5 Other Topics for Chapter 5

5.7 Dynamical and physical balance in the initial conditions

We saw in Chapter 1 that Richardson's (1922) experiment resulted in a disastrous estimation of the initial surface pressure tendency (a forecast of a change of 146 hPa in 6 h, whereas the actual pressure remained almost unchanged) because of noisy data and the presence of fast inertia-gravity waves in the solution of the primitive equations. If there are fast and slow waves in the solution of a model, $\omega_{fast} >> \omega_{slow}$, and $u = U_{slow}e^{-i\omega_{slow}t} + U_{fast}e^{-i\omega_{fast}t}$, then $\partial u/\partial t = -i\omega_{slow}U_{slow}e^{-i\omega_{slow}t} - i\omega_{fast}U_{fast}e^{-i\omega_{fast}t}$. As shown schematically in Figure 1.2.1 it is clear that, unless the amplitude of the fast waves component is made very small, the fast waves will dominate the initial tendency.

We saw in Section 2.5 that in the SWEs, the simplest example of primitive equations, there are two types of wave solutions: (a) steady or slowly evolving, quasi-geostrophically balanced "weather" modes, satisfying $\omega \approx 0$, and (b) fast inertia-gravity waves with a frequency dispersion relationship $\omega^2 \approx f^2 + gD(k^2 + l^2)$, where k, l are the horizontal wavenumbers, and D is the mean depth of the model. For the external mode, with about 10 km equivalent depth, inertia-gravity waves travel at a speed of about 300 m/s, and unless they are filtered out of the initial conditions, they can indeed produce a very noisy forecast. After a while, though, the inertia-gravity waves subside as the solution evolves towards quasi-geostrophic balance (a process known as geostrophic adjustment), so that, unless they interact nonlinearly with the slower "weather waves", the inertia-gravity waves do not necessarily ruin the forecast.

In this section we first consider the geostrophic adjustment process that takes place within a linear SWE system. This also allows us to assess what types of observations are most useful for NWP. We then consider the nonlinear case and describe the nonlinear normal mode initialization method, which was used for many years to reduce the imbalance in the initial conditions. Finally we describe a more recently introduced type of dynamical initialization, denoted digital filtering, which is simple and very effective. We should note that if the analysis is out of balance, all of the initial information that projects on inertia-gravity waves will be lost, whether the inertia-gravity waves are filtered out during the model integration through geostrophic adjustment, or through the other initialization methods. For this reason it is preferable to enforce balance within the analysis as done in 3D-Var, since this reduces the loss of information.

5.7.1 Geostrophic adjustment and the relative importance of different observations

If the initial conditions of a model are not in quasi-geostrophic balance, the balanced portion of the initial field will project on the quasi-geostrophic mode, and the unbalanced portion will project onto inertia-gravity waves. These waves have large horizontal divergence and propagate horizontally, dispersing quite fast. Because of horizontal dispersion, after a while the amplitude of the inertia-gravity waves becomes much smaller, and the leftover fields remain in quasi-geostrophic balance. Rossby (1936, 1938) first described this process of "*g*eostrophic adjustment." The time scale for geostrophic adjustment is of the order of f^{-1} (about 12 h). Arakawa (1997) provided an analytic solution of the linear geostrophic adjustment problem for the SWEs.

In Section 2.4 we showed that the potential vorticity is conserved for individual parcels:

$$\frac{d}{dt}\frac{\boldsymbol{\nabla}^2\psi + f}{\Phi} = 0 \tag{5.7.1}$$

Consider small perturbations on a basic state of rest. The linearized SWE potential vorticity is obtained assuming that the relative vorticity $\boldsymbol{\nabla}^2\psi$ is small compared with the Coriolis parameter and that the total geopotential height $\Phi = gD + \phi$, where the perturbations to the free surface are small compared to the mean depth of the fluid. In that case, the conservation of potential vorticity becomes

$$\frac{d}{dt}\eta = \frac{d}{dt}\left(f + \boldsymbol{\nabla}^2\psi - \frac{f_0}{gD}\phi\right) = 0 \tag{5.7.2}$$

Parcels will evolve conserving their *i*nitial potential vorticity $\eta(0) = f + \boldsymbol{\nabla}^2\psi - (f_o/gD)\phi$ even as they undergo the geostrophic adjustment process. This important conservation property allows us to assess how much of the initial mass and wind increments will project on the slow modes, and be "remembered" by the model after geostrophic adjustment, and how much information will be lost through inertia-gravity waves. If we introduce a perturbation $\delta\psi(0), \delta\phi(0)$ in the initial conditions through data assimilation, the perturbation in potential vorticity will remain in the solution even after geostrophic adjustment: $\delta\eta_g = \delta\eta(0)$. Recall that after geostrophic adjustment, the winds become geostrophic, so that $\psi_g = \phi_g/f_0$.

Assume that within a univariate analysis, the introduction of observations results in an *a*nalysis increment field associated with either mass observations $\delta\phi(x, y)$ or wind observations $\delta\psi(x, y)$. After about 12–24 h the initial unbalanced field will disperse away as inertia-gravity waves, and the remaining increments will be in geostrophic balance:

$$\delta\phi \rightarrow \delta\phi_g, \delta\psi \rightarrow \delta\psi_g \text{with } \delta\psi_g = \delta\phi_g/f_o \tag{5.7.3}$$

Assume that the analysis increment field was of the form $\delta\phi = Ae^{i(kx+ly)}$, with $k^2 + l^2 = n^2 = (2\pi/L)^2$. From (5.7.2), the final increment of potential vorticity after geostrophic adjustment is equal to the initial analysis increment $\delta\eta_g(x, y) = \delta\eta(x, y)$. Consider the effect of introducing only mass observations (performing a univariate analysis). They

will result in an analysis increment $\delta\phi$, and the potential vorticity initial (and final) increments are then

$$\delta\eta = -\frac{f_0}{gD}\delta\phi = \boldsymbol{\nabla}^2\delta\psi_g - \frac{f_0}{gD}\delta\phi_g = -n^2\delta\psi_g - \frac{f_0}{gD}\delta\phi_g \tag{5.7.4}$$

The *impact of the mass observation* after geostrophic adjustment is therefore

$$\delta\phi_g = \frac{1}{n^2R^2+1}\delta\phi \tag{5.7.5}$$

where $R = \sqrt{gD/f_0^2}$ is the *R*ossby radius of deformation, the natural quasi-geostrophic horizontal scale given a rotation rate and mean depth. Equation (5.7.5) indicates that the response depends strongly on whether the waves are short or long compared with R. For *long waves*, for which $n^2R^2 \ll 1$, the model "remembers" the mass data: $\delta\phi_g \approx \delta\phi$, i.e., it retains it after geostrophic adjustment. For *short waves*, for which $n^2R^2 \gg 1$, on the other hand, the model "forgets" the mass information: $\delta\phi_g \approx 0$.

The situation is reversed for a univariate analysis of only wind data, leading to an analysis increment $\delta\psi$. After geostrophic adjustment,

$$\delta\psi_g = \frac{n^2R^2}{n^2R^2+1}\delta\psi \tag{5.7.6}$$

so that for *long waves*, $n^2R^2 \ll 1$, the model "forgets" the wind data: $\delta\psi_g \approx 0$, whereas wind information is retained in *short waves*: $\delta\psi_g \approx \delta\psi$.

