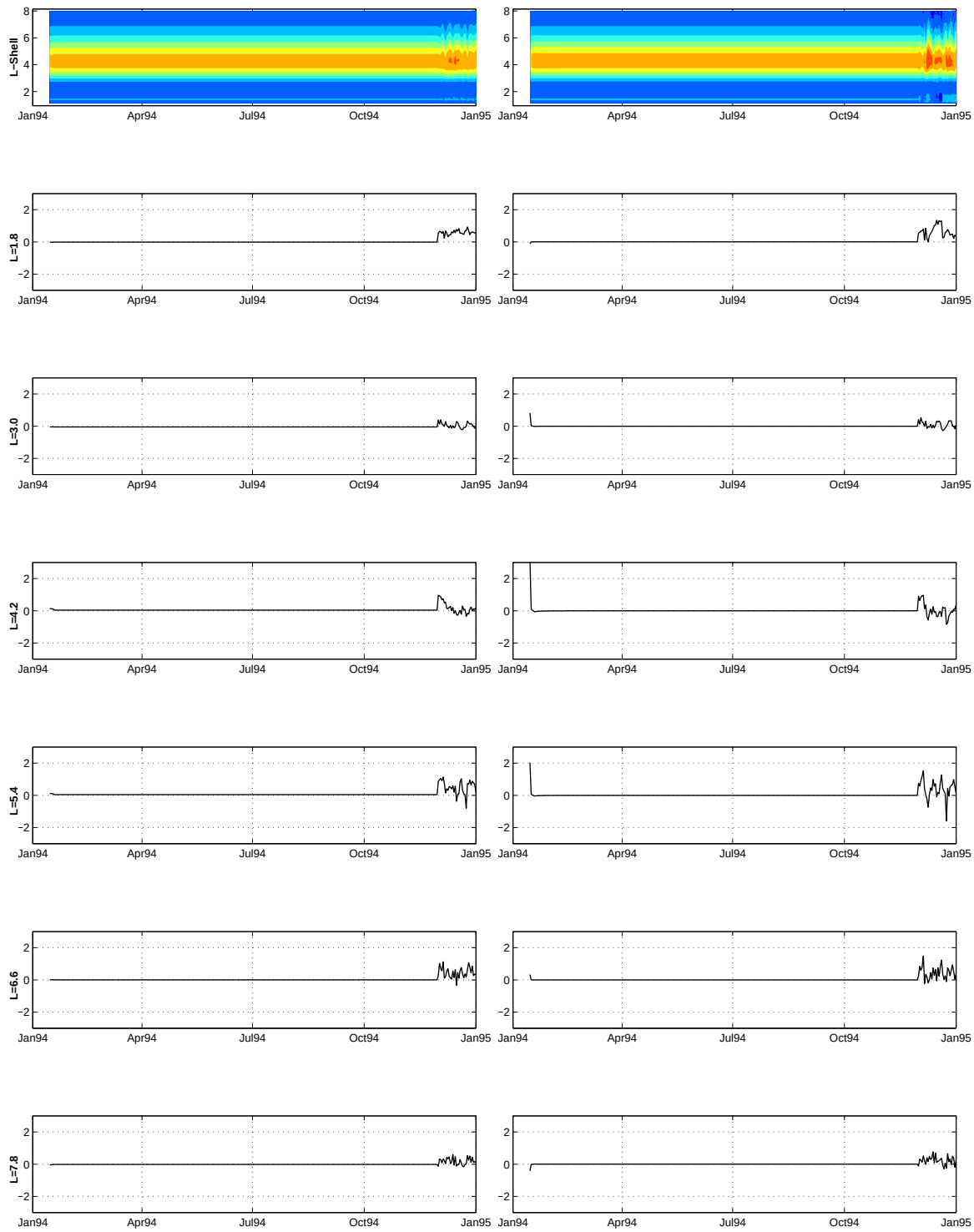


Figure B.20: Stationary (left) and Adaptive (right) $|\mathbf{B}| \rightarrow \log(\text{flux})$ predictions are plotted vs. L-shell, along with the resulting residuals for representative L-shells in 1994.

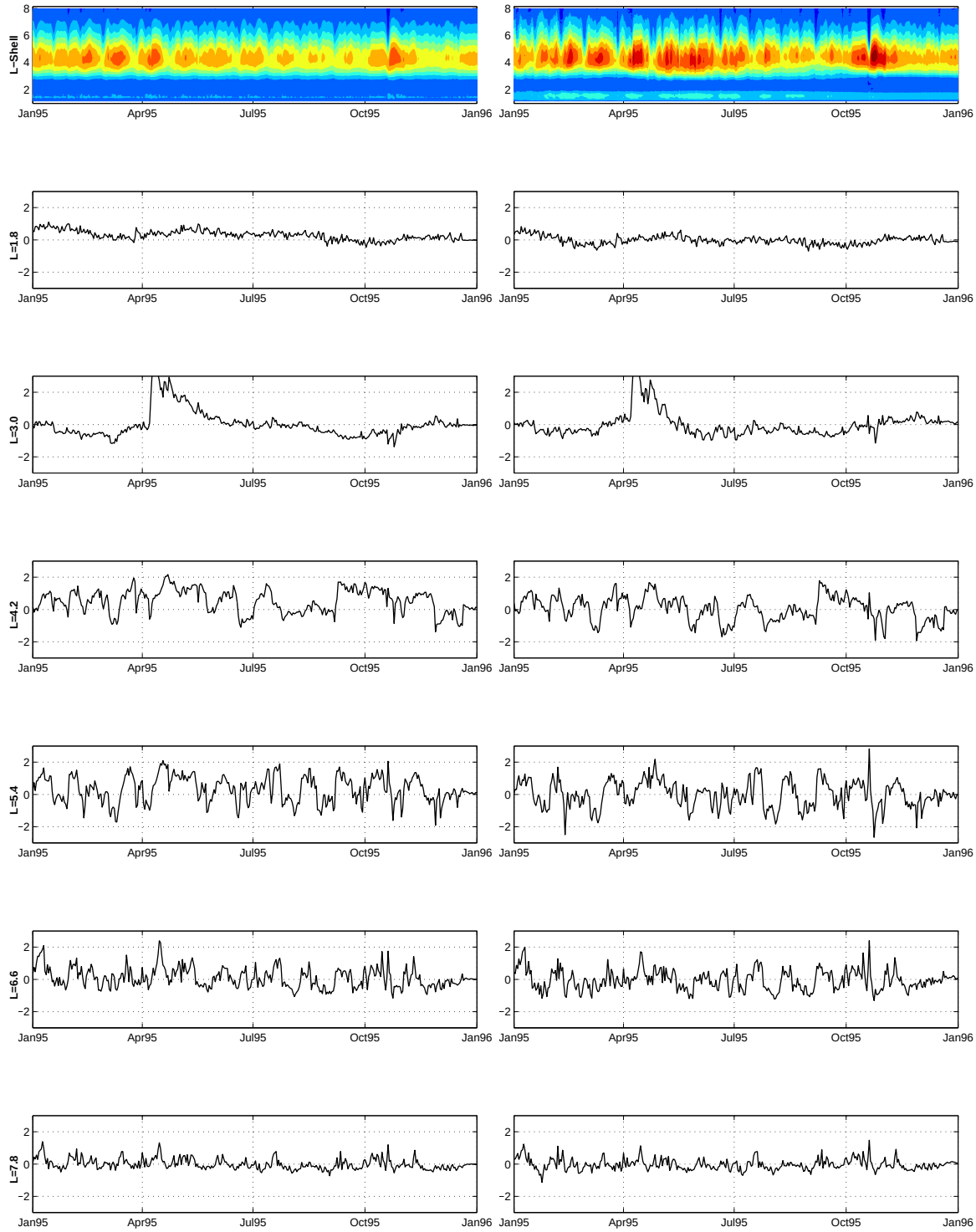


Figure B.21: Stationary (left) and Adaptive (right) $|\mathbf{B}| \rightarrow \log(\text{flux})$ predictions are plotted vs. L-shell, along with the resulting residuals for representative L-shells in 1995.

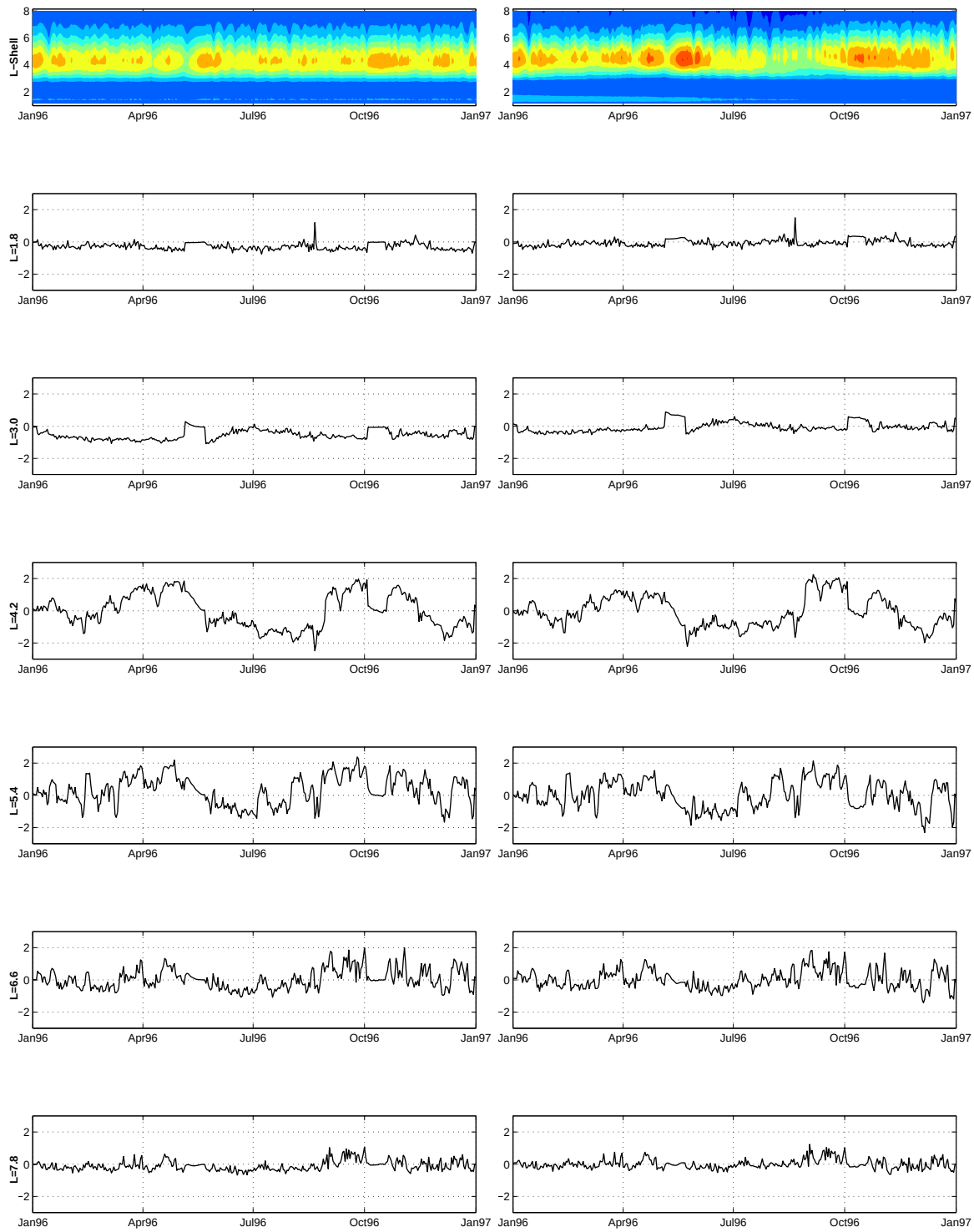


Figure B.22: Stationary (left) and Adaptive (right) $|\mathbf{B}| \rightarrow \log(\text{flux})$ predictions are plotted vs. L-shell, along with the resulting residuals for representative L-shells in 1996.

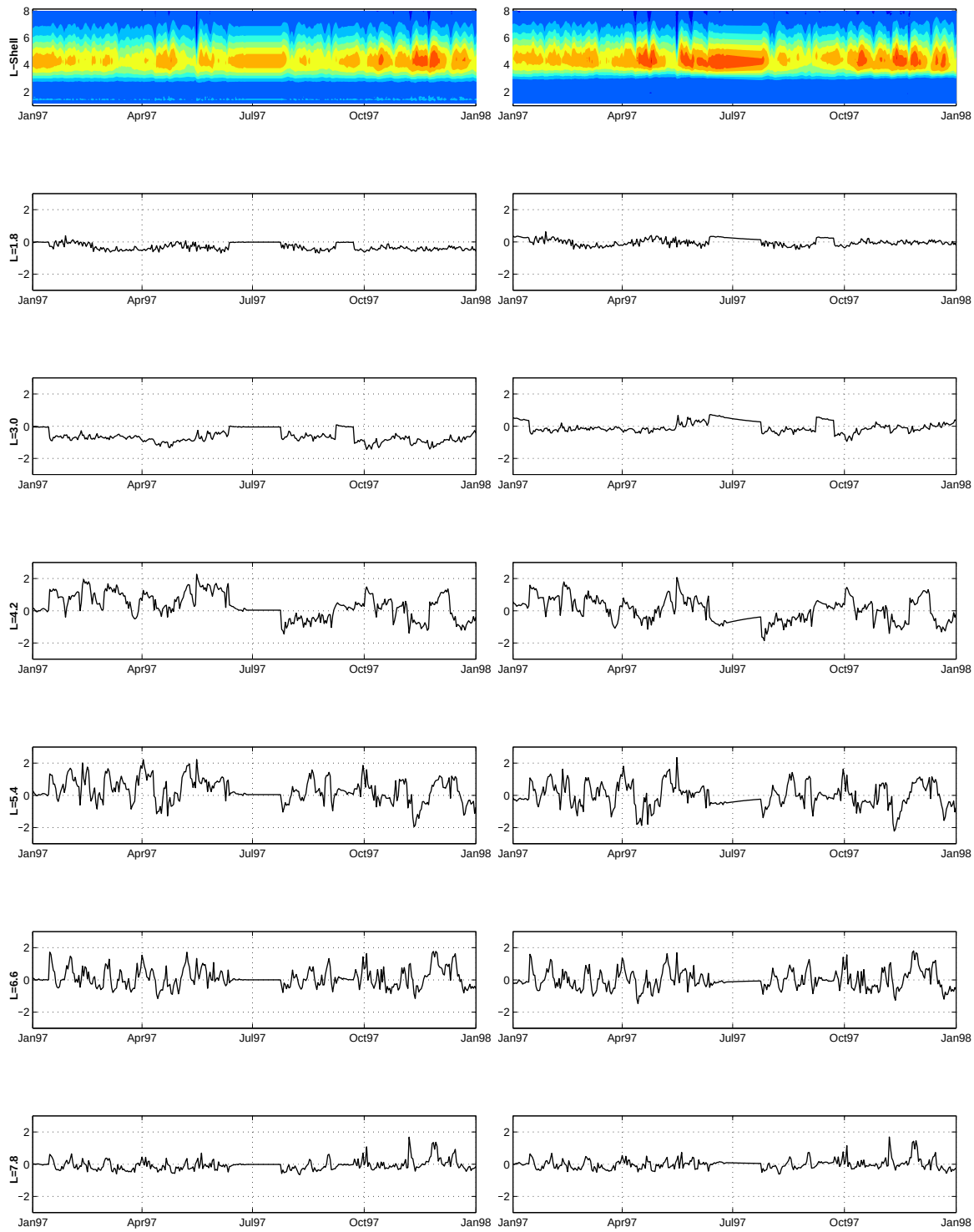


Figure B.23: Stationary (left) and Adaptive (right) $|\mathbf{B}| \rightarrow \log(\text{flux})$ predictions are plotted vs. L-shell, along with the resulting residuals for representative L-shells in 1997.

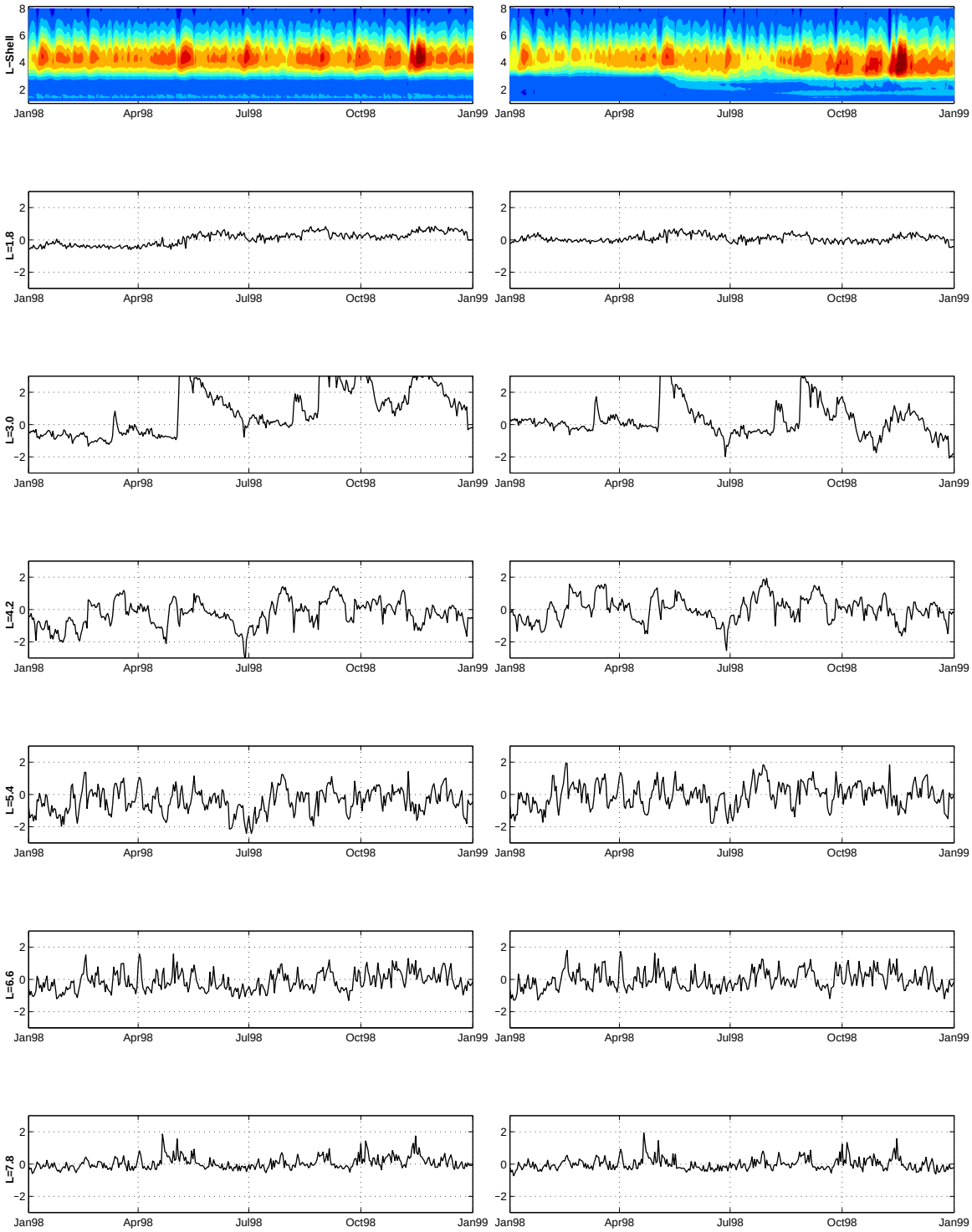


Figure B.24: Stationary (left) and Adaptive (right) $|\mathbf{B}| \rightarrow \log(\text{flux})$ predictions are plotted vs. L-shell, along with the resulting residuals for representative L-shells in 1998.