Now we have to determine which waves are "short" and which "long". In mid-latitudes, with $f_0 \sim 10^{-4}$, for the external or barotropic mode $D \sim$10 km, so that short waves are those for which $n^2R^2 \gg 1$ or $L \ll \sqrt{4\pi^2gD/f_0^2} \approx 20000$ km, i.e., *all but planetary waves are very short*. This implies that the model essentially will ignore surface pressure data, which is the mass data corresponding to the external mode, and it will adjust its surface pressure to the barotropic component of the wind. For the first internal mode $D \sim 1$ km, and waves are "short" if shorter than about 6000 km. In other words, most of the energy in the mid-latitude atmosphere is actually in short waves. In the tropics, where f is an order of magnitude smaller, the statement that "most waves are short" applies even more strongly. For this reason, *winds tend to be more effective in providing initial conditions for an NWP model than mass data*, but *temperature data are more important for shallower vertical modes*, for which the wind will adjust to the temperature observations.

The "acceptance" of wind and/or mass data is enhanced by the use of multivariate analysis schemes, because the geostrophic correlation assumed in the background error covariance enforces an approximate geostrophic balance in the analysis increments. However, even with multivariate analysis, because nonlinearities do not allow for a perfect balance, wind observations still have the strongest impact on the skill of a forecast in modern data assimilation systems. The full impact of an observing system depends on the extent to which they contribute to define the potential vorticity, as discussed above, and on other factors:

(a) Precision Assume we observe winds and mass, and in a multivariate scheme we combine them optimally, in which case their precisions (inverse of the error variances) are additive. This allows a simple estimation of the relative contributions to the analysis precision of mass and wind measurements. If we observe heights $\phi = f_0\psi^{\text{mass}}$ with an error $\delta\phi_{ob} = g\delta z_{ob}$, and winds $\mathbf{v} = \mathbf{k}\times\boldsymbol{\nabla}\psi^{\text{wind}}$ with an error $|\delta\mathbf{v}_{ob}|$, they both contribute to the streamfunction. If we combine them optimally, the analysis error precision will be

$$\frac{1}{|\delta\psi|^2} = \frac{f_o^2}{g^2|\delta z_{ob}|^2} + \frac{n^2}{|\delta\mathbf{v}_{ob}|^2} \tag{5.7.7}$$

suggesting that the wind field contributes more accuracy to the analysis for short waves or in the tropics. The higher accuracy of winds for short waves is a result of the fact that they measure a gradient field, and is *independent of the geostrophic adjustment argument.*

(b) Data coverage, both in the horizontal and in the vertical Obviously, the denser an observing system is, the more it will contribute to the accuracy of the analysis. This is true in the horizontal as well as in the vertical, so that vertical profiles of winds, temperature, or moisture are found to be more useful than single level observations. An observing system will also contribute more to the skill of the forecasts in the absence of other observing systems. For example, in the Northern Hemisphere, which has a relatively good network of rawinsondes, the contribution of satellite data to the improvement of NWP forecasts is much lower than in the Southern Hemisphere, where rawinsondes are much fewer.

(c) Physical model adjustments and model spin-up In the same way that mass and wind initial fields undergo a dynamical adjustment towards geostrophic balance, other variables quickly evolve towards thermal and hydrological balance within the model. For example, because of the low heat capacity of the surface layer, surface air temperatures (at 2 m) adjust very rapidly towards equilibrium with the sea or land surface temperatures. As a result, it is difficult to effectively use surface air temperature observations since the model will tend to "forget" them (replace them by model adjusted values). Similarly, if the moisture analysis profiles are wetter or drier than what the model hydrological equilibrium would require (i.e., the model has a wet or dry bias), the analysis moisture profiles are quickly replaced by model adjusted profiles. Therefore, a model with a dry climatological bias will produce excessive rain during the analysis cycle, as the initial conditions bring in higher observed moisture profiles every 6 h, and the model rains out what it perceives as "excessive moisture". This adjustment process, which is also affected by other physical parameterizations such as surface fluxes and radiation, is known as the "spin-up" or "spin-down" of the model, depending on whether there is an initial increase or decrease in the precipitation. The spin-up process is strongest immediately after the analysis and takes between 12 and 36 h before reaching model balance. It can be reduced by "physical initialization" or assimilation of precipitation, in which temperature and moisture profiles are modified so that the model during the

analysis cycle is forced to produce precipitation similar to the observed precipitation (e.g., Krishnamurti *et al.*, 1988, Treadon, 1996, Falkovich *et al.*, 2000).

Figure 5.7.1 shows the results of a comparison of different 5-day forecasts performed every 12 h for February 1998 using the 2000 operational NCEP data assimilation system (courtesy of Michiko Masutani and Stephen Lord). The forecasts are based on different data assimilation experiments, in which: (a) all available data was assimilated (ALL); (b) the satellite radiances were not assimilated (No TOVS 1B); (c) the rawinsonde temperatures were not assimilated (No RAOB Temp); and (d) the rawinsondes winds were not assimilated (No RAOB Wind). The results illustrate several conclusions, in agreement with the discussion above:

- If every statistical assumption about errors was perfect, the ALL experiments should have the best forecasts. This is generally but not always true.
- In the Northern Hemisphere where the rawinsonde network is fairly abundant, the forecasts from the assimilation without satellite data No TOVS 1B, are, on the average, slightly worse than the ALL data experiments.
- Eliminating the rawinsonde winds from the data assimilation has a much larger negative impact than eliminating the rawinsonde temperatures in the Northern Hemisphere.
- In the Southern Hemisphere, where there are relatively few rawinsondes, the satellite radiances are the backbone of the information needed in the data assimilation. Without satellite radiances, the forecasts in the Southern Hemisphere substantially deteriorate.
- We should also note that there is substantial day-to-day variability in the 5-day forecast skill. This can be attributed to the changes in atmospheric predictability, i.e., on some days the atmosphere is simply easier to predict than on others. This will be studied in detail in Chapter 6.

5.7.2 Normal modes initialization

Because atmospheric motion is nonlinear, a simple linear geostrophic balance as discussed in the previous section is not enough to ensure balanced initial conditions. An approach that has been widely used to improve the initial imbalance is that of nonlinear normal modes initialization, introduced by Machenhauer (1977) and by Baer and Tribbia (1977). As indicated by its name, it requires the determination of the (linear) normal modes of a model as a first step. Daley (1991), Temperton and Williamson (1981), and others give a complete discussion about how this procedure is carried out in a three-dimensional model. Here we only illustrate how it would be applied to a simple SWE model on a periodic f-plane model. The SWEs are written separating them into