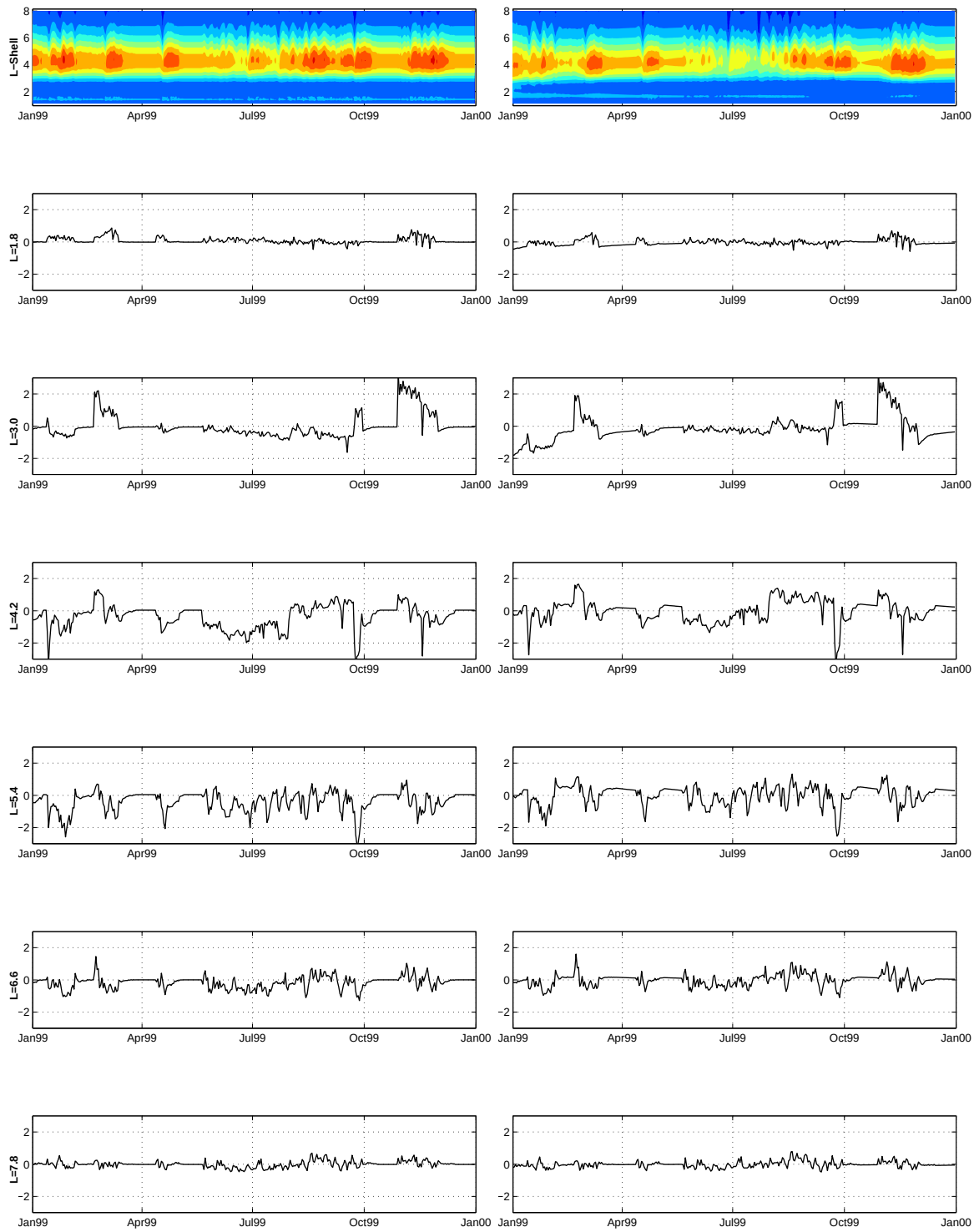


Figure B.25: Stationary (left) and Adaptive (right) $|\mathbf{B}| \rightarrow \log(\text{flux})$ predictions are plotted vs. L-shell, along with the resulting residuals for representative L-shells in 1999.

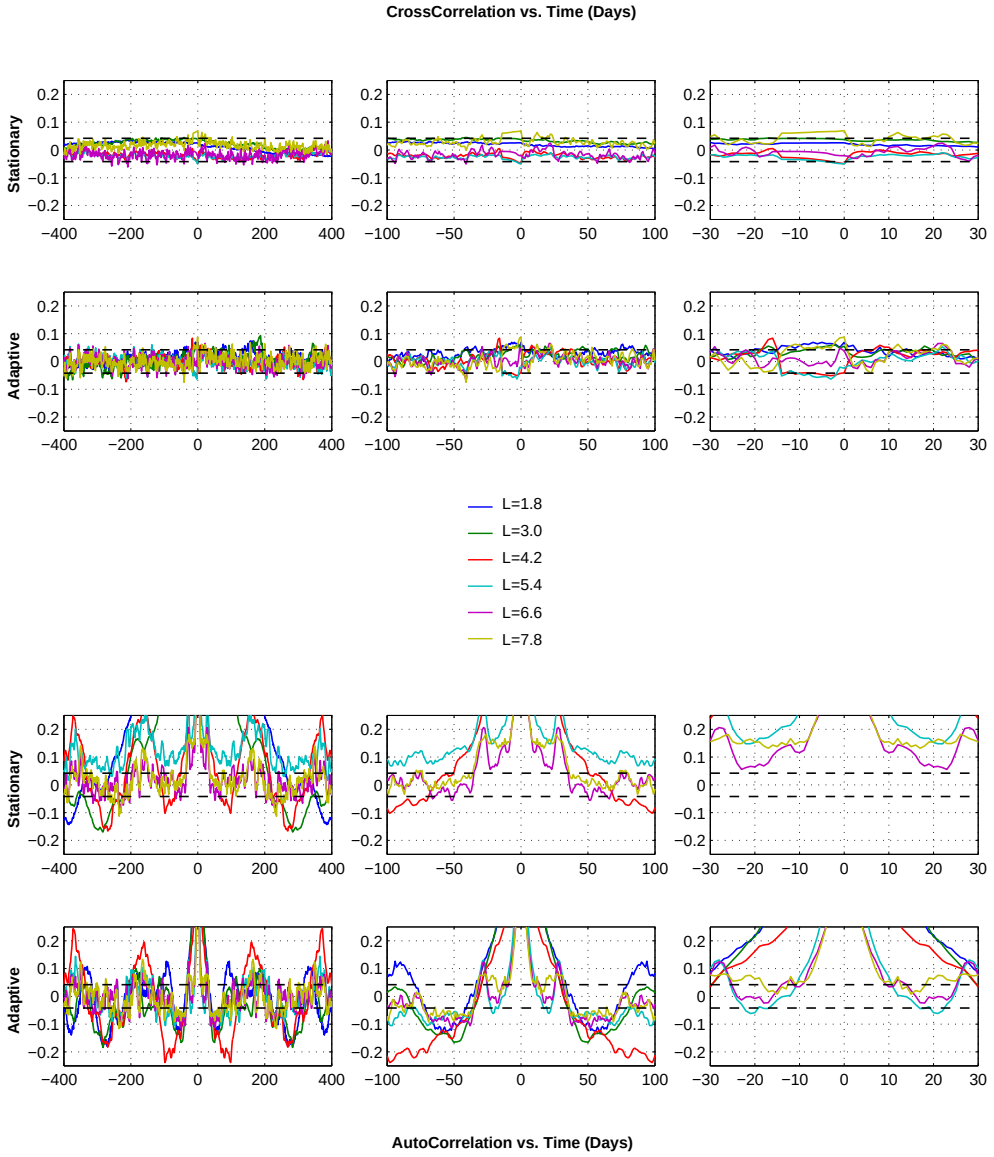


Figure B.26: Cross- and Auto-correlation functions of the $|\mathbf{B}| \rightarrow \log(\text{flux})$ residuals for stationary and adaptive predictions. These are essentially the same as the plots presented in Chapter 4, and simply provide additional "zoomed" views of the time-dependent correlation coefficients.

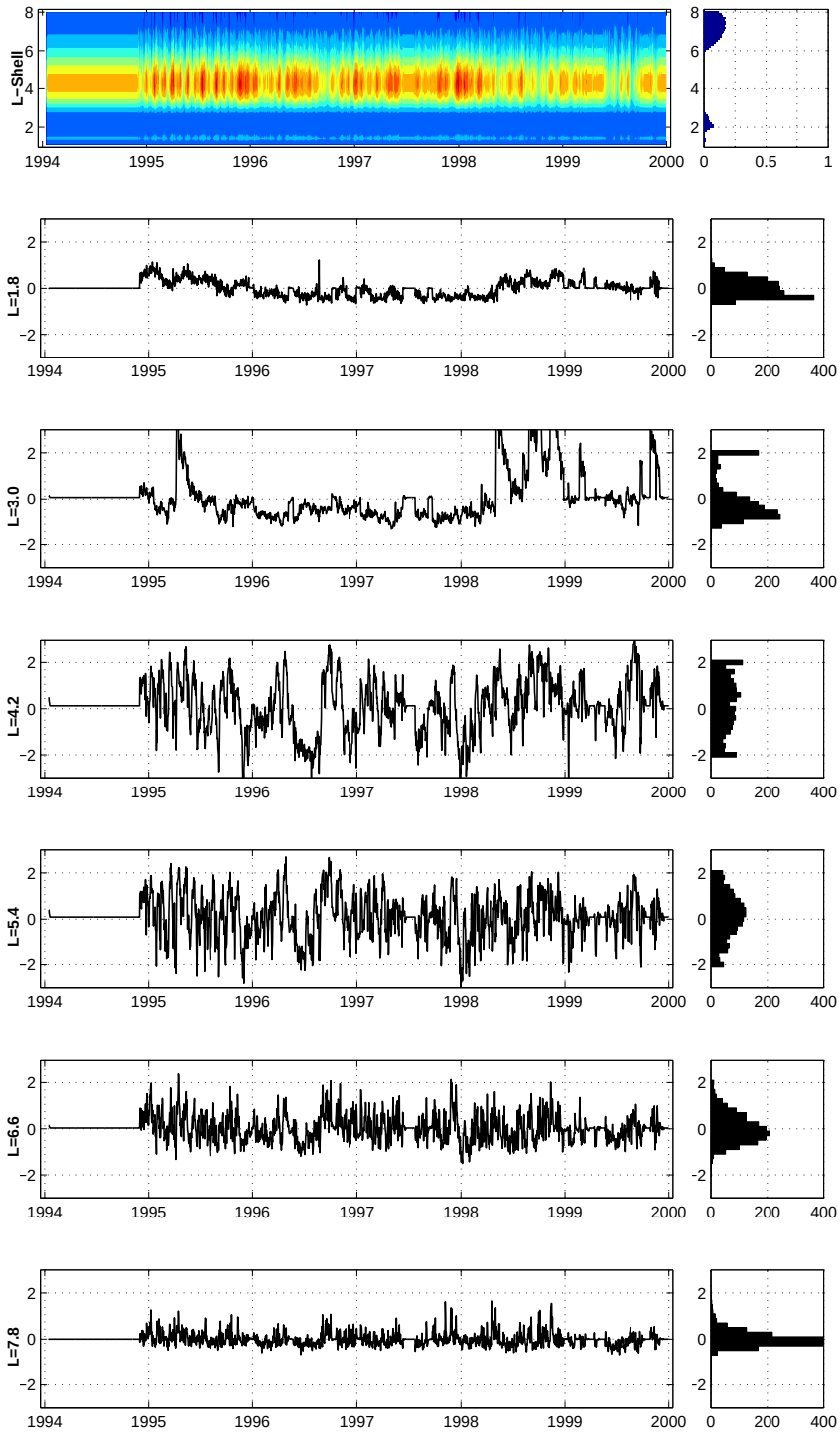


Figure B.27: Stationary $\rho_{ion} \rightarrow \log(\text{flux})$ predictions are plotted vs. L-shell, along with the resulting residuals and their histograms for L-shells representative of the inner electron radiation belt (L=1.8), the slot region (L=3.0), GPS orbits (L=4.2), the heart of the outer-belt (L=5.4), geostationary orbit (L=6.6), and the outer edge of the stable trapping region (L=7.8).

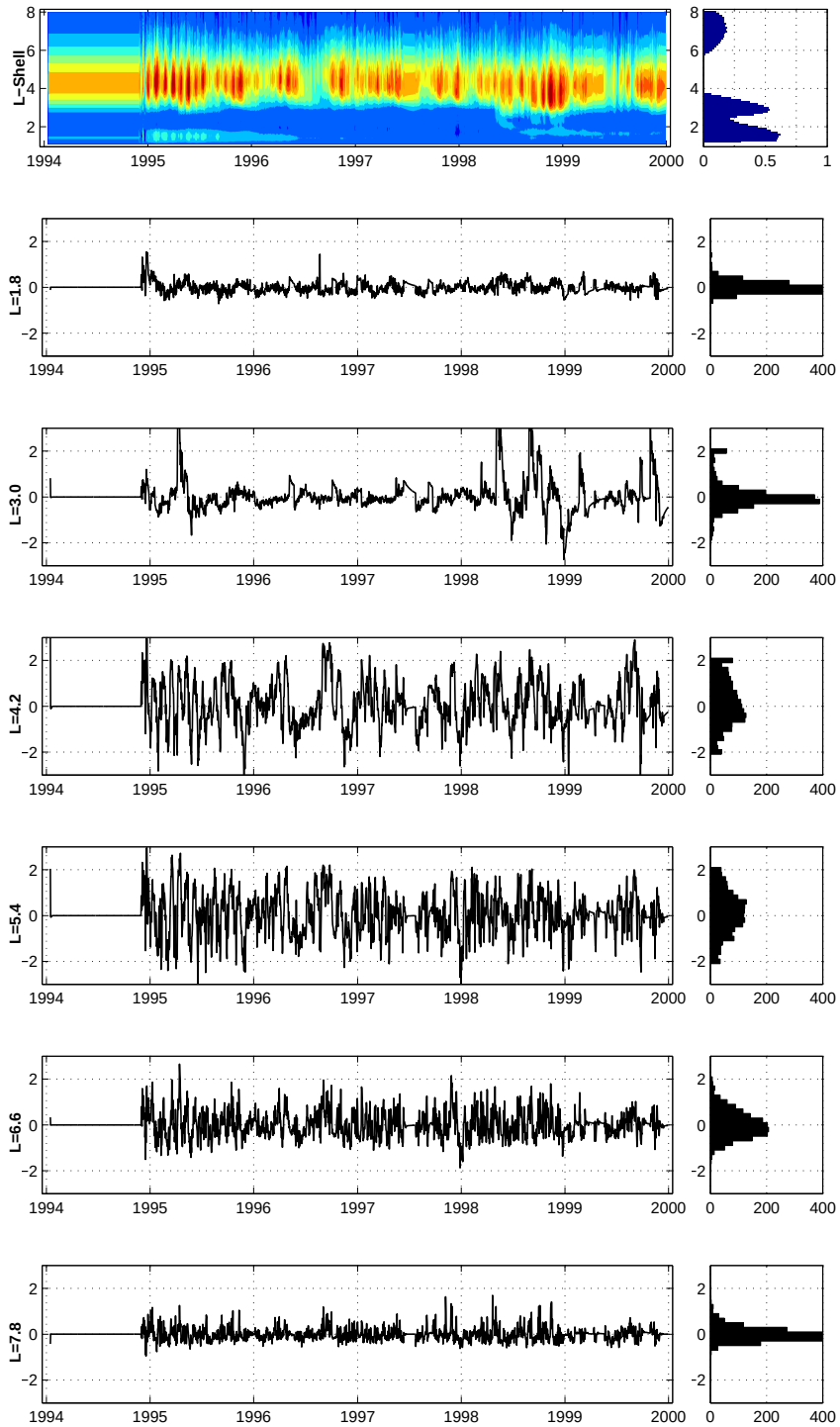


Figure B.28: Adaptive $\rho_{ion} \rightarrow \log(\text{flux})$ predictions are plotted vs. L-shell, along with the resulting residuals and their histograms for L-shells representative of the inner electron radiation belt (L=1.8), the slot region (L=3.0), GPS orbits (L=4.2), the heart of the outer-belt (L=5.4), geostationary orbit (L=6.6), and the outer edge of the stable trapping region (L=7.8).

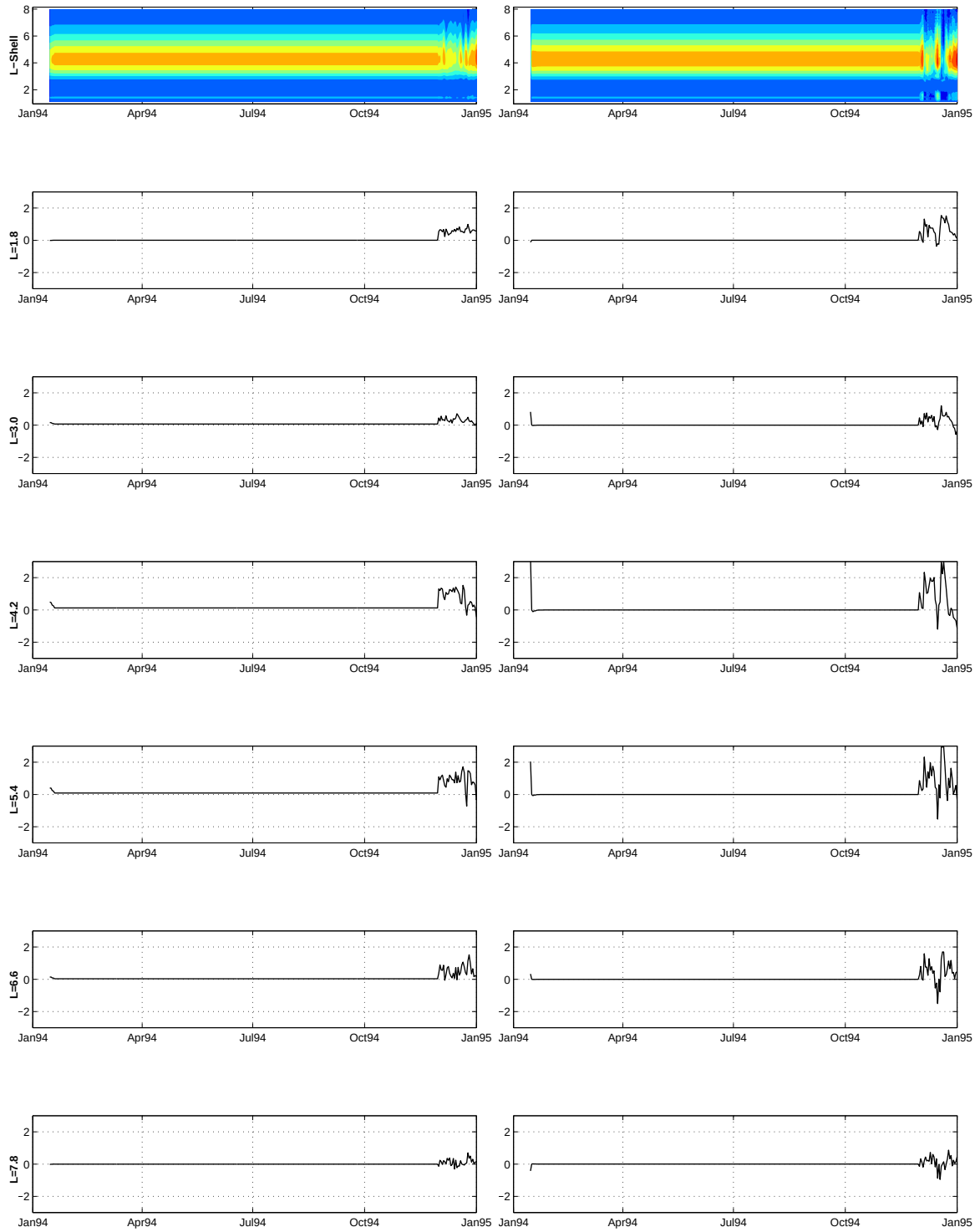


Figure B.29: Stationary (left) and Adaptive (right) $\rho_{ion} \rightarrow \log(\text{flux})$ predictions are plotted vs. L-shell, along with the resulting residuals for representative L-shells in 1994.

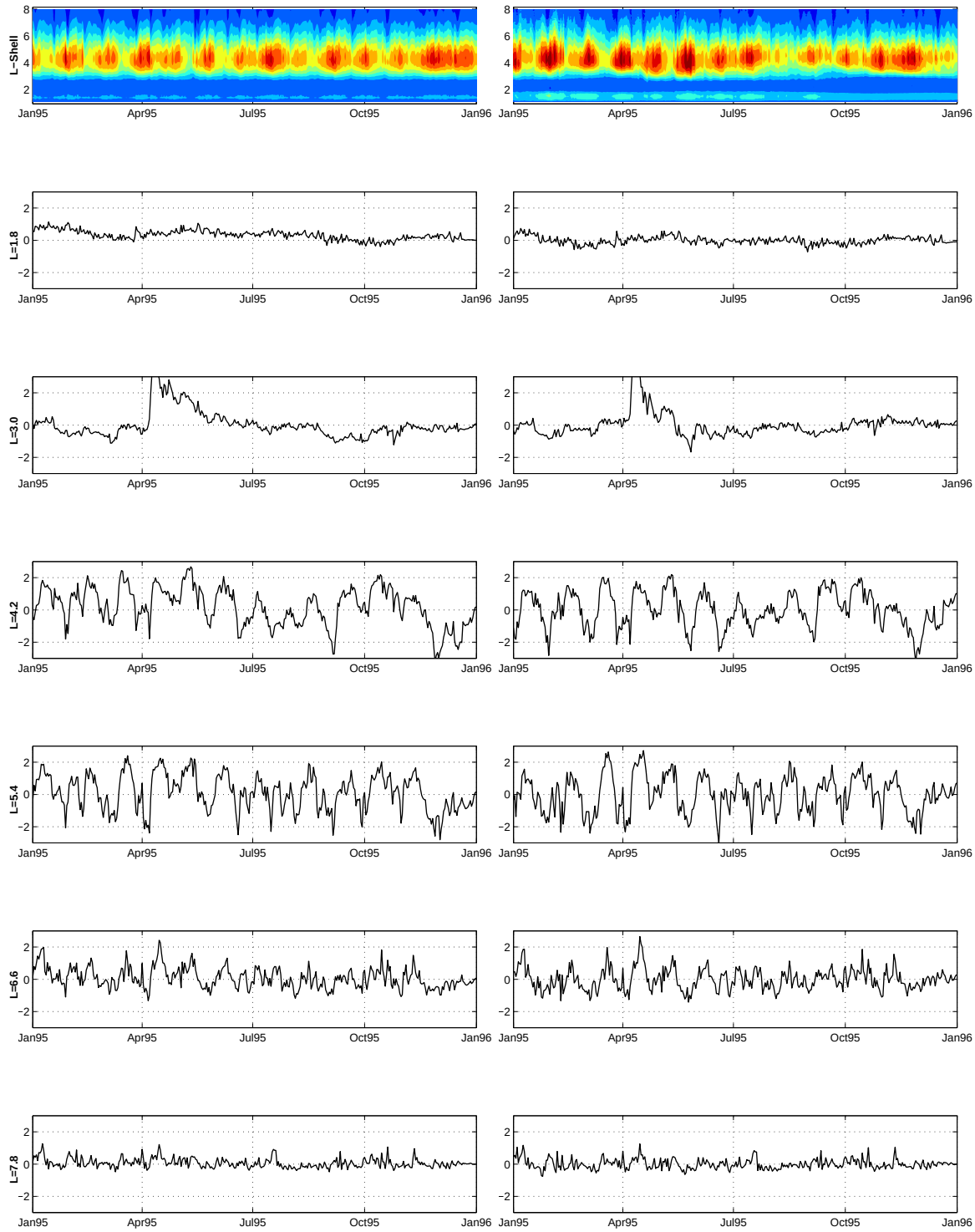


Figure B.30: Stationary (left) and Adaptive (right) $\rho_{ion} \rightarrow \log(\text{flux})$ predictions are plotted vs. L-shell, along with the resulting residuals for representative L-shells in 1995.

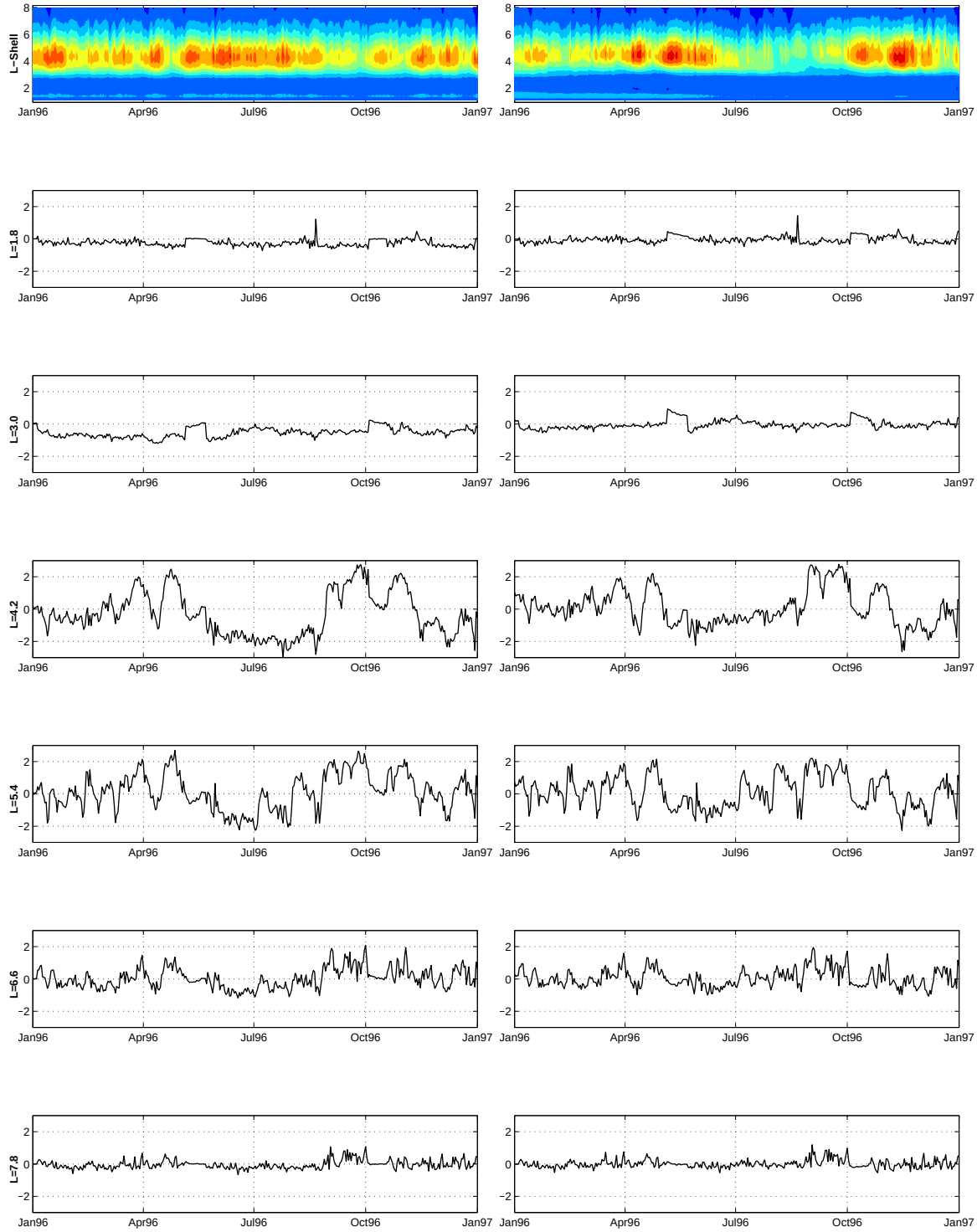


Figure B.31: Stationary (left) and Adaptive (right) $\rho_{ion} \rightarrow \log(\text{flux})$ predictions are plotted vs. L-shell, along with the resulting residuals for representative L-shells in 1996.

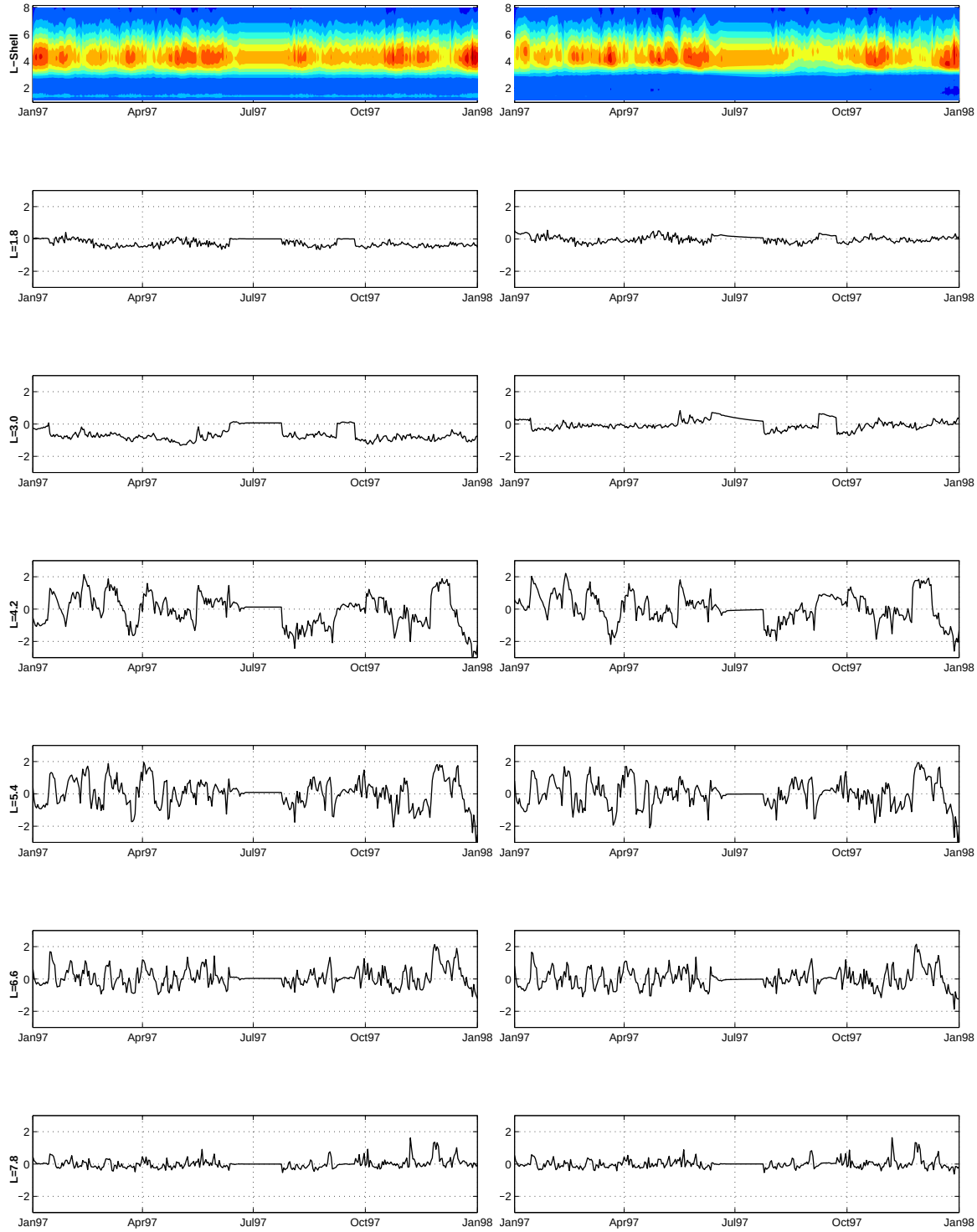


Figure B.32: Stationary (left) and Adaptive (right) $\rho_{ion} \rightarrow \log(\text{flux})$ predictions are plotted vs. L-shell, along with the resulting residuals for representative L-shells in 1997.