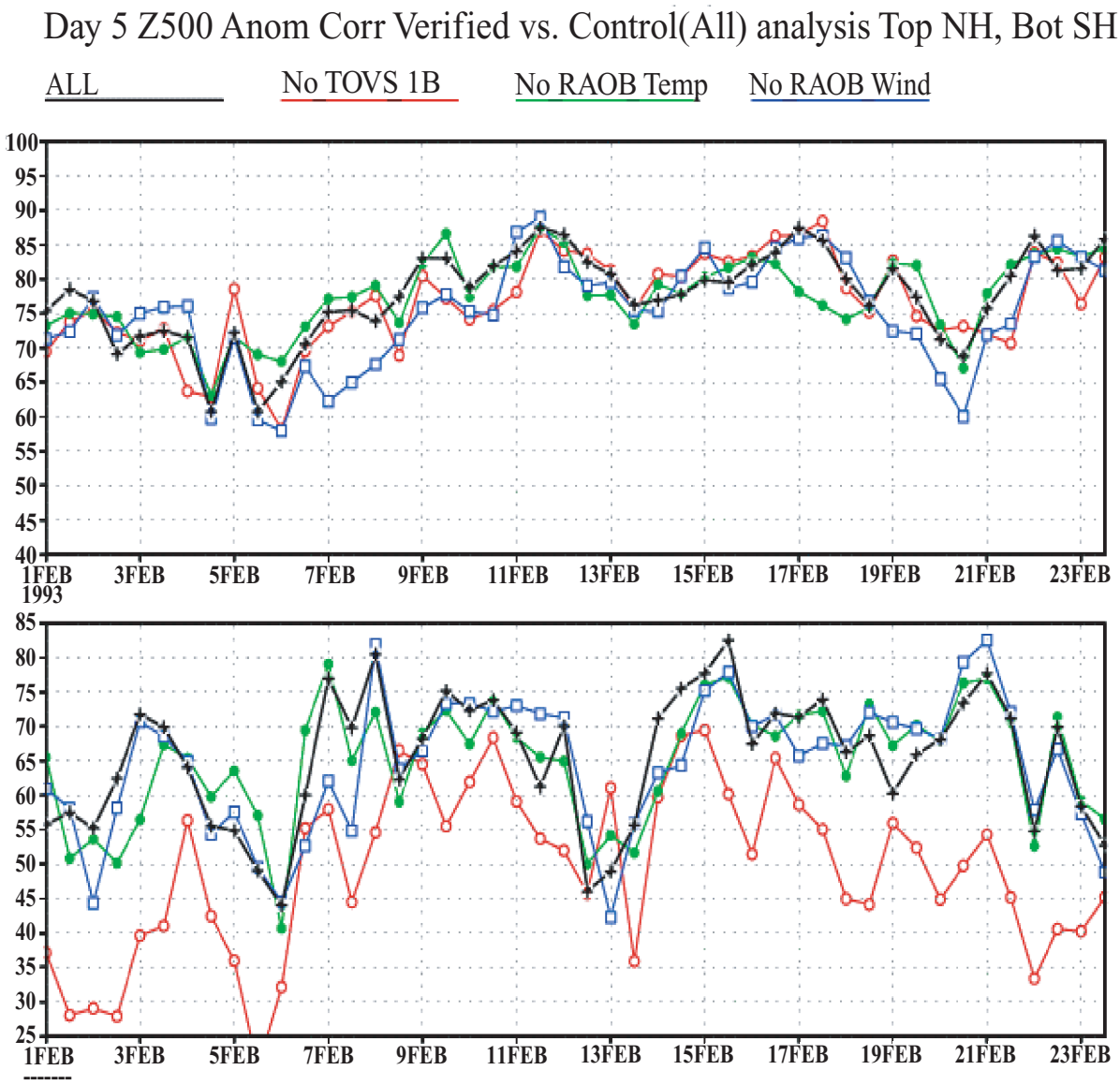

Figure 5.7.1: Verification (anomaly correlation) for the 5-day forecasts performed every 12 h with the NCEP data assimilation system with different combinations of data: crosses, all data; open circles, no TOVS radiances; open squares: no rawinsonde winds; full circles, no rawinsonde temperatures. (Courtesy of M. Masutani.)

their linear terms (on the left-hand sides) and the nonlinear terms (on the right-hand sides):

$$\left.\begin{aligned} \frac{\partial u}{\partial t} &- f\mathrm{v} + \frac{\partial \phi}{\partial x} = R_u \\ \frac{\partial v}{\partial t} &+ f\mathrm{u} + \frac{\partial \phi}{\partial y} = R_v \\ \frac{\partial \phi}{\partial t} &+ g\mathrm{D}\left(\frac{\partial u}{\partial x} + \frac{\partial v}{\partial y}\right) = R_\phi \end{aligned}\right\} \tag{5.7.8}$$

For the SWEs the nonlinear terms R_u, R_v, R_ϕ are

$$\left.\begin{aligned} R_u &= -u\frac{\partial u}{\partial x} - v\frac{\partial u}{\partial y} \\ R_v &= -u\frac{\partial v}{\partial x} - v\frac{\partial v}{\partial y} \\ R_\phi &= -u\frac{\partial \phi}{\partial x} - v\frac{\partial \phi}{\partial y} \end{aligned}\right\} \tag{5.7.9}$$

Note that the left-hand sides contain linear terms about a basic state of rest.

The first step is to determine the (linear) normal modes, and for this purpose we find the eigensolutions or normal modes of (5.7.8) setting $R_u = R_v = R_\phi = 0$. In the case of a doubly periodic domain the normal modes are simply of the form

$$\begin{bmatrix} U_{kl} \\ V_{kl} \\ F_{kl} \end{bmatrix} e^{i(kx+ly)} e^{-i\omega t} = X_{kl} e^{-i\omega t} \tag{5.7.10}$$

In a three-dimensional model there would be an additional functional dependence on the vertical, which can be represented by a vertical wavenumber. If we plug (5.7.10) into (5.7.8) with zero on the right-hand side, we obtain three solutions for the frequency:

$$\left.\begin{aligned} &\omega_S = 0 \\ &(\omega_F)^2 = f^2 + gD(k^2 + l^2) \end{aligned}\right\} \tag{5.7.11}$$

where the subscripts S and F refer to the slow (quasi-geostrophic) and the fast (inertia-gravity wave) modes. Because the equations are homogeneous, (U, V, F) are related to each other and the amplitude of the normal mode is arbitrary. For the slow modes, the relationship is $U_{klS} = -ilF_{klS}/f$, $V_{klS} = ikF_{klS}/f$ (geostrophic balance). For the fast modes, a similar but more complicated relationship exists which, if $f^2 \ll gD(k^2 + l^2)$ reduces to

$$U_{\pm klF} = \pm\frac{kF_{\pm klF}}{\sqrt{gD(k^2 + l^2)}} \qquad V_{\pm klF} = \pm\frac{lF_{\pm klF}}{\sqrt{gD(k^2 + l^2)}}$$

The normal modes in (5.7.10) can therefore be written as a set of slow (Y_{kl}) and fast (Z_{kl}) modes:$X_{klm} = (Y_{kl}, Z_{kl})$ where the subscript m is S for slow modes and F for fast modes. The normal modes X_{klm} constitute a complete orthonormal basis if we normalize them by their total energy:

$$\int_x \int_y [gD({U_{klm}}^2 + {V_{klm}}^2) + {F_{klm}}^2]dxdy = 1 \tag{5.7.12}$$

Any time-dependent field of winds and heights can be expanded in terms of the normal modes:

$$\begin{bmatrix} u(x,y,t) \\ v(x,y,t) \\ \phi(x,y,t) \end{bmatrix} = \sum_k \sum_l \sum_{m=S,F} a_{klm}(t) \begin{bmatrix} U_{kl} \\ V_{kl} \\ F_{kl} \end{bmatrix}_m e^{i(kx+ly)} \tag{5.7.13}$$

where the (time-dependent) coefficients can be determined from a back Fourier transform of (5.7.13):

$$a_{klm}(t) = \int_x \int_y [gD(uU_{klm} + vV_{klm}) + \phi F_{klm}]e^{-i(kx+ly)}dydx \tag{5.7.14}$$