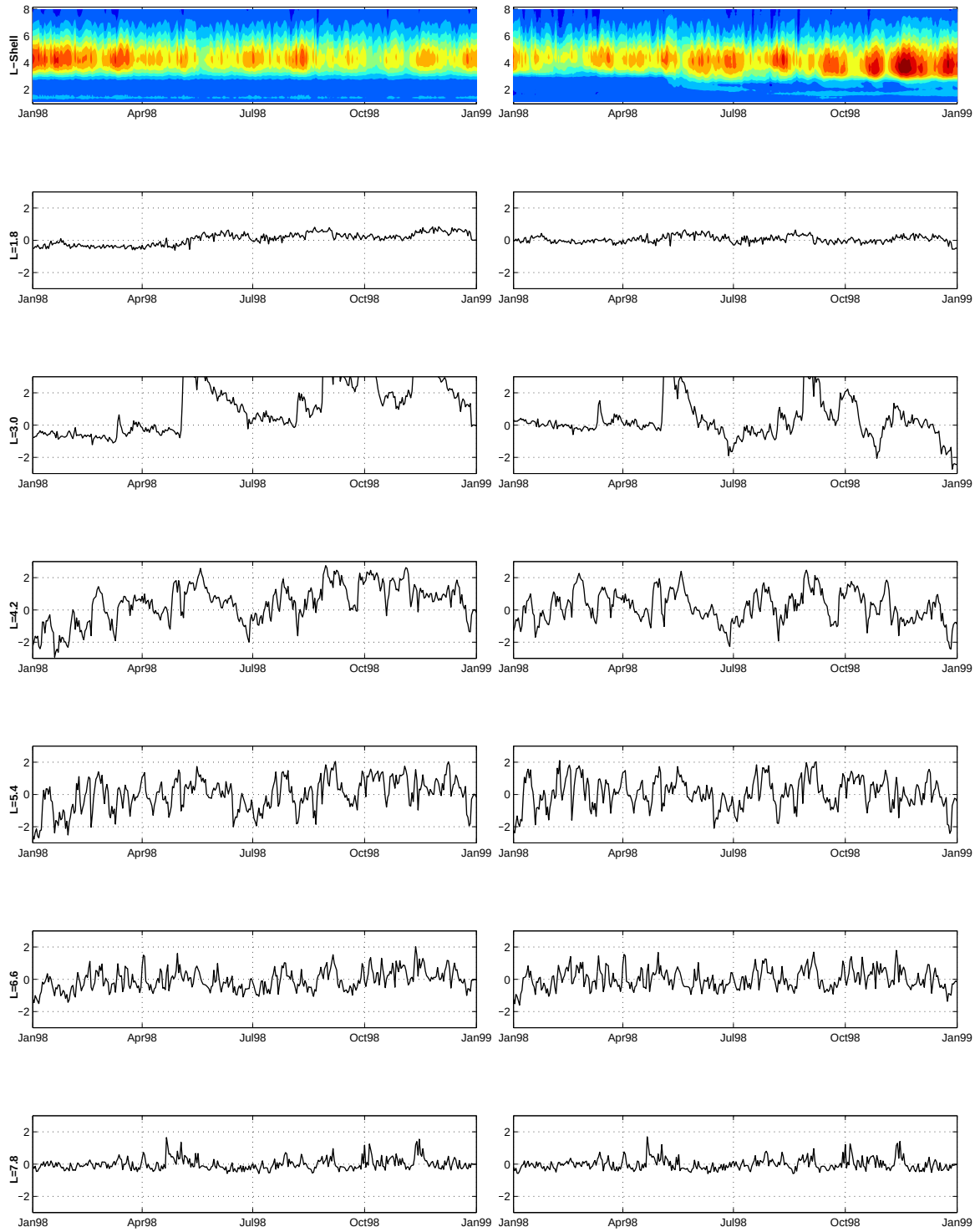


Figure B.33: Stationary (left) and Adaptive (right) $\rho_{ion} \rightarrow \log(\text{flux})$ predictions are plotted vs. L-shell, along with the resulting residuals for representative L-shells in 1998.

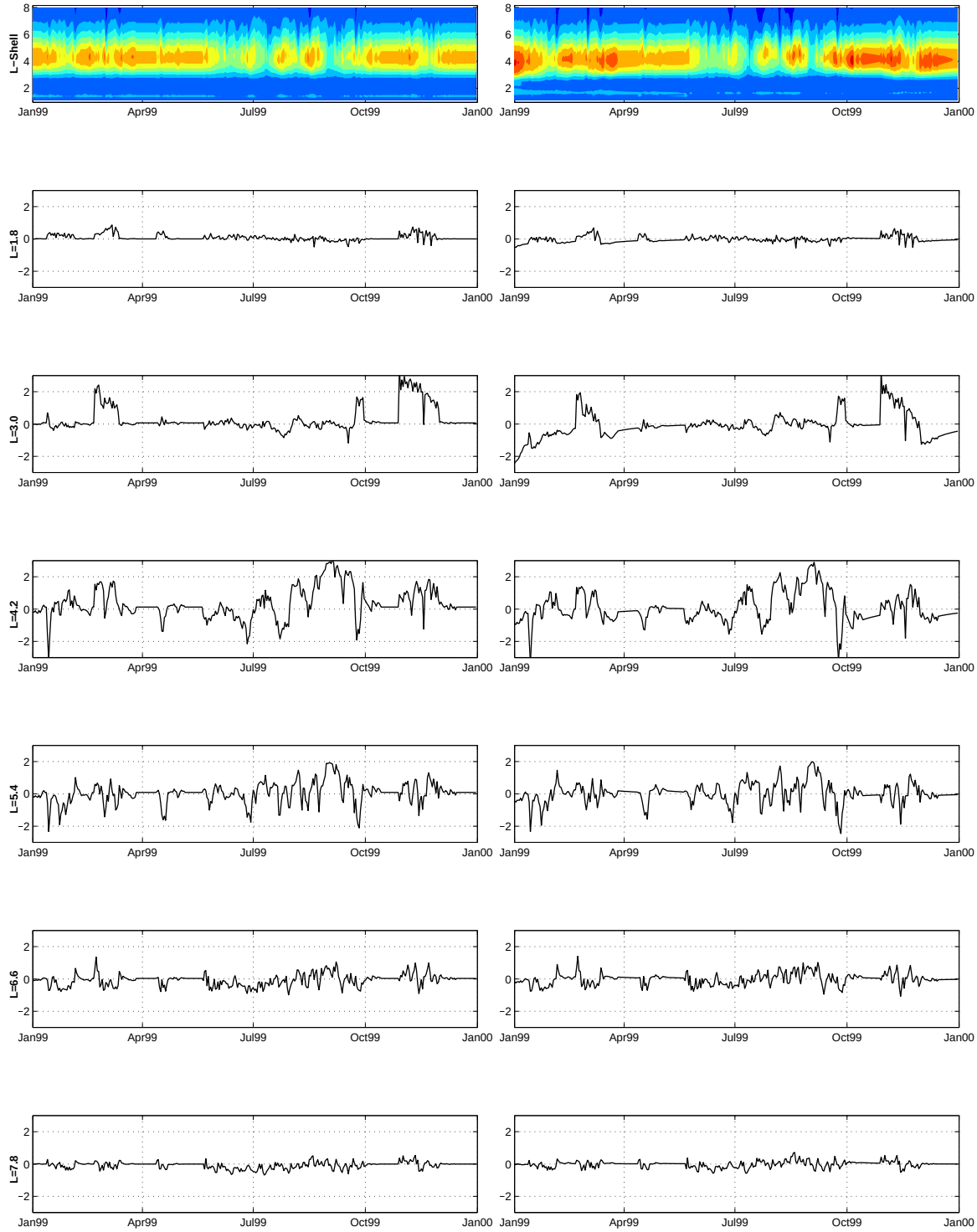


Figure B.34: Stationary (left) and Adaptive (right) $\rho_{ion} \rightarrow \log(\text{flux})$ predictions are plotted vs. L-shell, along with the resulting residuals for representative L-shells in 1999.

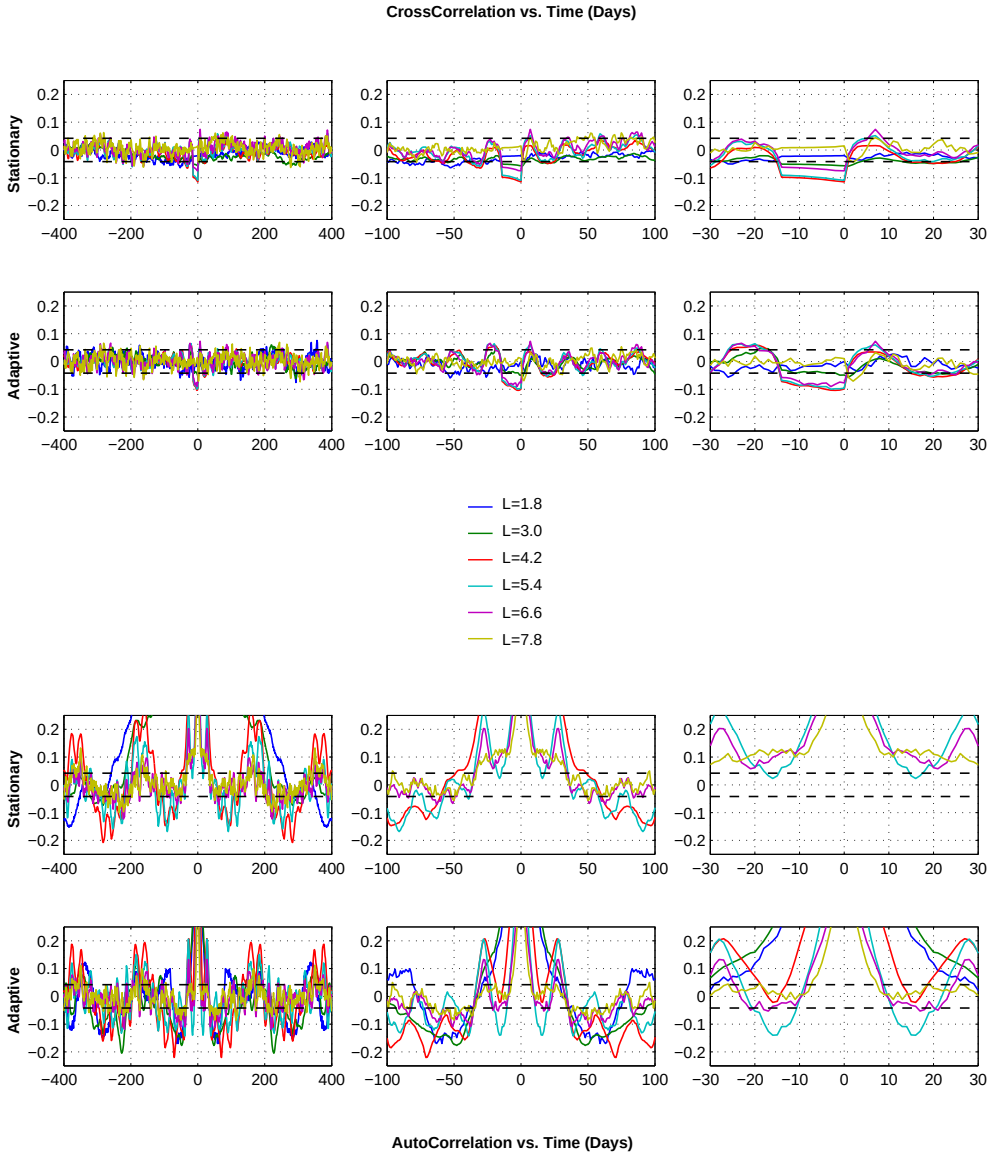


Figure B.35: Cross- and Auto-correlation functions of the $\rho_{ion} \rightarrow \log(\text{flux})$ residuals for stationary and adaptive predictions. These are essentially the same as the plots presented in Chapter 4, and simply provide additional "zoomed" views of the time-dependent correlation coefficients.

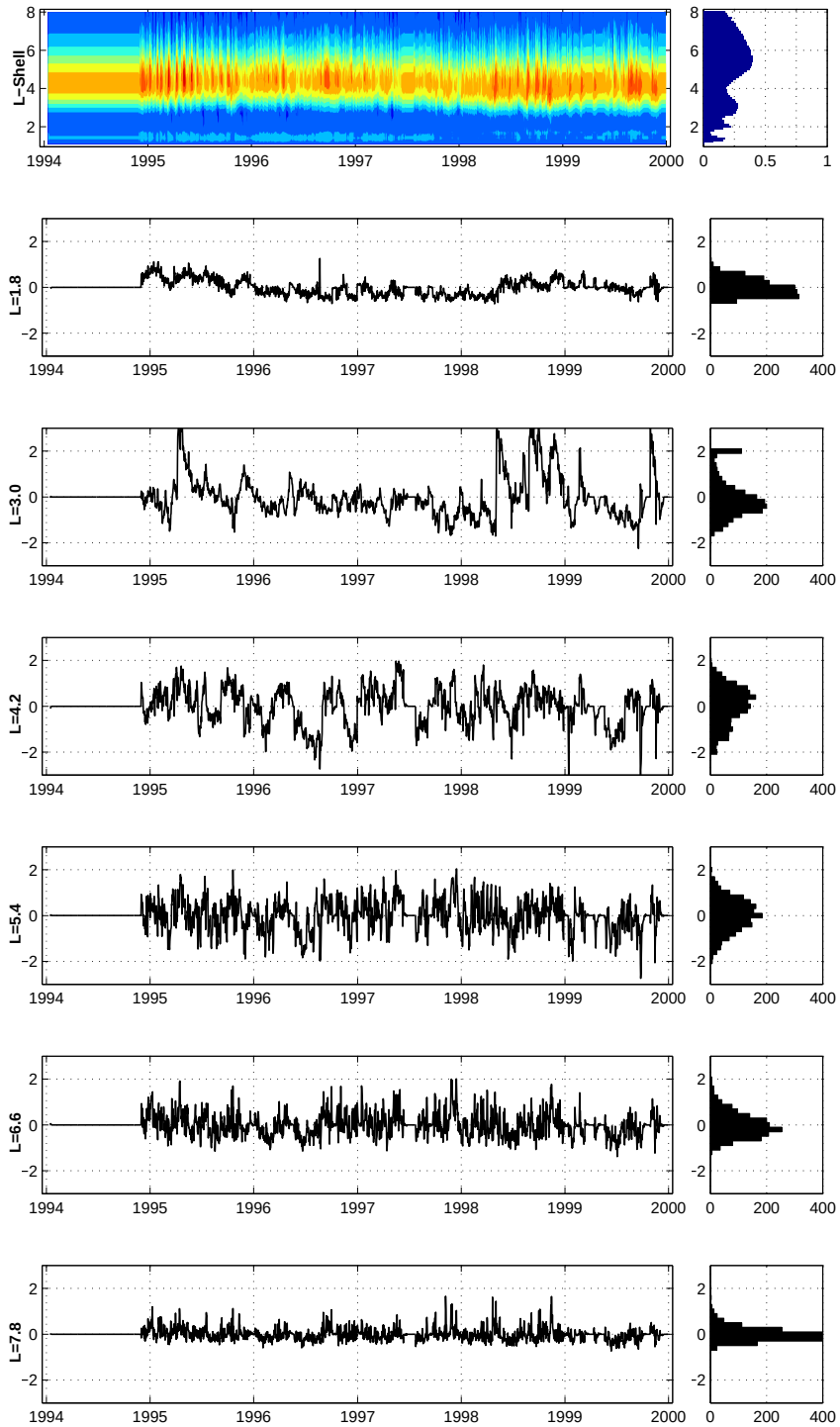


Figure B.36: Stationary $[V_{sw}, |\mathbf{B}|, \rho_{ion}] \rightarrow \log(\text{flux})$ predictions are plotted vs. L-shell, along with the resulting residuals and their histograms for L-shells representative of the inner electron radiation belt (L=1.8), the slot region (L=3.0), GPS orbits (L=4.2), the heart of the outer-belt (L=5.4), geostationary orbit (L=6.6), and the outer edge of the stable trapping region (L=7.8).

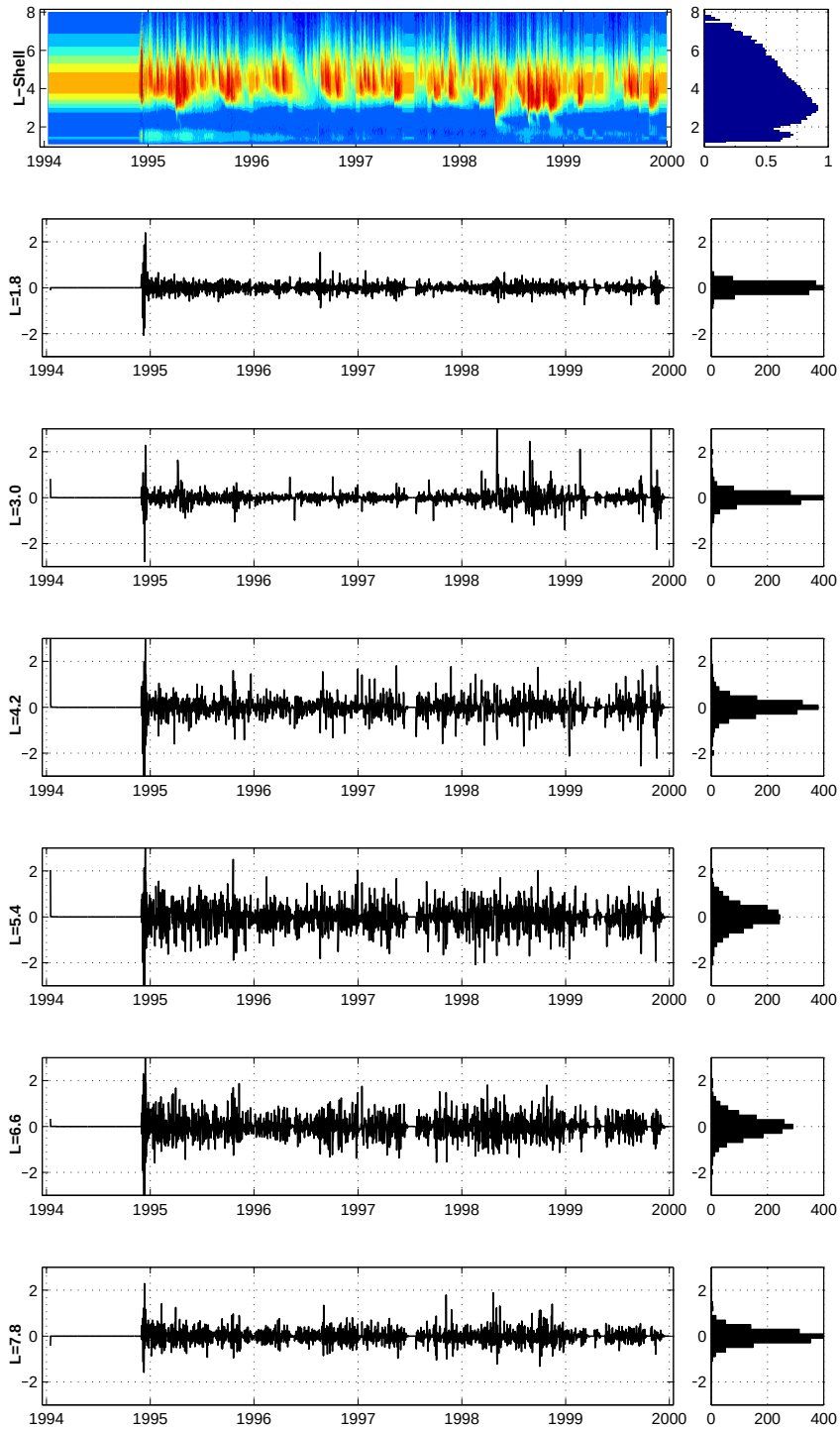


Figure B.37: Adaptive $[V_{sw}, |\mathbf{B}|, \rho_{ion}] \rightarrow \log(\text{flux})$ predictions are plotted vs. L-shell, along with the resulting residuals and their histograms for L-shells representative of the inner electron radiation belt (L=1.8), the slot region (L=3.0), GPS orbits (L=4.2), the heart of the outer-belt (L=5.4), geostationary orbit (L=6.6), and the outer edge of the stable trapping region (L=7.8).

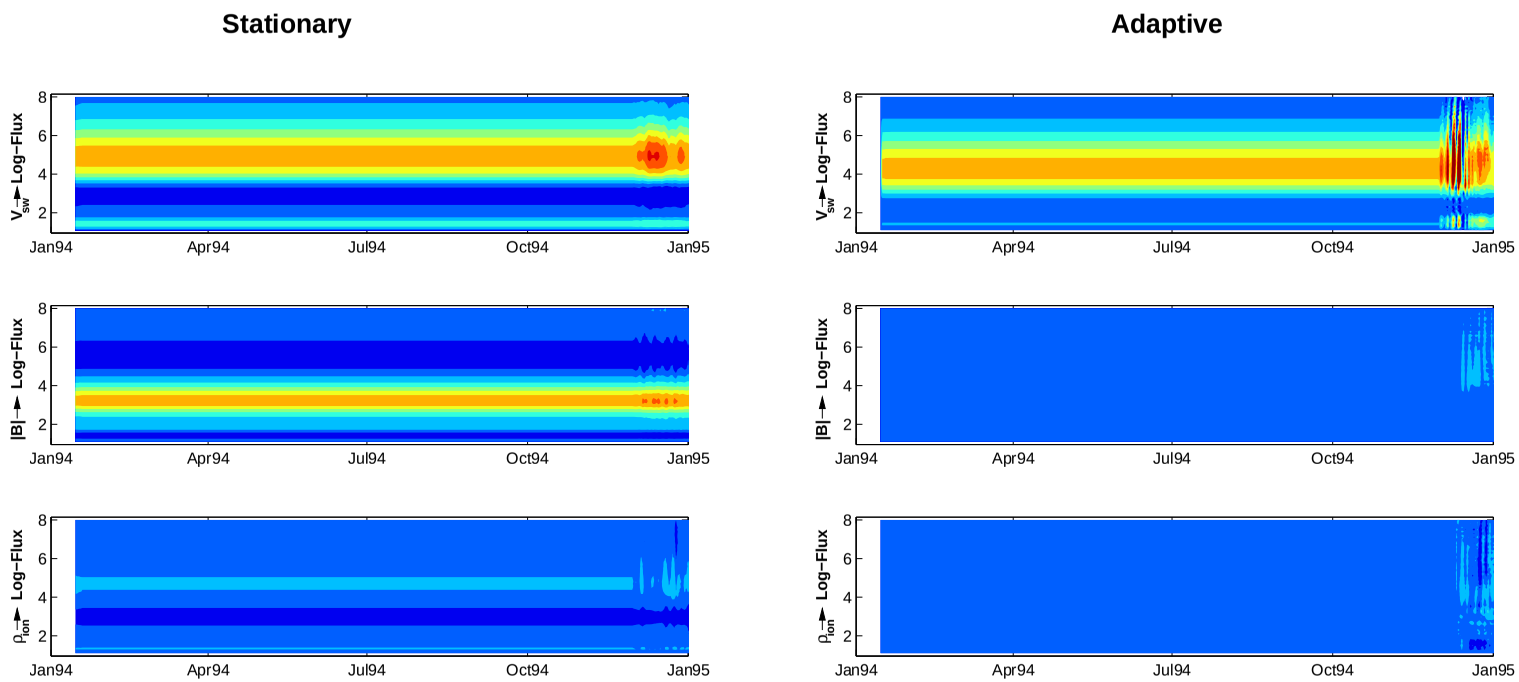


Figure B.38: Stationary and Adaptive $[V_{sw}, |\mathbf{B}|, \rho_{ion}] \rightarrow \log(\text{flux})$ component predictions are plotted vs. L-shell in 1994.

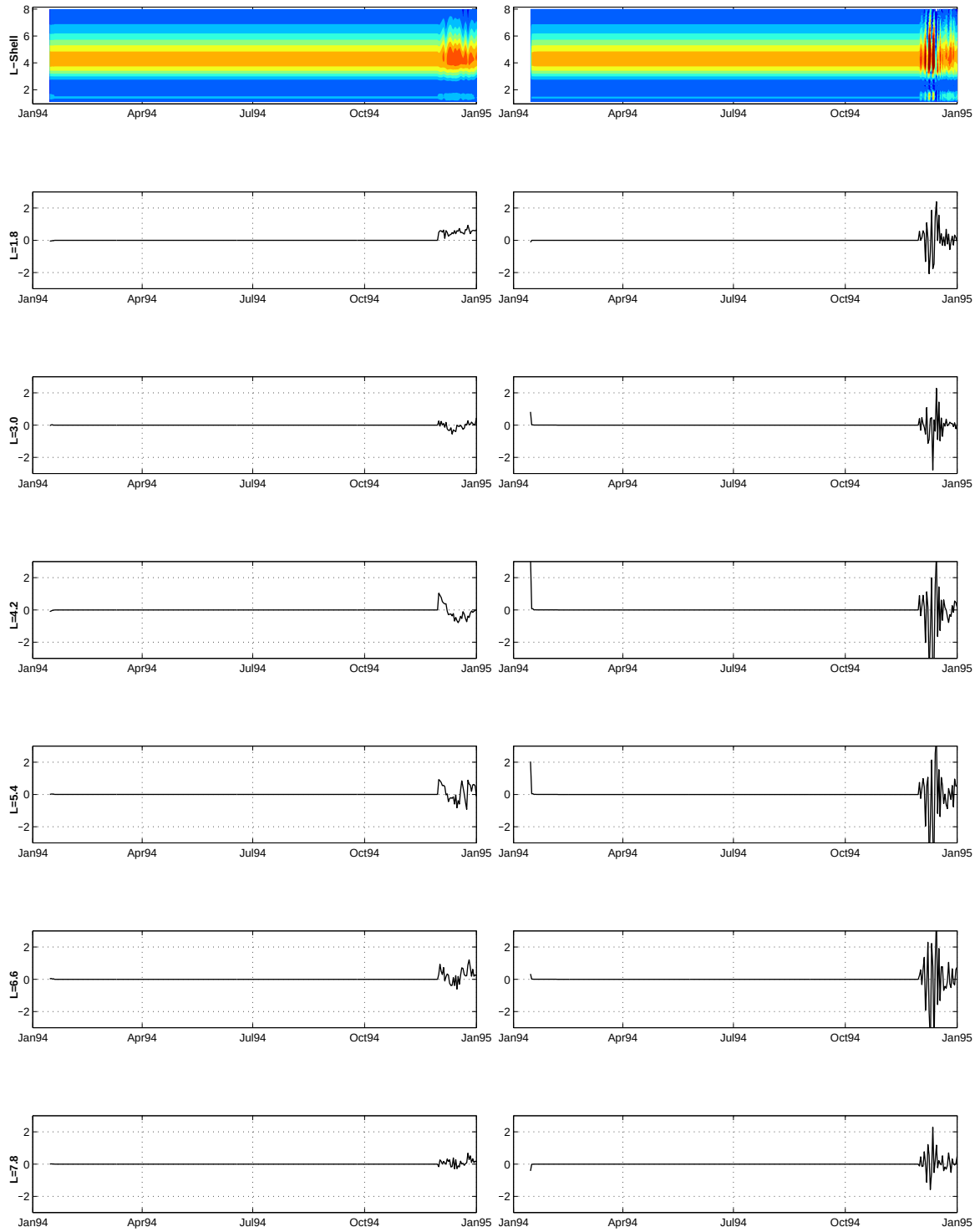


Figure B.39: Stationary (left) and Adaptive (right) $[V_{sw}, |\mathbf{B}|, \rho_{ion}] \rightarrow \log(\text{flux})$ total predictions are plotted vs. L-shell, along with the resulting residuals for representative L-shells in 1994.

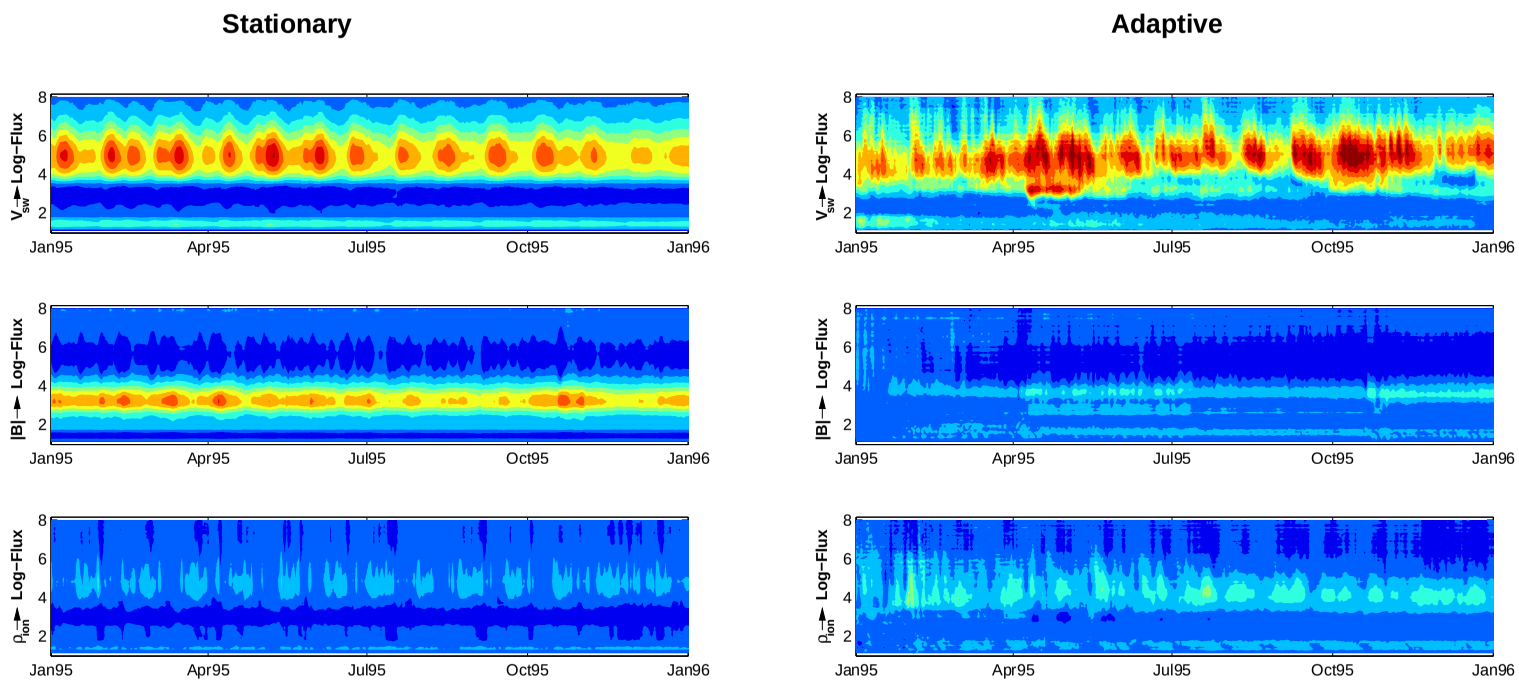


Figure B.40: Stationary and Adaptive $[V_{sw}, |\mathbf{B}|, \rho_{ion}] \rightarrow \log(\text{flux})$ component predictions are plotted vs. L-shell in 1995.

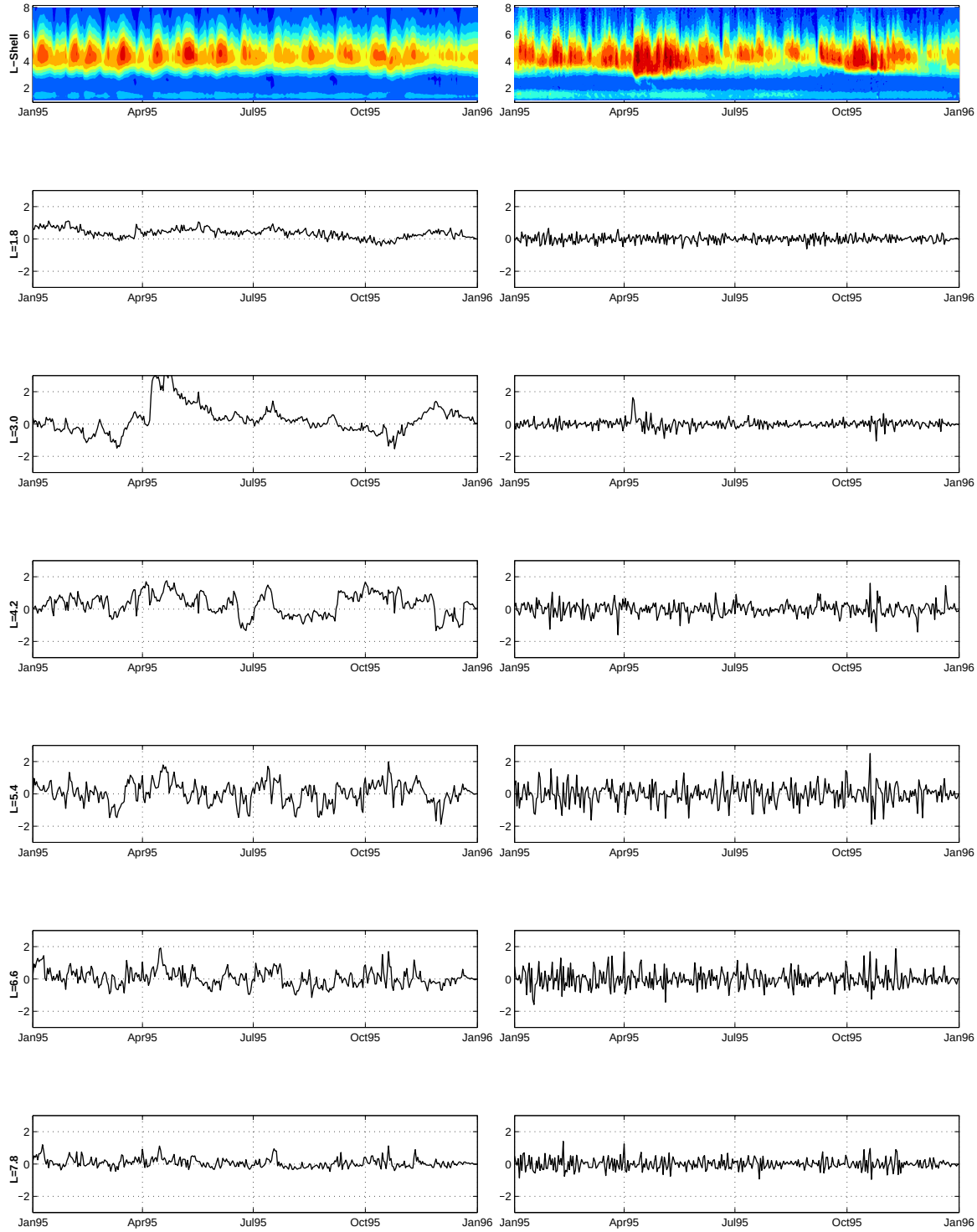


Figure B.41: Stationary (left) and Adaptive (right) $[V_{sw}, |\mathbf{B}|, \rho_{ion}] \rightarrow \log(\text{flux})$ predictions are plotted vs. L-shell, along with the resulting residuals for representative L-shells in 1995.

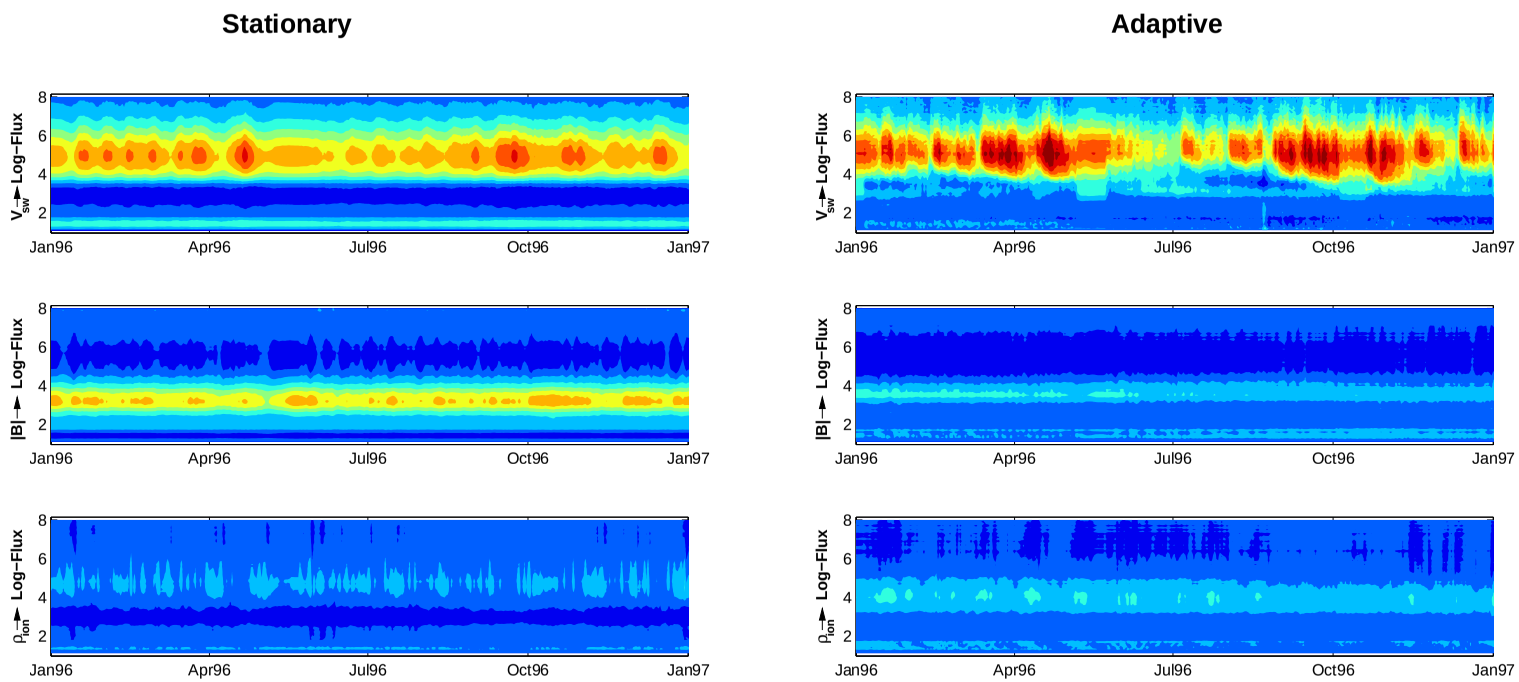


Figure B.42: Stationary and Adaptive $[V_{sw}, |\mathbf{B}|, \rho_{ion}] \rightarrow \log(\text{flux})$ component predictions are plotted vs. L-shell in 1996.