Now we return to the full nonlinear SWE (5.7.8), multiply the three equations by

$$\left.\begin{array}{l} gDU_{klm}e^{-i(kx+ly)}e^{+i\omega_{klm}t} \\ gDV_{klm}e^{-i(kx+ly)}e^{+i\omega_{klm}t} \\ F_{klm}e^{-i(kx+ly)}e^{+i\omega_{klm}t} \end{array}\right\} \tag{5.7.15}$$

respectively, add them and integrate over the domain to obtain the nonlinear equations for the amplitudes of the slow and fast modes:

$$\left.\begin{array}{rcl} \dfrac{da_{klS}}{dt} & + & i\omega_{klS}a_{klS} = R_{klS} \\ \dfrac{da_{klF}}{dt} & + & i\omega_{klF}a_{klF} = R_{klF} \end{array}\right\} \tag{5.7.16}$$

Here we have separated slow and fast modes. $R_{klm}(Y,Z)$ is the result of applying this operation to the right-hand side nonlinear terms and depends on both the vector $\mathbf{Y}$ of slow modes coefficients and on the vector $\mathbf{Z}$ of fast modes coefficients. Recall (5.7.11) that for this simple geometry, $\omega_{klS} = 0$.

$$\frac{da_{klF}}{dt} + i\omega_{klF}a_{klF} = 0 \tag{5.7.17}$$

is the linear equation for the fast modes.

We can choose to perform a *linear normal mode initialization* by zeroing out the initial amplitude of the fast modes: $a_{klF}(t) = a_{klF}(0)e^{-i\omega_{klF}t} = 0$ in (5.7.17). This will

make the *linear* time derivative of the fast modes equal to zero. But (5.7.16) shows that this initialization, which is equivalent to a perfect geostrophic balance, is not accurate enough for the realistic nonlinear case. The presence of nonlinear forcing will generate fast oscillations even if the basic state is geostrophic. Therefore Machenhauer (1977) suggested instead *to zero out the time derivative* of the fast modes in the nonlinear equation for the fast modes in (5.7.16), whose right-hand side depends on both slow and fast modes. Therefore, from

$$\frac{da_{Fkl}}{dt} + i\omega_{Fkl}a_{Fkl} = R_{Fkl}(Y(0), Z(0)) \tag{5.7.18}$$

we obtain the *nonlinear normal mode initialization* condition:

$$a_{Fkl}(0) = \frac{R_{Fkl}(Y(0), Z(0))}{i\omega_{Fkl}} \tag{5.7.19}$$

Since the coefficient $a_{klF}(0)$ appears both in the left- and right-hand sides of (5.7.19) in a component of the vector of fast coefficients Z, Machenauer (1977) suggested iterating equation (5.7.19) until convergence.

Nonlinear normal mode initialization (NLNMI) has been widely used in many operational data assimilation systems, since it is quite effective in substantially reducing the amplitude of the inertia-gravity waves from the initial conditions, much better than a simple geostrophic balance. It requires the determination of the linear normal modes of a model, but Temperton (1988) derived a formulation denoted "implicit NLNMI" without this requirement.

NLNMI has some problems, however:

- In the tropics, diabatic heating plays a fundamental role, essentially balancing the vertical advection of static stability. Therefore, diabatic forcing has to be included in the nonlinear terms, and this requires estimating the heating from short-term forecasts (Wergen, 1988).

- There is some arbitrariness in defining which "fast" modes need to be initialized. For example, inertia-gravity waves with high vertical wavenumbers are quite slow (see the discussion in the previous subsection), so that only the first few vertical modes are usually initialized. On the other hand, NLNMI eliminates the high-frequency but real atmospheric tides from the solution, since they appear as fast modes. This requires a special handling of these modes.

- NLNMI is only an approximation of the true slow evolution of the atmosphere: if we apply NLNMI to a model that has been running for a day or longer, ideally it should not modify it, since it has already reached slow modes equilibrium. However, NLNMI will change the initial fields significantly.

Ballish *et al.* (1992) developed a modification of the procedure denoted *incremental* NLNMI, in which the initialization is applied to the analysis increments, rather than to

the full analysis field. This procedure is able to substantially solve the three problems indicated above.

In recent years the use of NLNMI after the analysis step has become less popular because of the development of three alternative approaches. The first one is the use of 3D-Var, which allows the introduction *within the cost function* of a term that penalizes the lack of balance. Parrish and Derber (1992) included a penalty term based on the global linear balance equation applied on the analysis increments. They found that this, combined with the use of a more realistic global background error covariance based on differences between 24- and 48-h forecasts verifying at the same time, yielded an analysis that was well balanced. As a result, the NLNMI step became unnecessary in the NCEP system. This is a major advantage of 3D-Var over the standard OI procedure followed by NLNMI. It eliminates the artificial separation of the analysis step, which produces fields that are close to the observations but out of balance, and the initialization step, which produces fields that are balanced but further away from the observations.

Another method that seems to achieve similar balanced results and minimize the spin-up problems is the *incremental analysis update* (Bloom *et al.*, 1996), in which the analysis increment is added in small "drips" throughout the 6-h forecast rather than once as a large change at the analysis time. Assume that there are n time steps in the 6-h forecast. In the incremental analysis update the analysis increments are computed at the analysis time. Then the forecast at the analysis time minus 3 h is integrated for 6 h adding at each time step the analysis increment divided by n, until the forecast reaches the analysis time plus 3 h. At that time a preliminary 3-h integration of the model without analysis increments is performed until the next analysis time is reached, and the cycle is repeated. The overhead of this method is only the additional preliminary integration of the model during the second half of the interval between analysis times.

A third development is the introduction of digital filter initialization, a variation of dynamic initialization that has proven to be very simple and efficient.

5.7.3 Dynamic initialization using digital filters

Some numerical schemes, like the Euler-backwards or Matsuno scheme, damp high frequencies, and this property has been used in order to reduce the accumulation of high-frequency noise within an assimilation cycle. Assume we have an equation

$$\frac{du}{dt} = -i\omega u \tag{5.7.20}$$

The Matsuno scheme is a predictor–corrector type of scheme (see Table 3.2.1) where

$$\left.\begin{aligned} \tilde{u} &= u_n - i\omega\Delta t u_n \\ u_{n+1} &= u_n - i\omega\Delta t\tilde{u} \end{aligned}\right\} \tag{5.7.21}$$

so that $u_{n+1} = \rho u_n$, where the amplification factor is $\rho = (1-\omega^2\Delta t^2) - i\omega\Delta t$. Therefore

$$|\rho|^2 = (1 - \omega^2\Delta t^2 + \omega^4\Delta t^4) \tag{5.7.22}$$

It is evident from (5.7.22) that as long as the CFL stability condition $|\omega_{max}\Delta t_{CFL}| \leq 1$ is satisfied, high frequencies are damped at every Matsuno time step. This has been found to be reasonably satisfactory for avoiding excessive accumulation of noise in the analysis cycle (e.g., Halem *et al.*, 1982), but the damping is slow except for high frequencies close to $1/\Delta t$. Using a damping time scheme does not balance the initial fields; rather the balance is achieved only after integrating the model for a while (e.g., 6 h in the analysis cycle). Several dynamical initialization methods using forward/backward integrations were suggested to balance the initial field (e.g., Nitta and Hovermale, 1969, Okamura, 1969). Grant (1975) suggested a more efficient dynamic initialization based on linear combinations of forward/backward integrations, with combinations of time steps some of which are longer than allowed by the CFL condition. Dynamic initialization never became widely used, despite its simplicity, because it is not efficient, requiring many forward/backward iterations to substantially reduce medium frequency waves.

Exercise 5.7.1. *Show that the net damping at $t = 0$ after the application of a Matsuno time step followed by another Matsuno time step integrating backwards in time (changing the sign of Δt) is given by (5.7.22).*

Exercise 5.7.2. *Find the damping of the Okamura scheme: $u^+ = u^n(1 + i\omega\Delta t); u^- = u^+(1 - i\omega\Delta t); u^{n+1} = 2u^n - u^-$. The Okamura–Rivas scheme is the same except that the time step cycles over three iterations: $\Delta t = \Delta t_{CFL}; 1.4\Delta t_{CFL}; 2\Delta t_{CFL}$, resulting in an even faster damping (Grant, 1975).*

The introduction of dynamic initialization based on digital filtering by Lynch and Huang (1992) and Lynch (1997) has changed this situation substantially, and essentially eliminated the need for NLNMI. In a digital filter, the model is integrated forward and backward in time between $-t_M$ and t_M, as in regular dynamic initialization. The difference is that the model fields are used at every time step to compute a weighted average valid at the initial time $t = 0$, and the weights are optimally chosen in order to damp high frequencies, rather than simply using a damping time scheme.

The idea of a digital filter is to choose the filtering weights in such a way that for low frequencies $\omega \ll \omega_s$ the amplitude of the solution with that frequency remains mostly unchanged, whereas for high frequencies $\omega \geq \omega_s$ the amplitude is substantially reduced. Given a time step Δt and the corresponding threshold computational frequency $\theta_s = \omega_s \Delta t$, low frequencies are characterized by $0 \leq |\theta| \ll \theta_s$, and the high frequencies that we want to filter by $\theta_s \leq |\theta| \leq \pi$.

One digital filter approach (Lynch, 1997) is based on a Dolph–Tchebychev filter, which is close to optimal, using the properties of Tchebychev polynomials:

$$T_n(x) = \begin{cases} \cos(n \cos^{-1} x) & \text{if } |x| \leq 1 \\ \cosh(n \cosh^{-1} x) & \text{if } |x| > 1 \end{cases} \tag{5.7.23}$$

from which $T_0(x) = 1$ and $T_1(x) = x$. The higher polynomials can be obtained from the recurrence relationship

$$T_n(x) = 2xT_{n-1}(x) - T_{n-2}(x) \quad n \geq 2 \tag{5.7.24}$$

The following function is ideal for such damping:

$$H(\theta) = \frac{T_{2M}[x_0 \cos(\theta/2)]}{T_{2M}(x_0)} \tag{5.7.25}$$

where $x_0 = 1/\cos(\theta_S/2)$, since for the low-frequency range, $H(\theta)$ falls from 1 to $r = 1/T_{2M}(x_0)$ as $|\theta|$ goes from 0 to θ_s, and for the high frequency range $\theta_S \le |\theta| \le \pi$, $H(\theta)$ oscillates within $\pm r$ (e.g., see Figure 5.7.2).

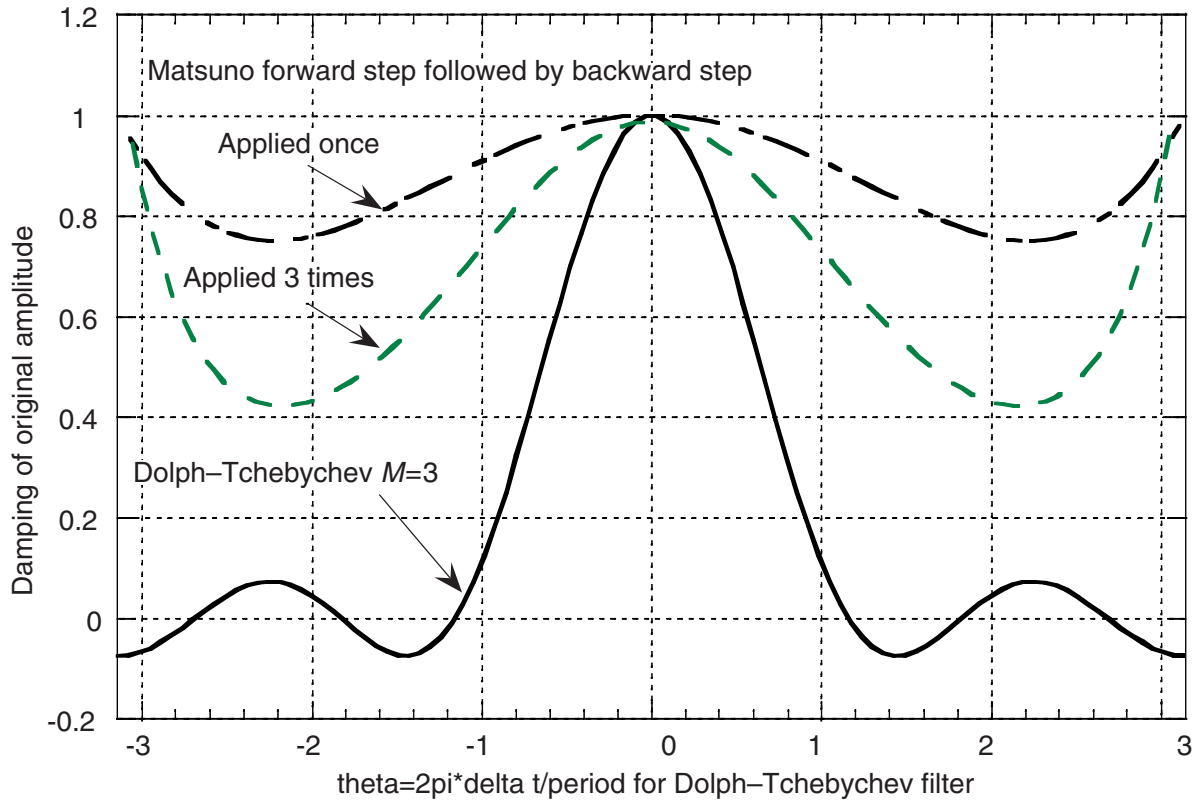

Figure 5.7.2: Comparison of the damping of high frequencies using a Dolph–Tchebychev filter with $\theta_s = \pi/3, M = 3$, and a forward/backward Matsuno dynamical initialization.

From the definition of the Tchebychev polynomials it can be shown that

$$H(\theta) = \sum_{n=-M}^{+M} h_n e^{in\theta} \tag{5.7.26}$$

where the coefficients are

$$h_n = \frac{1}{2M+1}\left[1 + 2r\sum_{m=1}^{M} T_{2M}\left(x_0 \cos\frac{\theta_m}{2}\right)\cos m\theta_n\right] \tag{5.7.27}$$

The solution of the model, integrated from $-t_M$ to t_M, is weighted averaged:

$$\overline{u}(0) = \sum_{n=-M}^{M} h_n u_n \tag{5.7.28}$$

so that at the end of this procedure the amplitude of each frequency in $\overline{u}$ is modulated by $H(\theta)$. The parameters are chosen in the following way: Choose a period τ_S such that waves with periods shorter than this are to be filtered. The cut-off frequency is then given by $\theta_s = 2\pi\Delta t/\tau_s$. The time span of the integration $T_s = 2t_M = 2M\Delta t$ gives a filter of order $N = 2M + 1$.