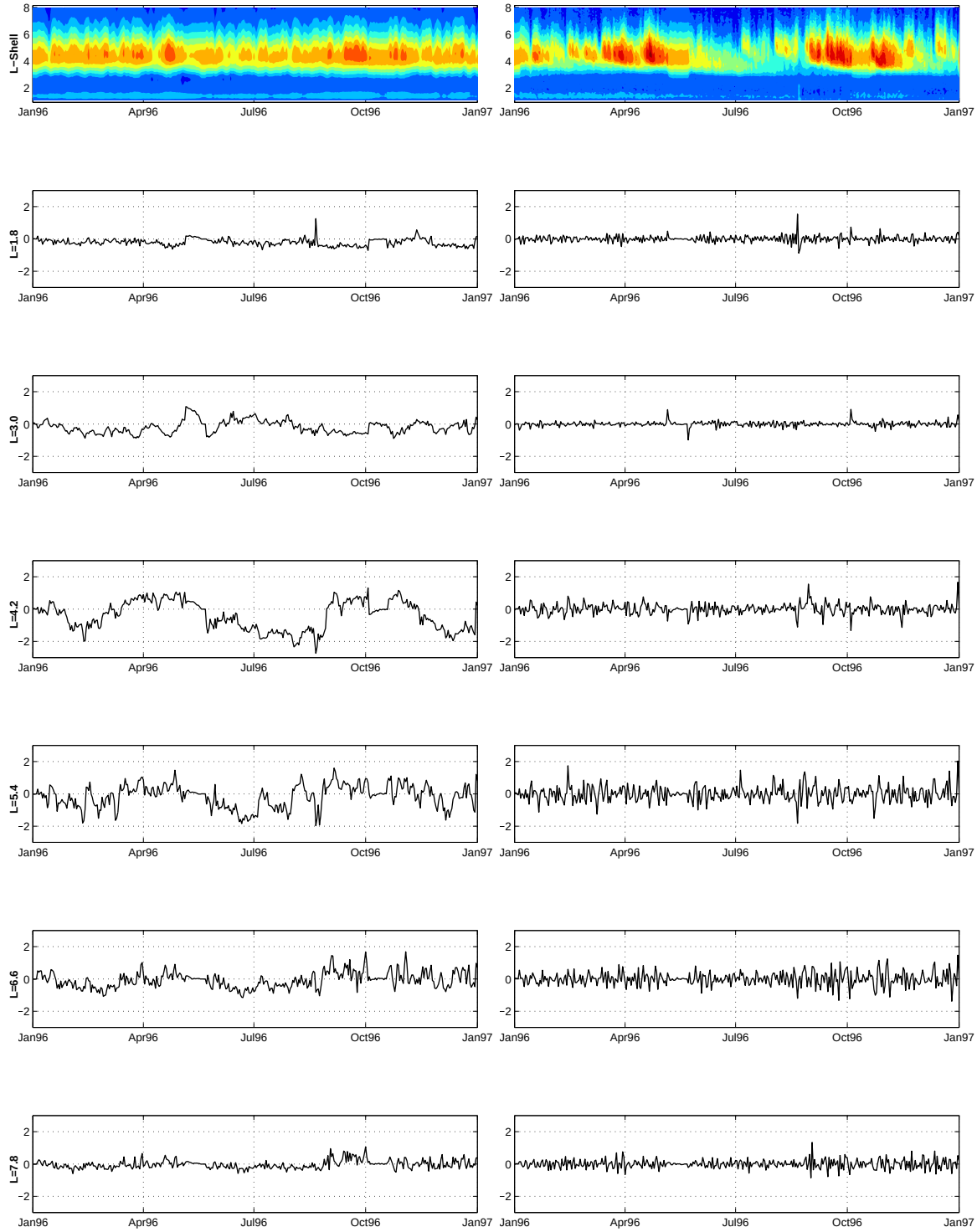


Figure B.43: Stationary (left) and Adaptive (right) $[V_{sw}, |\mathbf{B}|, \rho_{ion}] \rightarrow \log(\text{flux})$ predictions are plotted vs. L-shell, along with the resulting residuals for representative L-shells in 1996.

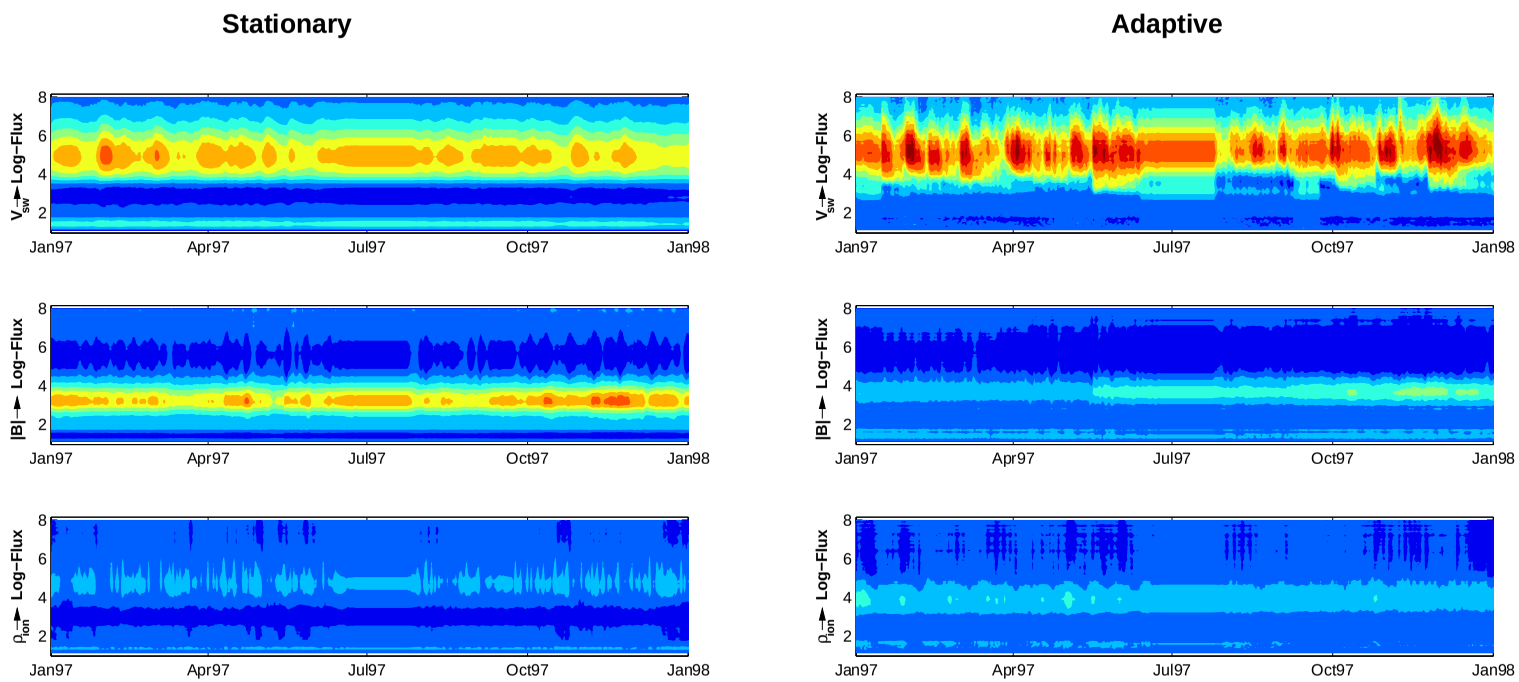


Figure B.44: Stationary and Adaptive $[V_{sw}, |\mathbf{B}|, \rho_{ion}] \rightarrow \log(\text{flux})$ component predictions are plotted vs. L-shell in 1997.

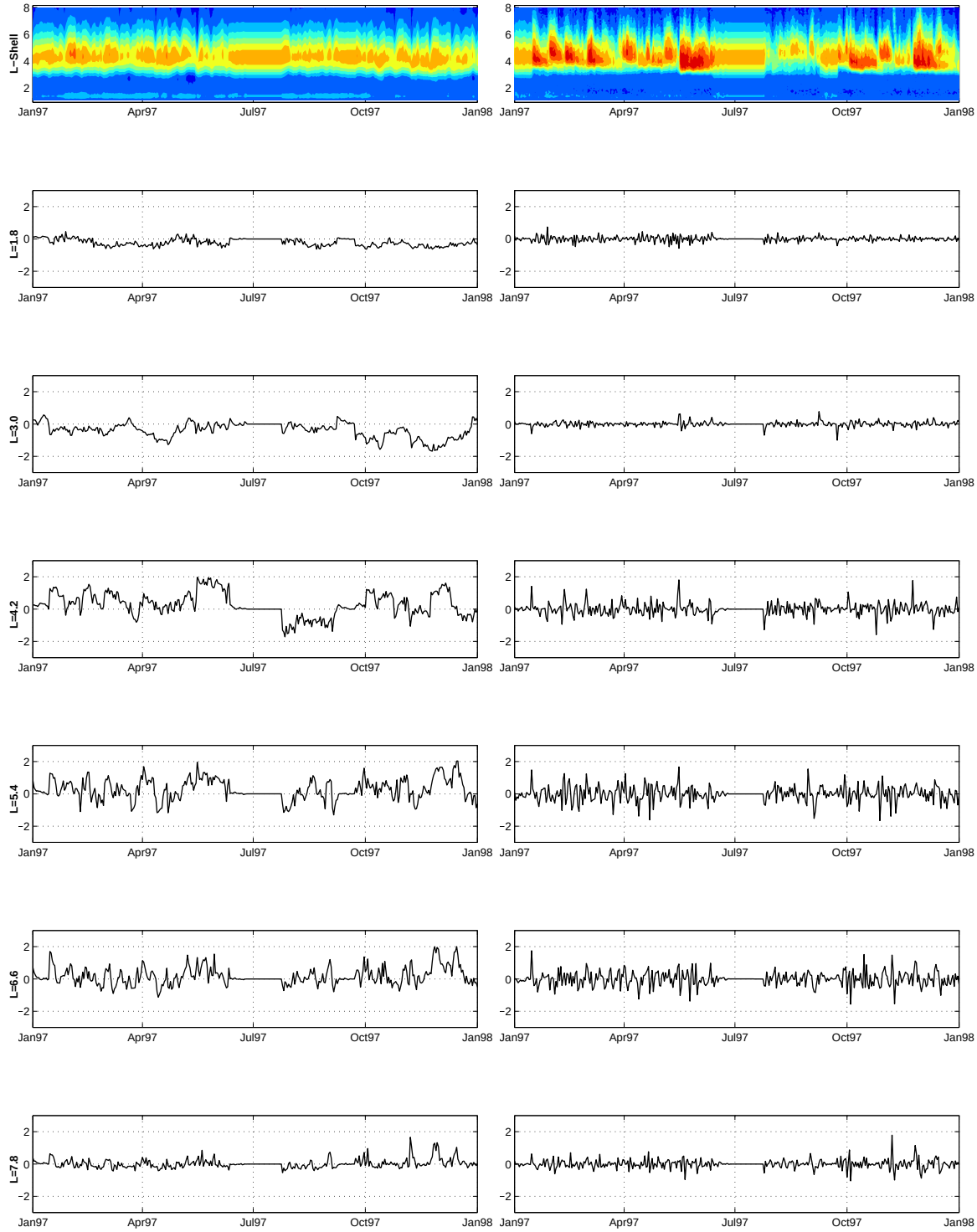


Figure B.45: Stationary (left) and Adaptive (right) $[V_{sw}, |\mathbf{B}|, \rho_{ion}] \rightarrow \log(\text{flux})$ predictions are plotted vs. L-shell, along with the resulting residuals for representative L-shells in 1997.

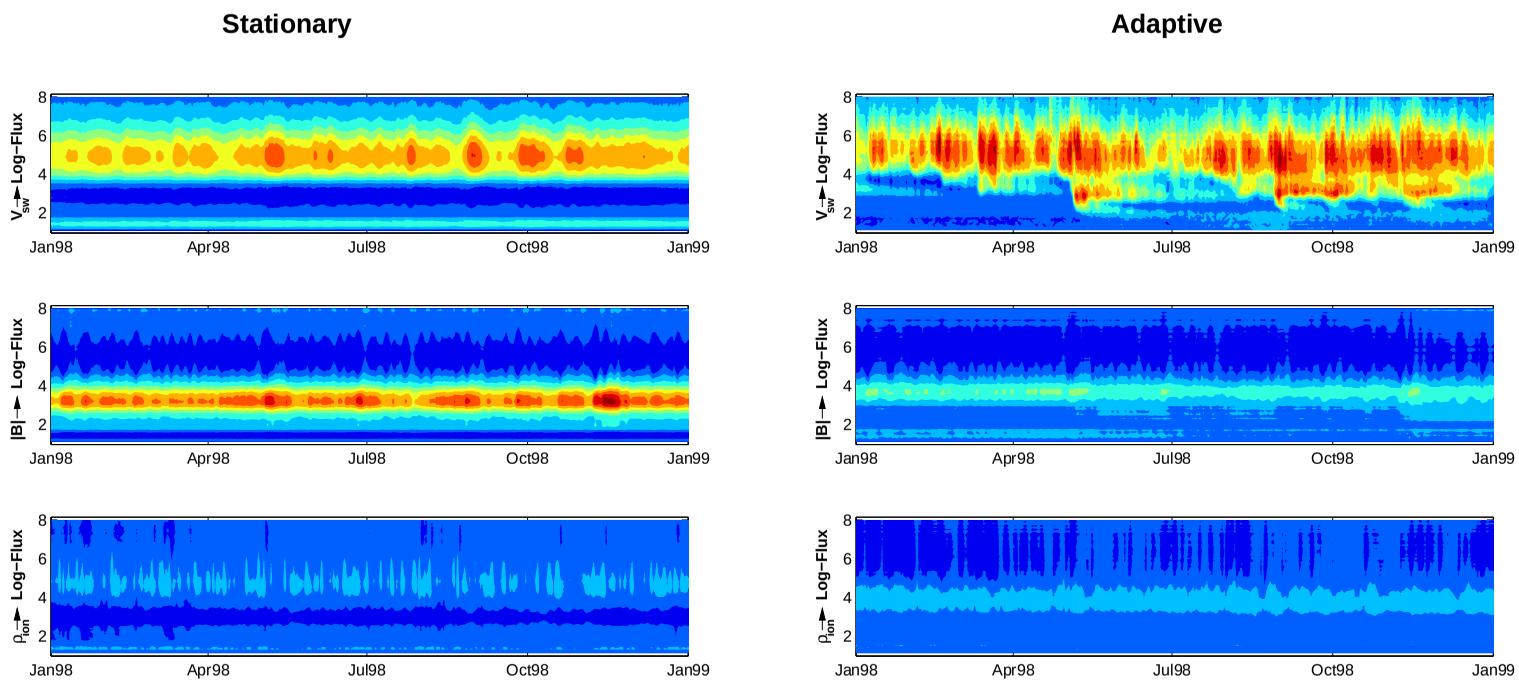


Figure B.46: Stationary and Adaptive $[V_{sw}, |\mathbf{B}|, \rho_{ion}] \rightarrow \log(\text{flux})$ component predictions are plotted vs. L-shell in 1998.