In practice, Lynch (1997) recommended to first perform a backward integration with just the *dry adiabatic dynamics* (since they are reversible), from $t = 0$ to $t = -T_S$. An application of the weighted average (5.7.28) gives a filtered field centered at $t = -T_S/2$. Then a forward integration from $t = -T_S/2$ to $t = +T_S/2$ using the *full model with physics* results in a field centered at $t = 0$, *filtered for the second time including filtering of the effects of irreversible diabatic processes*. In the high resolution-limited area model, with a time step of half an hour, and a filtering period of 3 h, the time span is also half an hour, with $N = 2M = 7$. The double filter gives a reduction in energy of the high frequencies of more than 99%.

Figure 5.7.2 shows the response of dynamic initialization for this case, using the Dolph–Tchebychev filter with $\Delta t = 30\,\text{min}, \theta_s = \pi/3$ (filtering periods shorter than 3 h) and $M = 3$. It compares this with the result of using one and three iterations of the forward/backward Matsuno time step.

Another digital filter is based on the Lanczos filter, which is widely used to filter out high frequencies from time series (Duchon, 1979, Lynch and Huang, 1992). The filter is similar to the Dolph–Tchebychev one, but now the Lanczos weights in (5.7.28) are given by

$$h_n = \frac{\sin(n\theta_{crit}\Delta t)}{\pi n}\,\frac{\sin(n\pi/M)}{n\pi/M} \tag{5.7.29}$$

Figure 5.7.3 compares the responses of the Dolph–Tchebychev and Lanczos filters requiring the same number of time steps (6), since for the Lanczos filter the coefficients for $n = 4$ and 5 are zero. The response for the Lanczos filter with $\theta_{crit} = \pi/4, M = 5$ is competitive with the Dolph–Tchebychev filter.

In summary, initialization using digital filtering is a very simple process that avoids the determination of the model normal modes and the need of NLNMI. Most importantly, it dampens the high-frequency solutions according to their actual model-determined frequency rather than from an arbitrary separation into inertia-gravity waves and quasi-geostrophic modes. It does not make any additional approximation and it can include the full diabatic and nonreversible effects to determine the classification of slow and fast modes.

Exercise 5.7.3. *Is it better to achieve balance within 3D-Var or to apply a digital filter afterwards?*

5.8 Quality control of observations

The reported atmospheric observations used in data assimilation are not perfect; they contain several kinds of errors, including instrumental errors and errors of human origin.

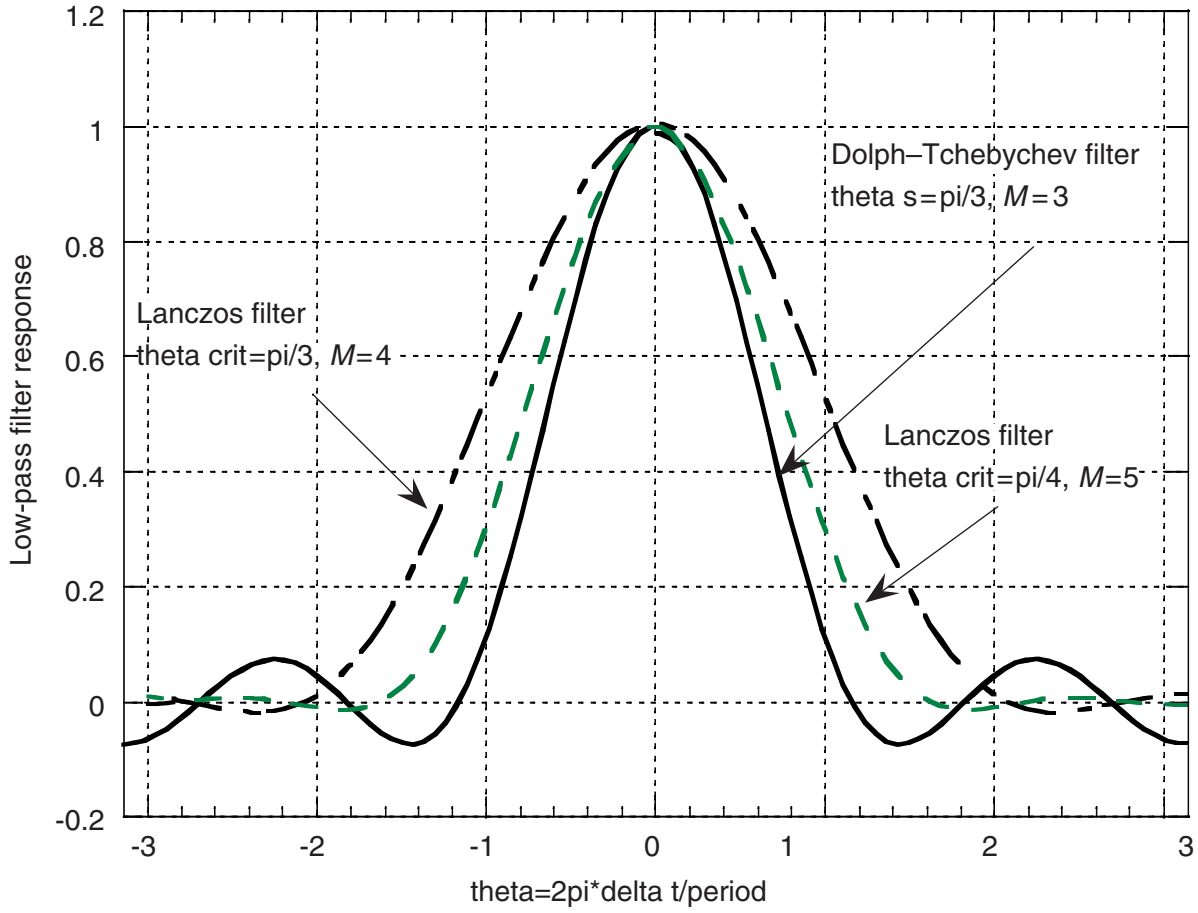

Figure 5.7.3: Comparison of the responses of the Dolph–Tchebychev filter, with $\theta_s = \pi/3, M = 3$, and the Lanczos filter, with $\theta_{crit} = \pi/3, M = 4$ and $\theta_{crit} = \pi/4, M = 5$. All the choices require six time integrations and are therefore computationally comparable.

The reported observations may also contain "errors of representativeness", i.e., actually correct observations may reflect the presence of a subgrid-scale atmospheric phenomenon that cannot be resolved by the model or the analysis. The representativeness error indicates the observation is not representative of the areally averaged measurement required by the model grid. The instrumental and representativeness errors can be systematic or random. Systematic errors and biases should be determined by calibration or other means such as time averages. Random errors are generally assumed to be normally distributed.