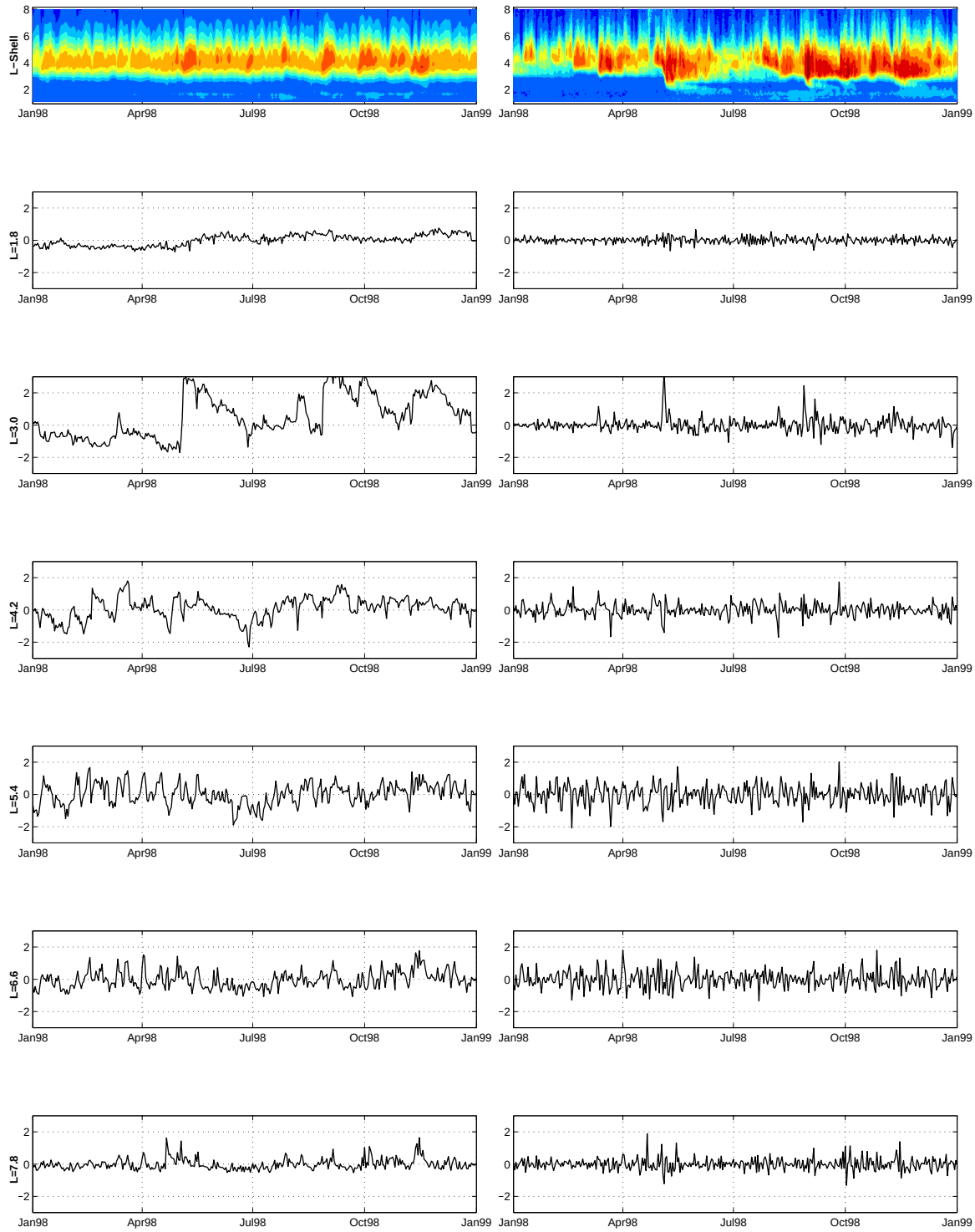


Figure B.47: Stationary (left) and Adaptive (right) $[V_{sw}, |\mathbf{B}|, \rho_{ion}] \rightarrow \log(\text{flux})$ predictions are plotted vs. L-shell, along with the resulting residuals for representative L-shells in 1998.

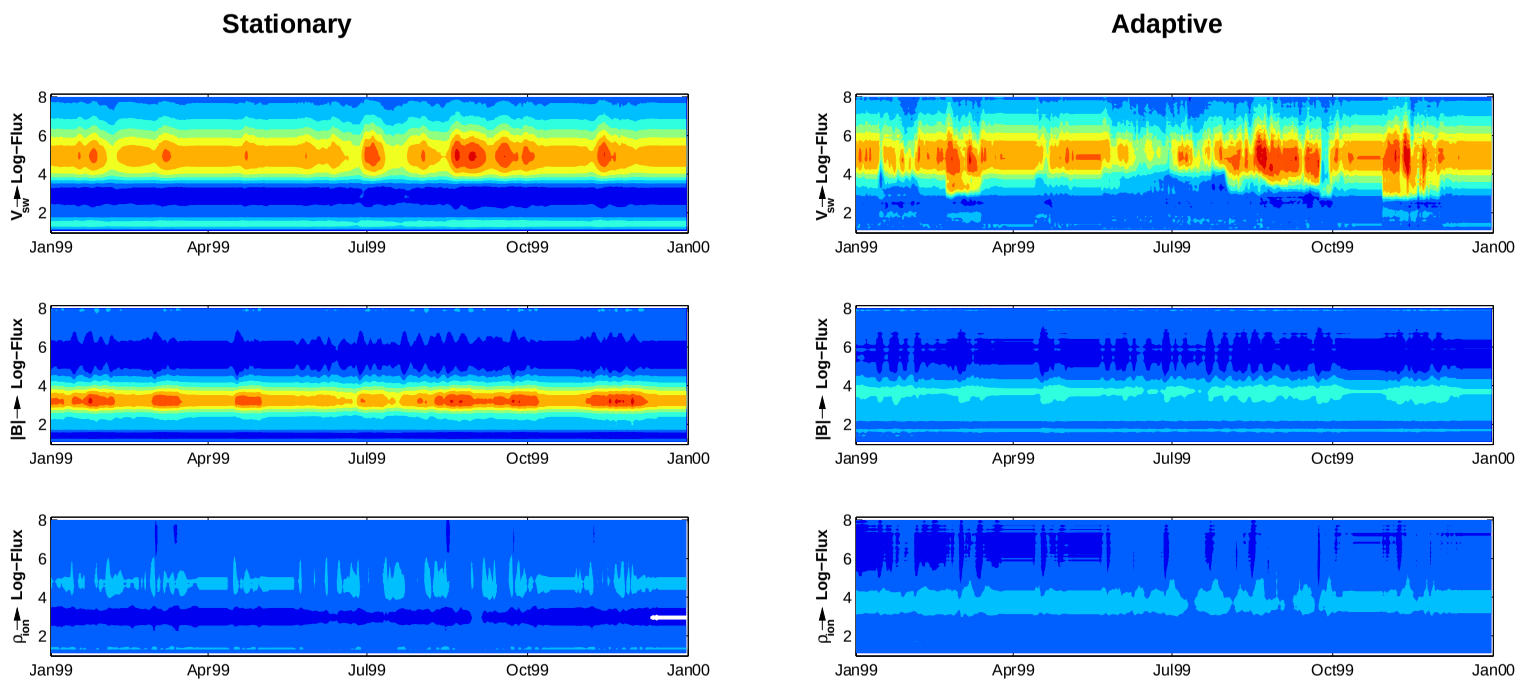


Figure B.48: Stationary and Adaptive $[V_{sw}, |\mathbf{B}|, \rho_{ion}] \rightarrow \log(\text{flux})$ component predictions are plotted vs. L-shell in 1999.

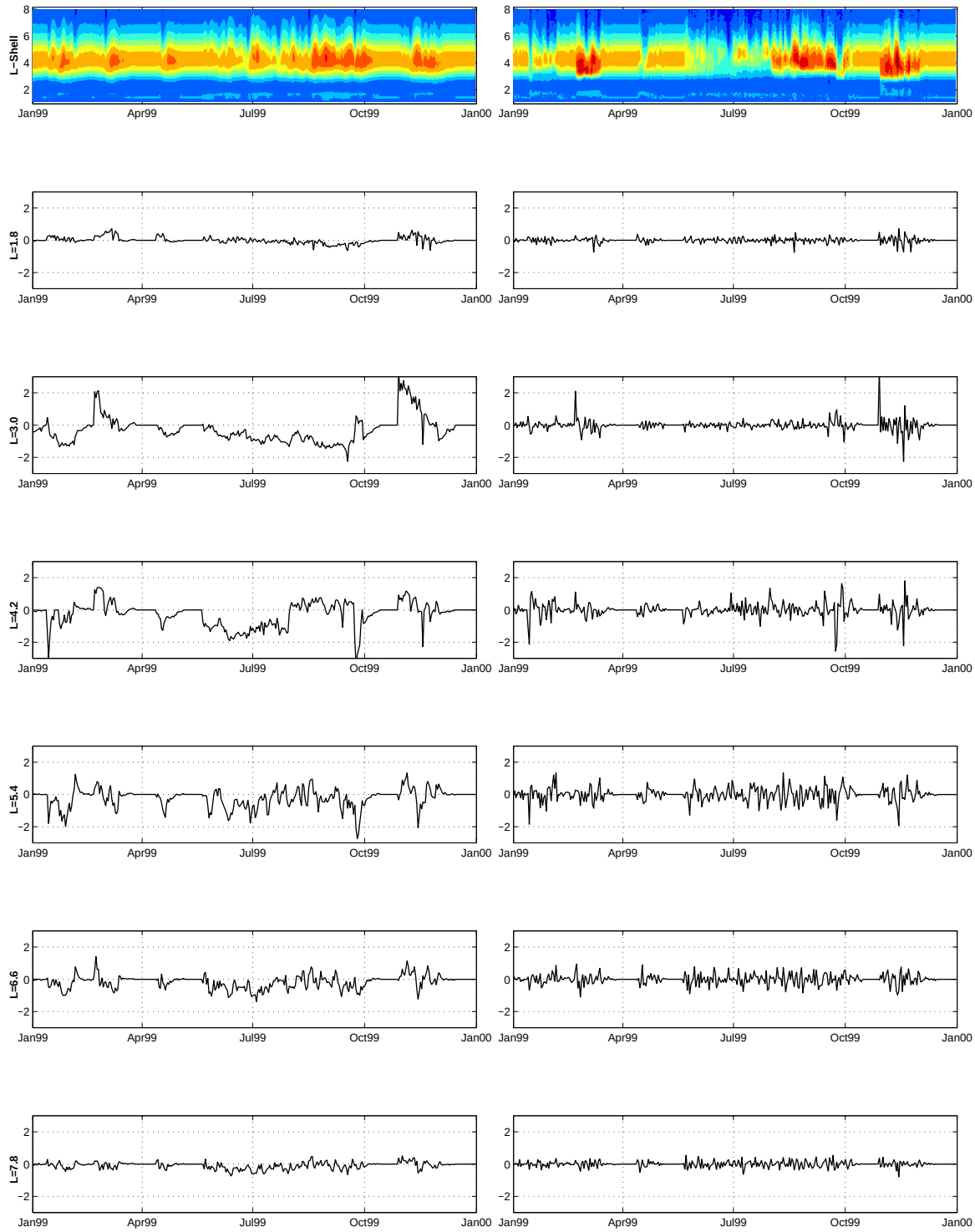


Figure B.49: Stationary (left) and Adaptive (right) $[V_{sw}, |\mathbf{B}|, \rho_{ion}] \rightarrow \log(\text{flux})$ predictions are plotted vs. L-shell, along with the resulting residuals for representative L-shells in 1999.

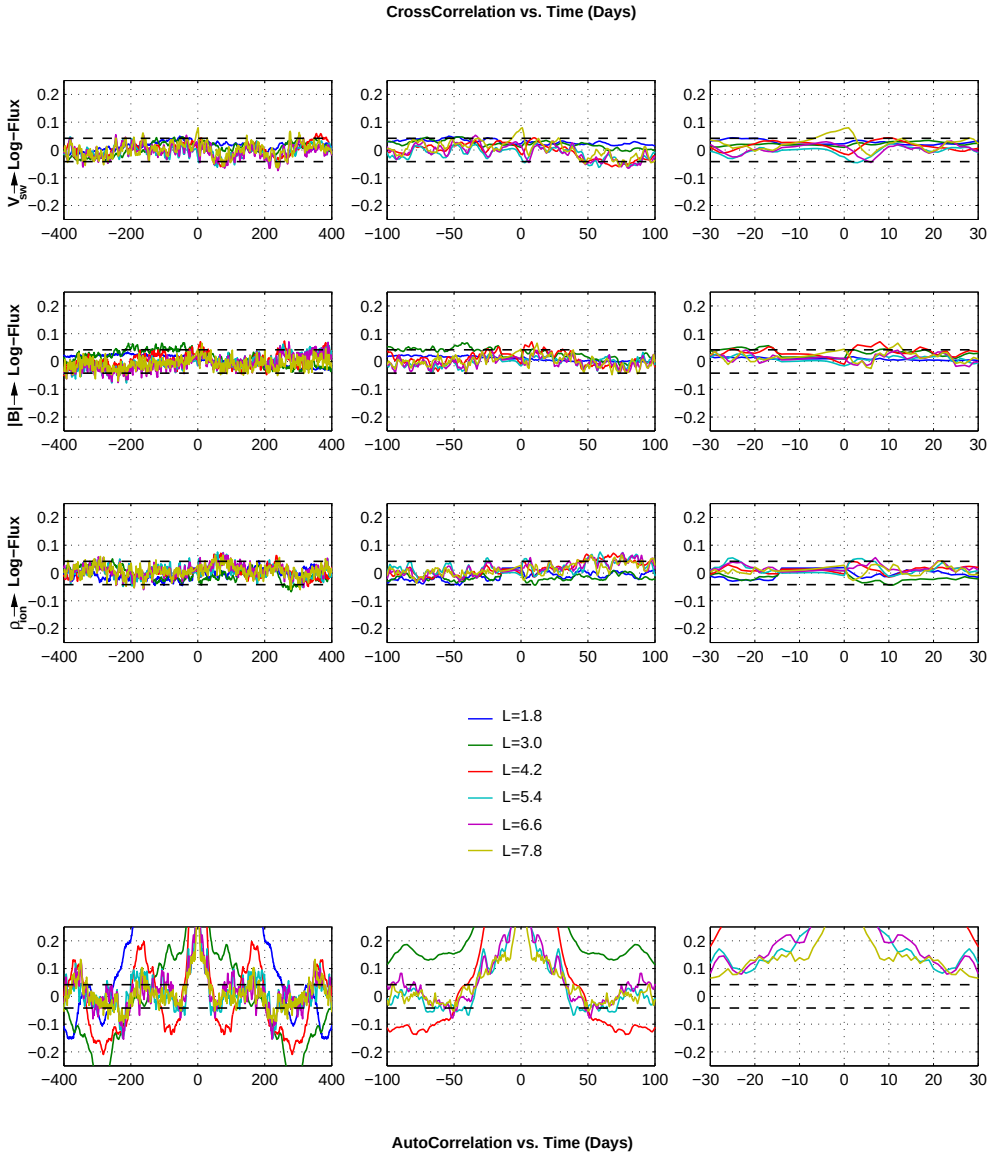


Figure B.50: Cross- and Auto-correlation functions of the $[V_{sw}, |\mathbf{B}|, \rho_{ion}] \rightarrow \log(\text{flux})$ residuals for **stationary** predictions. These are essentially the same as the plots presented in Chapter 4, and simply provide additional "zoomed" views of the time-dependent correlation coefficients.

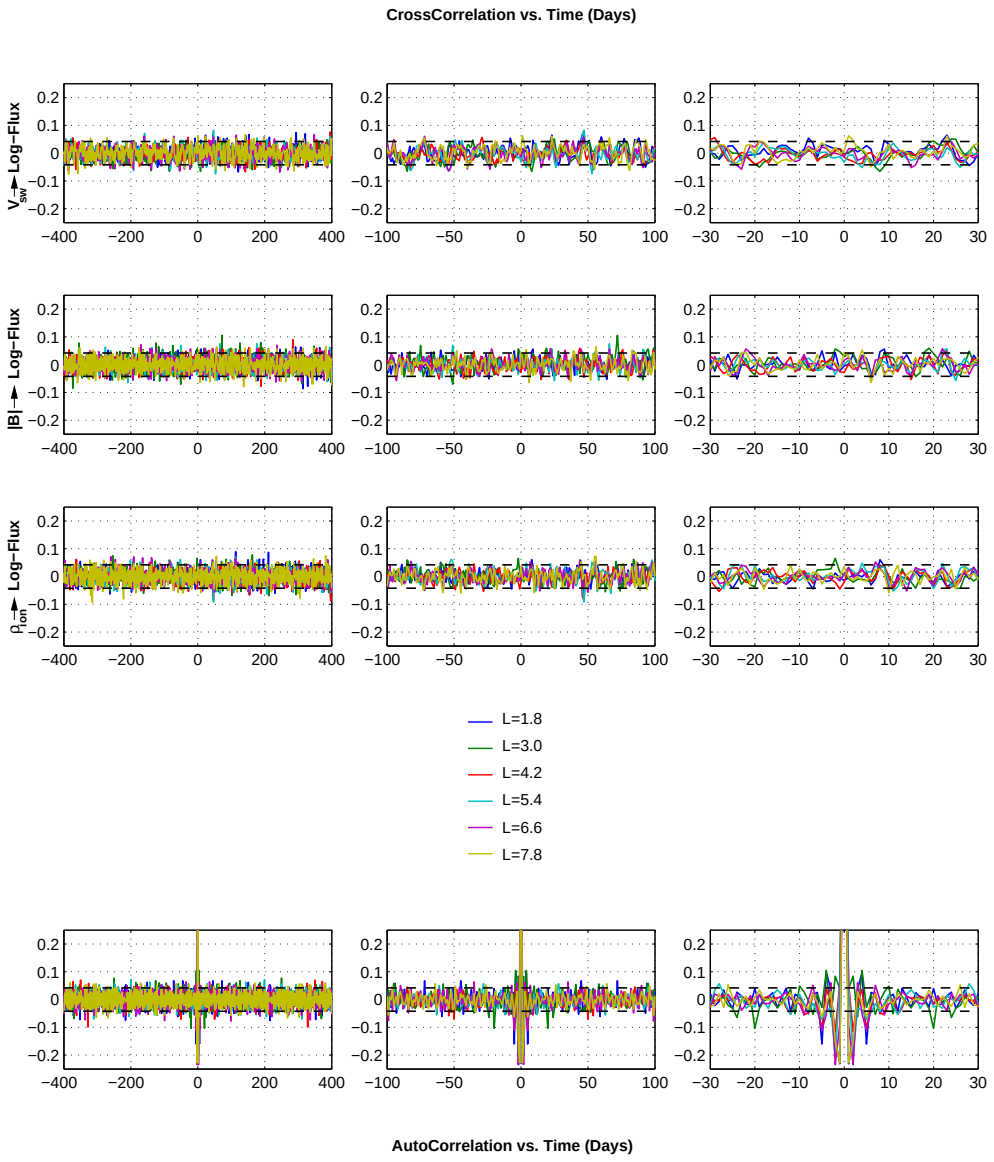


Figure B.51: Cross- and Auto-correlation functions of the $[V_{sw}, |\mathbf{B}|, \rho_{ion}] \rightarrow \log(\text{flux})$ residuals for **adaptive** predictions. These are essentially the same as the plots presented in Chapter 4, and simply provide additional "zoomed" views of the time-dependent correlation coefficients.

Appendix C

Polynomial-based Linear Prediction Filters

This appendix is included mostly for the author's benefit since he found it necessary to re-derive many of these relationships several times before realizing the value in recording things in some sort of "white-paper". It is assumed that the utility of the Kalman Filter as an adaptive system identification tool has been established. With this in mind, a step or two back is now taken in order to review general linear time series models in a bit more detail to determine if the prediction of radiation belt electron fluxes might be improved with more sophisticated model structures. The model types discussed in sequel are restricted to single-input, single-output (SISO) systems for the sake of clarity, however most of these relationships have multi-input, multi-output (MIMO) counterparts.

C.1 Combined Deterministic/Stochastic Prediction Filters

The output of a linear system can be described as a weighted moving average of time-lagged inputs plus some error. For an nth-order system, this can be represented with the following discrete convolution:

$$y(t) = \sum_{\tau=0}^{n} b_\tau u(t-\tau) + e(t) \tag{C.1}$$

or, as a difference equation:

$$y(t) = b_0 u(t) + b_1 u(t-1) + \cdots + b_n u(t-n) + e(t) \tag{C.2}$$

Here, $y(t)$ is the process output, b_τ is the τth time-lagged coefficient of the system's impulse response function, $u(t)$ is the input, and $e(t)$ is the model's prediction residual at time t. If we define the prediction as $\hat{y}(t)$, then $e(t) = y(t) - \hat{y}(t)$. Substituting this expression into Equation C.2 and rearranging provides the familiar FIR prediction equation:

$$\hat{y}(t) = b_0 u(t) + b_1 u(t-1) + \cdots + b_n u(t-n) = \sum_{\tau=0}^{n} b_\tau u(t-\tau) \tag{C.3}$$

A perfect model in a noise-free environement will result in $e(t) = 0$. Even when the residuals are not zero, if they comprise a Gaussian distribution with a zero-mean (i.e., they can be described as "white noise"), one can assume that all useful information in the data used to drive the model has been used up, and no better predictions can be made using the data provided. Residuals from such a simple model as Equation C.3 are usually not white, however, but possess significant correlated structure. This implies a deficiency in either the assumptions made when choosing the data to drive the model, or in the model itself. The former was partially addressed in this thesis by using multiple simultaneous inputs, and will not be discussed further here. Deficiencies in the model itself were addressed by assuming that the linear coefficients b_τ could vary with time to track non-linear dynamical processes.

Often, however, correlated structure in prediction residuals is due to simple recurrence, not non-linear dynamics, and can be accounted for by assuming a more sophisticated model structure. If it is assumed that previous process outputs are available, they can also be filtered in order to reduce and possibly normalize the residuals. Smaller residuals reduces the numerical "strain" exerted by the adaptive identification algorithm, thereby allowing it to better modify the model parameters to track true non-linear changes in the system dynamics.

$$\sum_{\tau=0}^{n} a_\tau y(t-\tau) = \sum_{\tau=0}^{n} b_\tau u(t-\tau) + e(t) \tag{C.4}$$

$$a_0 y(t) + \cdots + a_n y(t-n) = b_0 u(t) + \cdots + b_n u(t-n) + e(t) \tag{C.5}$$

Again, since $e(t) = y(t) - \hat{y}(t)$, we can substitute and rearrange the previous expressions to provide an autoregressive prediction model with exogenous inputs, or an ARX model. Assuming that $a_0 \equiv 1$ gives the following predictor:

$$\begin{aligned} \hat{y}(t) &= b_0 u(t) + b_1 u(t-1) + \cdots + b_n u(t-n) \\ &\quad -a_1 y(t-1) - \cdots - a_n y(t-n) \end{aligned} \tag{C.6}$$

...or...

$$\hat{y}(t) = \sum_{\tau=0}^{n} b_\tau u(t-\tau) - \sum_{\tau=1}^{n} a_\tau y(t-\tau) \tag{C.7}$$

Interestingly, if one starts with $y(t) = \hat{y}(t) + e(t)$ instead, rearranging Equation C.5, while making the appropriate substitutions, results in the following predictor:

$$\begin{aligned} \hat{y}(t) &= b_0 u(t) + b_1 u(t-1) + \cdots + b_n u(t-n) \\ &\quad -a_1 \hat{y}(t-1) - \cdots - a_n \hat{y}(t-n) \\ &\quad -a_1 e(t-1) - \cdots - a_n e(t-n) \end{aligned} \tag{C.8}$$

First, note that $y(t)$ is no longer eplicitly included in this equation, but rather previous predictions $\hat{y}(t-\tau)$ are. This kind of dynamical feedback will be discussed in more detail shortly. Next, note that this predictor actually filters previous residuals which, if all assumptions that went into this choice of model structure are appropriate, should be white noise. Therefore, the ARX prediction filter has introduced a stochastic component that can actually help to improve predictions.

C.2 Dynamical Feedback and Transfer Functions

If, for the moment, the stochastic component of the ARX prediction filter is ignored, a trivial modification to Equation C.8 yields:

$$\begin{aligned} \hat{y}(t) &= b_0 u(t) + b_1 u(t-1) + \cdots + b_n u(t-n) \\ &\quad - f_1 \hat{y}(t-1) - \cdots - f_n \hat{y}(t-n) \end{aligned} \tag{C.9}$$

The parameter f has been introduced simply to distinguish it from a. This predictor looks surprisingly like the one for the ARX model, however as noted previously, there are no previous observations that are explicitly included, but rather only previous predictions. This constitutes a true feedback loop to the model which can potentially enhance its dynamic flexibility significantly. Rearranging once again, and defining q as the backward shift operator such that

$$q^n u(t) = u(t-n) \tag{C.10}$$

it becomes possible to rewrite Equation C.9 as

$$\left(\sum_{\tau=0}^{n} f_\tau q^\tau\right)\hat{y}(t) = \left(\sum_{\tau=0}^{n} b_\tau q^\tau\right)u(t) \tag{C.11}$$

Finally, setting the operation defined inside the parentheses as $F(q)$ and $B(q)$, respectively, we find that we are able to represent this recursive relationship as a transfer function using standard numerator and a denominator notation:

$$\hat{y}(t) = \frac{B(q)}{F(q)}u(t) \tag{C.12}$$

If the input is unknown, the output might still be described as colored, or filtered white noise.

$$y(t) = \frac{C(q)}{D(q)}e(t) \tag{C.13}$$

If $e(t) = y(t) - \hat{y}(t)$, or in other words it defines the prediction residuals of the model, we can once again substitute in the following manner:

$$D(q)y(t) = C(q)y(t) - C(q)\hat{y}(t) \tag{C.14}$$

Rearranging and isolating the prediction $\hat{y}(t)$ on the left gives

$$\hat{y}(t) = \frac{C(q) - D(q)}{C(q)}y(t) \tag{C.15}$$

which, assuming $c_0 = d_0 = 1$, is simply short-hand for the following difference equation.

$$\hat{y}(t) = (c_1 - d_1)y(t-1) + \cdots + (c_n - d_n)y(t-n) - c_1\hat{y}(t-1) - \cdots - c_n\hat{y}(t-n) \tag{C.16}$$

A model that combines both deterministic and non-deterministic linear dynamic models, if properly identified, will act to separate directly driven dynamical responses from those that are due to internal dynamical processes and/or colored noise in the system. Even still, the possibility remains that some portion of the output time series variability is shared by both deterministic and stochastic components, resulting in the most general linear SISO model structure.

$$A(q)y(t) = \frac{B(q)}{F(q)}u(t) + \frac{C(q)}{D(q)}e(t) \tag{C.17}$$

Assuming $e(t)$ is once again defined as the prediction residuals, this gives the following predictor:

$$\hat{y}(t|t-1) = \frac{B(q)D(q)}{F(q)C(q)}u(t) + \left(\frac{C(q) - A(q)D(q)}{C(q)}\right)y(t) \tag{C.18}$$

C.3 Common Black-Box Models

The model structure in Equations C.17 and C.18 is too general for most practical applications. It is common practice for one or more of the transfer functions to simply be set equal to 1, or that either the inputs or outputs are not available or used at runtime. A few of the more common black-box models that result are described here.

C.3.1 Pure Stochastic Models

Moving Average (MA)

If no input drivers are known or available, and $A(q) = D(q) = 1$, a moving average (MA) filter results:

$$y(t) = C(q)e(t) \tag{C.19}$$

The “stochastic” nature of this model should be clear, as it involves the filtering of a sequence of random values $e(t)$. If $e(t)$ is assumed to be previous prediction residuals,

the 1-step predictor will depend on observations up to and including $y(t-1)$. It is given by

$$\begin{aligned} \hat{y}(t|t-1) &= \tfrac{C(q)-1}{C(q)} y(t) \\ &= c_1 y(t-1) + \cdots + c_n y(t-n) - c_1 \hat{y}(t-1) - \cdots - c_n \hat{y}(t-n) \end{aligned} \tag{C.20}$$

The MA filter is not capable of generating any type of correlated behavior. It simply smooths the time series upon which it operates. In addition, this model is not "linear in its parameters" (due to the $\hat{y}$'s on the right-hand side of the prediction equation), and a non-linear estimation algorithm is required to determine the optimal coefficients in $C(q)$. It is therefore not surprising that, alone, the MA filter is used little in engineering or science applications.

Autoregressive (AR)

If no input drivers are known or available, and $C(q) = D(q) = 1$, an autoregressive (AR) filter results:

$$A(q)y(t) = e(t) \tag{C.21}$$

If $e(t)$ is assumed to be previous prediction residuals, the 1-step predictor will depend on observations up to and including $y(t-1)$. It is given by

$$\begin{aligned} \hat{y}(t|t-1) &= (1 - A(q))\, y(t) \\ &= -a_1 y(t-1) - a_2 y(t-2) - \cdots - a_n y(t-n) \end{aligned} \tag{C.22}$$

In General, an AR filter is capable of modeling exponential decay, oscillations, and other types of correlated behavior. It is linear in its parameters, so they can be estimated using standard least-squares techniques. It is actually the most common model used to reproduce time series data when no external input is known or available. Clearly, starting with the general linear model structure, if one sets $A(q) = 1$, and allows $1/D(q)$ instead, an identical model results.

Autoregressive Moving Average (ARMA)

If no input drivers are known or available, and $A(q) = 1$, an autoregressive moving average (ARMA) filter results:

$$y(t) = \frac{C(q)}{D(q)} e(t) \tag{C.23}$$

If $e(t)$ is assumed to be previous prediction residuals, the 1-step predictor will depend on observations up to and including $y(t-1)$. It is given by

$$\begin{aligned} \hat{y}(t|t-1) &= \left(\tfrac{C(q)-D(q)}{C(q)}\right) y(t) \\ &= (c-d)_1 y(t-1) + \cdots + (c-d)_n y(t-n) \\ &\quad -c_1 \hat{y}(t-1) - \cdots - c_n \hat{y}(t-n) \end{aligned} \tag{C.24}$$

Collecting like terms makes it clear that the MA portion of the filter, $C(q)$, smooths the model's previous residuals, while the AR portion of the filter acts to reproduce correlated behavior in the observations. This type of model is not linear in its parameters and therefore requires a non-linear estimation algorithm in order to determine the optimal coefficients. Despite this drawback it has tremendous dynamic flexibility and its use is widespread as a powerful spectral analysis tool.

C.3.2 Purely Deterministic Models

Finite Impulse Response (FIR)

If only system inputs are available, and $A(q) = C(q) = D(q) = F(q) = 1$, a finite impulse response (FIR) filter results:

$$y(t) = B(q)u(t) + e(t) \tag{C.25}$$

The optimal predictor, which does not depend on past output observations, is

$$\begin{aligned} \hat{y}(t) &= B(q)u(t) \\ &= b_0 u(t) + b_1 u(t-1) + \cdots + b_n u(t-n) \end{aligned} \tag{C.26}$$

This is our favorite linear prediction filter, used for all analyses and validation studies presented in this thesis. It shapes the input signal, and can even reproduce correlated behavior, so long as it happens within n time-steps after a a change in the input, because its coefficients fully represent a linear system's discrete impulse response out to the nth term. It is linear in its parameters, allowing a simple least-squares parameter estimation. Its primary drawback is that the model order must be chosen very large if one wishes reproduce an extended response at a high time resolution. Today's available computer speed and memory makes this less of an issue than it once was.

Output Error (OE)

If only system inputs are available, and $A(q) = C(q) = D(q) = 1$, a so-called output error (OE) filter results:

$$y(t) = \frac{B(q)}{F(q)}u(t) + e(t) \tag{C.27}$$

The optimal predictor, which does not depend on past output observations, is

$$\begin{aligned} \hat{y}(t) &= \tfrac{B(q)}{F(q)}u(t) \\ &= b_0u(t) + b_1u(t-1) + \cdots + b_nu(t-n) \\ &\qquad - f_1\hat{y}(t-1) - \cdots - f_n\hat{y}(t-n) \end{aligned} \tag{C.28}$$

This filter can theoretically produce an infinite impulse response (IIR), usually with only a small number of parameters relative to the FIR filter. It is not linear in its parameters, however, and requires a non-linear estimation algorithm in order to determine the optimal coefficients. This type of filter is popular in engineering and scientific applications because it is capable of representing a discrete driven damped harmonic oscillator, a very common type of dynamical system in nature.

C.3.3 Combined Deterministic and Stochastic Models

Autoregressive with eXogenous Inputs (ARX)

If both system inputs and recent output observations are available, and $F(q) = C(q) = D(q) = 1$, an autoregressive filter with exogenous inputs (ARX) results:

$$A(q)y(t) = B(q)u(t) + e(t) \text{ ...or... } y(t) = \frac{B(q)}{A(q)}u(t) + \frac{1}{A(q)}e(t) \tag{C.29}$$

If $e(t)$ is assumed to be previous prediction residuals, the 1-step predictor will depend on observations up to and including $y(t-1)$. It is given by

$$\begin{aligned} \hat{y}(t|t-1) &= B(q)u(t) + (1 - A(q))\,y(t) \\ &= b_0u(t) + b_1u(t-1) + \cdots + b_nu(t-n) \\ &\qquad -a_1y(t-1) - \cdots - a_ny(t-n) \end{aligned} \tag{C.30}$$

This is the most common linear filter that combines both a deterministic and stochastic component in an attempt to separate directly driven from internal dynamics and colored noise. It is popular mostly because it is actually linear in its parameters and its coefficients can easily be determined using simple least-squares estimation techniques. In fact, this model structure is often assumed for identification purposes, then the $A(q)$ coefficients are used in place of the $F(q)$ coefficients of the OE model. The astute reader will however recognize that the noise assumptions are wrong (the OE model assumes additive white noise, while the ARX assumes that the noise is filtered through $1/A(q)$). Such an approach can lead to severely biased model output in some instances.

Autoregressive Moving Average with eXogenous Inputs (ARMAX)

If both system inputs and recent output observations are available, and $F(q) = A(q) = 1$, an autoregressive moving average filter with exogenous inputs (ARMAX) results:

$$y(t) = B(q)u(t) + \frac{C(q)}{D(q)}e(t) \tag{C.31}$$

If $e(t)$ is assumed to be previous prediction residuals, the 1-step predictor will depend

on observations up to and including $y(t-1)$. It is given by

$$\begin{aligned}\hat{y}(t|t-1) &= \tfrac{B(q)}{C(q)}u(t) + \left(\tfrac{C(q)-D(q)}{C(q)}\right)y(t) \\ &= b_0u(t) + b_1u(t-1) + \cdots + b_nu(t-n) \\ &\qquad + (c-d)_1y(t-1) + \cdots + (c-d)_ny(t-n) \\ &\qquad -c_1\hat{y}(t-1) - \cdots - c_n\hat{y}(t-n)\end{aligned} \tag{C.32}$$

This linear filter is not used as often as those mentioned previously, simply because it is comprised of a large number of coefficients that must be determined using a non-linear estimation algorithm non-linearity of its parameters. It is extremely powerful, however, due to the flexible ARMA portion of the model that can recreate a wide variety of correlated stochastic dynamics, plus the additional input filter $B(q)$ that helps account for what would otherwise appear as non-stationary sequences in the prediction residuals (i.e. extended sequences of residuals without a zero-mean).

Box-Jenkins (BJ)

If both system inputs and recent output observations are available, and there are no shared dynamics between the deterministic and stochastic portions of the model (i.e. $A(q) = 1$) a so-called Box-Jenkins (BJ) filter results:

$$y(t) = \frac{B(q)}{F(q)}u(t) + \frac{C(q)}{D(q)}e(t) \tag{C.33}$$

If $e(t)$ is assumed to be previous prediction residuals, the 1-step predictor will depend on observations up to and including $y(t-1)$. It is given by

$$\hat{y}(t|t-1) = \frac{B(q)D(q)}{F(q)C(q)}u(t) + \left(\frac{C(q)-D(q)}{C(q)}\right)y(t) \tag{C.34}$$

The difference equation is not shown because it is extremely complex. the terms $B(q)D(q)$ and $F(q)C(q)$ are actually the full convolution of two filters, resulting in product filters with numbers of coefficients equal to $n_{BD} = n_B + n_D - 1$ and $n_{FC} = n_F + n_C - 1$. This type of filter is highly non-linear in its parameters and is rarely

used in practice due to its complexity. It has at least two important characteristics that make it an attractive option for modeling dynamical systems though: 1) it offers an extremely flexible combined deterministic-stochastic structure capable of separating directly driven from internal correlated dynamics; 2) it includes denominator dynamics in the deterministic term, resulting in fewer necessary inputs to reproduce a full impulse response. This means fewer inputs need to be dropped when accounting for small data gaps in the input data (see Chapter 2 for a full explanation). More often, a structurally simpler but equivalent state-space model is used when such sophisticated dynamics are being modeled.