In addition to randomly distributed errors, the reported observations may contain errors that are so large that the observations have no useful information content and should be tossed out. Frequently, these rough or gross errors are of human origin, and take place during the computation or the transmission of the observation. There are other sources of observation errors (wrong date, time or location, uncalibrated instruments, etc.). The use of an observation with a rough error can cause a disproportionally big error in the analysis, so there has been a tendency to use observations conservatively

("when in doubt, throw it out"). In recent years, however, quality control systems have become more sophisticated and many observations that would have been thrown out in the past are now corrected, resulting in an improvement in the initial conditions, and hence in the forecasts. Newer quality control systems allow for a continuous weighting of the observation suspected of having a gross error, rather than a "yes or no" decision to toss it out.

Quality control is based on a comparison between observations and some kind of expected value (which could be based on climatology, an average of nearby observations, or the first guess). The difference between the expected value and the reported observation is denoted the "residual". If the residual is very large (measured in standard deviations of the estimate), the observation may be considered to be erroneous. The more sophisticated the estimate of the expected value (i.e., the smaller its expected standard deviation with respect to the true value), the more discriminating the quality control algorithm will be, i.e., the better it will be able to distinguish between observations with large errors, which should not be given credence, and correct observations reporting unusual states of the atmosphere, such as very low pressures or unusually high winds reported within an area affected by a tropical cyclone. In the latter case it is important to keep the observation in order to improve the initial conditions of the forecast.

Earlier quality control systems were based on several checks performed in series (one after another) before the analysis. For example DiMego *et al.* (1985) compared each observation with a climatological distribution to see whether it was within a reasonable range ("gross check"). If the reported value was outside a prescribed range, or differed from the climatological mean by more than, for example, five standard deviations, the observation was tossed out. If the observation survived this test, it was then compared to the average of nearby observations, and again tossed out if outside a reasonable range. This check (called a "buddy check", comparing the observations with their "buddies") could also salvage an observation previously tossed out even if it was quite different from the expected climatological value.

An analysis is, in principle, more accurate than either the first guess or the observation. This led to the development of an "OI" quality control: each observation is compared with a simple OI value that would be obtained at the observation location using the first guess (background) field and nearby observations, but without including the observation being checked (Lorenc, 1981, Woollen, 1991). When the residual (difference between the observed and analyzed value) is larger than a certain number of analysis error standard deviations, it is tossed out. The analysis is iterated so some observations may be salvaged after first being tossed out.

Gandin (1988) introduced the idea of *complex quality control* (where the word "complex" means that it uses several tests simultaneously rather than in series). The basic idea is to estimate several independent residuals and then apply a decision making algorithm based on the information provided by all the independent residuals, rather than performing decisions either in a sequential order, or as a single OI quality control check using all the information at once. Collins and Gandin (1990) and Collins (1998) applied this approach to rawinsonde heights and temperatures with great success. The power of this method lies in the fact that several independent checks can support each

other and reduce the level of uncertainty. Furthermore, if the residuals are large and agree reasonably well with each other, they provide the basis for a correction of the observation.

The independent residuals obtained from the different checks used at NCEP for the complex quality control of rawinsondes temperature and heights are: (1) Incremental check (the residual is the increment between the reported observation and the 6-h forecast) (2) Horizontal check based on a simple OI horizontal analysis of the increments, using one observation per quadrant, and only observations within 1000 km of the reported observation. The horizontal residual is the difference between the increment at the observation location and the horizontally interpolated value. (3) Vertical check: the vertical residual is the difference between the observed increment and the increment interpolated vertically from the nearest data points for the same station, one above and one below. (4) Hydrostatic check, the most powerful of the checks since it takes advantage of the redundancy between temperature and height information reported in rawinsonde observations. The hydrostatic residual is the difference between the values of the thickness of a layer between mandatory level heights, calculated using the reported heights, and the thickness calculated independently from the reported virtual temperatures. (5) Baseline check computed by making a hydrostatic computation downward, from the first mandatory level above the surface with complete heights and temperatures, to the reported surface pressure. The baseline residual is the difference between the station elevation, given by the report, and the hydrostatically determined height at the surface pressure. Another possible check used at NCEP is based on a temporal interpolation of observations at the same station 12 or 24 h before and after the observation time. This check is particularly useful for isolated stations within a "reanalysis" mode. It is also possible to perform a check of the stability of the lapse rate.

Collins and Gandin (1990) and Collins (1998) developed a sophisticated decision making algorithm that makes generally confident decisions correcting computation or communication errors of human origin. They assumed that human errors have a simple structure: a single digit or a sign is wrong or missing. The following example of such error detected at a single level in the heights is typical:

Example of CQC and the decision making algorithm correction of gross errors (Collins and Gandin, 1998)

- Reported 1000 hPa height: 8 m
- Computed residuals:
  - Incremental residual: −72 m;
  - Vertical residual: −66 m;
  - Hydrostatic residual (using significant levels): 60 m;
  - Hydrostatic (using only mandatory levels): 65 m;
  - Baseline residual: −58 m.

With this information, the decision making algorithm concluded that a simple correction changing one digit (adding 60 m to make the observation 68 m instead of the reported 8 m) could be confidently made.

The NCEP global system has also an OI-based complex quality control for all other data (winds, moisture, satellite retrievals) that makes the first three checks discussed above and rejects observations with large residuals but does not attempt to make corrections, since the strong redundancy of the hydrostatic check is not available for variables other than rawinsonde temperatures and heights (Woollen, 1991).

The effect of modern quality control systems is difficult to gauge, but Kistler *et al.* (2001) showed an impressive example of the positive impact from the modern approach compared with the quality control that was operational at NCEP in the 1970s. They pointed out that in 1974 NMC (now NCEP) introduced a modern observation formatting system (known as Office Note 29, ON29), which later became the basis of the official World Meteorological Organization (WMO) Binary Universal Format Representation system for the encoding of observations. ON29 included more information about the observation than previously used encoders. This change in formatting required a complete overhaul of the NMC decoding system, and errors must have been introduced during this complex reprogramming process. The NMC operational forecast skill actually went down and it took a few years before it recovered to the pre-1974 error levels (Kalnay *et al.*, 1998). During the production of the NCEP/NCAR reanalysis (Kalnay *et al.*, 1996), both the complex quality control for rawinsonde heights and temperatures and the OI quality control for other observations were used to screen observations. The complex quality control found and corrected an unusually large number of rawinsondes errors starting in 1974, and, presumably as a result of this correction and the OI quality control screening of the other information, the benefit of using the more advanced formatting system ON29 became realized. In the reanalysis, the forecast skill increased substantially in 1974, rather than deteriorating as in the operational forecasts.

Another approach that has also become popular is variational quality control performed within 3D-Var or 4D-Var, rather than before the analysis, like OI and complex quality controls (Purser, 1984, Lorenc and Hammon, 1988, Ingleby and Lorenc, 1993, Andersson and Jarvinen, 1999, Collins, 2001a,b). It has the advantage that it is performed as part of the analysis itself, rather than as a preprocessing step like OI quality control, but because it computes a single (iterative) residual for each observation, it is not able to correct observations like complex quality control.

The variational quality control approach is based on modifying the observational component J_o of the variational cost function $J = J_b + J_o$ to take into account the possibility of gross errors. Note that in the variational analysis approach, the gradient $\boldsymbol{\nabla}_{y_o} J_o$ of the cost function with respect to an observation y_o determines how quickly the analysis estimate $\mathbf{x}$ will shift towards that observation (Figure 5.8.1(b)).

Consider the cost function term for a single uncorrelated observation y_o without al-

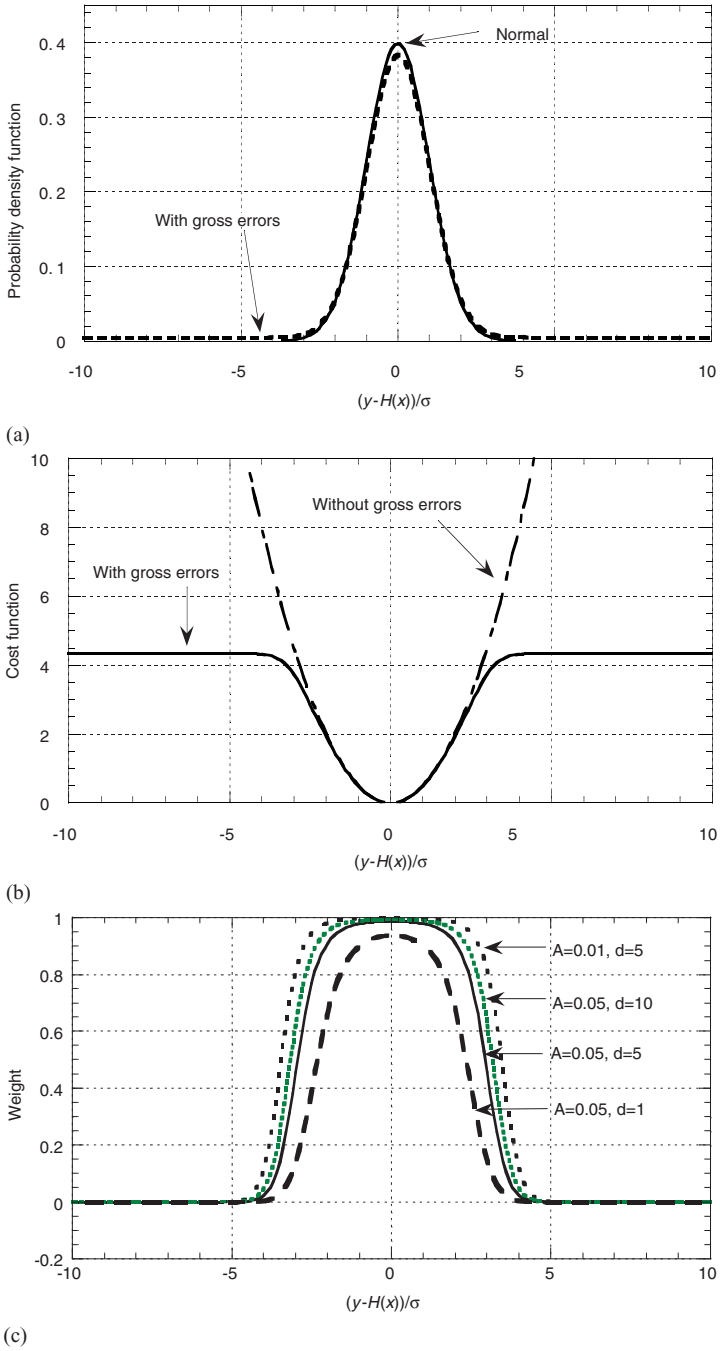

Figure 5.8.1: (a) Probability density function for an observation with a given scaled residual, without gross errors (normal distributions) and with gross errors. $A = 0.05, d = 5$. (b) Observational cost function derived from the logarithm of the error distributions in (a). (c) Weight factor applied to the gradient of the cost function for different values of the *a priori* probability of gross errors A and the width of the flat distribution.

lowing for gross errors:

$$J_o^N = \frac{1}{2}\left[\frac{y_o - H(x)}{\sigma_o}\right]^2 \text{ with } \boldsymbol{\nabla} J_o^N = \left[\frac{y_o - H(x)}{\sigma_o^2}\right] \tag{5.8.1}$$

where x is a close approximation of the true value (analysis). In variational quality control one assumes that there is an *a priori* probability A of having a gross error, estimated from past statistics. $1 - A$ is then the *a priori* probability of not having a gross error, in which case the observation is assumed to have random errors with a Gaussian distribution. Without gross errors ($A = 0$), the probability that an observation y_o, if $H(x)$ is the true value, is given by the normal distribution is

$$p(y_o) = N = \frac{1}{\sqrt{2\pi}\sigma_o} e^{-J_o^N} \tag{5.8.2}$$

where (as discussed in Section 5.3) $J_o^N = -\ln N + const.$ (Figure 5.8.1(a), full line) and the constant is chosen arbitrarily to make $J^N(0) = 0$.

If there are gross errors with an *a priori* probability A (let's say, 0.05), then the probability of an observation y_o (Figure 5.8.1(a), bold dashes) is modified:

$$p^{QC}(y_o) = (1 - A)N + AF \tag{5.8.3}$$

where F is a flat (uniform) distribution for the gross errors:

$$F = \begin{cases} \dfrac{1}{2d\sigma_o} & \text{if } |y_o - H(x)| < d\sigma_o \\ 0 & \text{otherwise} \end{cases} \tag{5.8.4}$$

Here d is the maximum number of standard deviations allowed for gross errors (e.g., $d = 5$). If $|y_o - H(x)| > d\sigma_o$, it is assumed that the observation was so obviously wrong that it was eliminated in a preliminary check against climatology or the background. The integral of F is therefore equal to 1.

We can then modify the contribution to the cost function made by the observation by including the probability of gross errors: $J_o^{QC} = -\ln p^{QC}(y_o) + const.$ (Figure 5.8.1(a), full line). Since the flat probability distribution of gross errors does not depend on the value y_o, the modified gradient of the cost function is:

$$\begin{aligned} \boldsymbol{\nabla}_{y_o} J_o^{QC} &= \boldsymbol{\nabla}[-\ln p^{QC}(y_o)] = \frac{1 - A}{p^{QC}(y_o)} N \boldsymbol{\nabla} J_o^N \\ &= \frac{(1 - A)N}{(1 - A)N + AF} \boldsymbol{\nabla} J_o^N = (1 - P)\, \boldsymbol{\nabla} J_o^N \end{aligned} \tag{5.8.5}$$

where

$$P = \frac{AF}{(1 - A)N + AF} \tag{5.8.6}$$

is the *a posteriori* probability of having a gross error (after making the observation y_o). In other words, the gradient of the cost function including variational quality control is the gradient without variational quality control multiplied by a weight

$$W^{QC} = \frac{(1-A)N}{(1-A)N + AF} \tag{5.8.7}$$

(the probability of NOT having a gross error), which is close to 1 for $|y_o - H(x)|/\sigma_o < d$ and goes to zero for $|y_o - H(x)|/\sigma_o \geq d$. Figure 5.8.1(c) shows that the weights are not very sensitive to the choice of parameters. A small value for the *a priori* probability of gross errors A will result in a steeper reduction of weights. Because of the shape of the normal probability function, the weights are not too sensitive to d unless the value chosen is rather small (e.g., $d = 1$ in the graph).

Because $H(x)$ is assumed to be close to the truth, at ECMWF variational quality control is not turned on (i.e., the weight multiplying the gradient during the minimization is $W = 1$) during the first 40 iterations of the 4D-Var algorithm, while the solution starts to converge towards the analysis. It is then turned on ($W = W^{QC}$) for the last 30 iterations of the 4D-Var, thus giving less weight to observations that are likely to contain gross errors (Andersson and Jarvinen, 1999).