6 Model initialization

6.1 Background

As we have seen in Chapter 3, solving the equations that govern the physical systems that we are modeling is an initial- and boundary-value problem. The lateral, upper, and lower boundary conditions are discussed in Chapters 3 and 5. In this chapter will be described the procedure by which observations are processed to define initial conditions for the model dependent variables, from which the model integration begins. This process is called model *initialization*.[1] There are essentially two requirements for the initialization. First, the dependent variables defined on the model grid must faithfully represent conditions in the real atmosphere (e.g., fronts should be in the correct location), and second, the gridded mass-field variables (temperature, pressure) and momentum-field variables (velocity components) should be dynamically consistent, as defined by the model equations. An example of the mass–momentum consistency requirement is that, on the synoptic scale, the gridded initial conditions should be in approximate hydrostatic and geostrophic balance. If they are not, the model will generate potentially large-amplitude inertia–gravity waves after the initialization shock, and these nonphysical waves will be overlaid on the meteorological part of the model solution until the adjustment process is complete. The final adjusted condition will prevail after the inertia–gravity waves have been damped, or have propagated off the grid of a LAM. However, the model solution will be typically unusable during this adjustment period, which is one reason for the common, historical recommendation that model output not be used for about the first 12 h of the integration. On the smaller mesoscale and convective scales, ageostrophic circulations, such as associated with horizontally differential surface forcing and convection, should ideally exist in the initial conditions. Otherwise, such features will need to *spin up* during the early period of the model integration.

Historically, there have been two approaches for accomplishing the initialization, although modern methods have blurred the distinctions between them. One is called *static initialization*, where observations are first interpolated to a model grid (*data analysis*), and then the resulting variables may be adjusted using diagnostic, dynamical constraints to

[1] The terminology here is not used consistently in the community. Some employ the term initialization to refer only to the process of defining a dynamic balance in the initial conditions.

make them more consistent with each other and with the model equations. In contrast to this diagnostic method, *dynamic initialization* involves the preforecast integration of the model to produce an initial state that is dynamically consistent with the equations used for the forecast.

The commonly used terms *data assimilation* and data analysis both refer to processes that employ observations to construct a gridded data set that defines the spatial variability of model dependent variables at the initial time of a forecast. However, the expression data assimilation typically means that a meteorological model is employed, where many approaches will be discussed throughout this chapter. The objective of data assimilation can be the production of initial conditions for operational forecasts, or the construction of long-term reanalyses of the state of the atmosphere (see Chapter 16 for an explanation of the latter).

Because the initialization of the land surface was treated in the last chapter in the context of LDASs, this subject will not be treated here. But, it should be remembered that these land-surface variables, such as substrate temperature and moisture, the state of the vegetation, etc., are time dependent, and their accurate specification in the model initial conditions is an important part of the initialization process.

6.2 Observations used for model initialization

6.2.1 Sources of observations used for model initialization

Meteorological observations can be classified as either *in situ* or remotely sensed. Obtaining the former involves the use of sensors that measure the local value of a variable. Remote sensing employs sensors that perform measurements from a distance, through the use of either active or passive methods. Passive methods employ the measurement of naturally emitted radiation. With active methods, the sensing system emits radiation and measures the response of the atmosphere to that radiation. Radiosondes are examples of *in-situ* sensors. Satellite-borne radiometers that measure radiances (the emissions spectrum) from the atmosphere are passive remote sensors, while radars that emit microwave energy and measure that fraction which is reflected by hydrometeors are active remote sensors. For either remote-sensing method, a *retrieval algorithm* is often needed in order to translate the information obtained by the sensor into meteorologically useful information (values of dependent variables). In the case of the radiometer, the algorithm translates the sensor data into temperature, and for the radar data the echo strength is converted to precipitation intensity. In contrast to the use of retrieval algorithms, variational-analysis methods, to be discussed later, allow the direct use of raw sensor information in the analysis process. Non-satellite-based measurement platforms that are commonly used to provide initial conditions for NWP models are listed below.

- *Radiosondes* – Measure temperature, relative humidity, and pressure; and tracking the balloon displacement provides wind speed and direction. This is still the primary method for defining the three-dimensional structure of the atmosphere on the synoptic and global scales, for model initialization. Even though the most-common frequency for radiosonde ascents is every 12 h, at 0000 UTC and 1200 UTC, in some countries it is every 24 h. These two standard radiosonde launch times define the most-common initialization times for models.

- *Near-surface weather stations* – Typically measure temperature, humidity, pressure, wind speed, wind direction, and precipitation. A challenge associated with using these observations is that it is difficult to estimate how far into the model boundary layer to spread their influence (i.e., over how many model levels) when defining the initial conditions on the model grid. This is important to know because vertical mixing in the model can quickly eliminate near-surface information that is incorporated in the initial conditions, if the atmosphere above is not analyzed with vertically consistent structures. Another challenge is that it is difficult to consider these observations in the context of any dynamic balance, simply because of the dominance of local forcing. Near-surface variables may be reported at intervals of 5 minutes, 15 minutes, 1 hour, 3 hours, or 6 hours. For near-surface winds, the averaging that is done to remove turbulence can be over periods of 5 minutes to 15 minutes. The height above ground at which near-surface measurements are made also varies. The standard is for winds to be measured at 10-m AGL and temperature and humidity to be measured at 2-m AGL; however, some observation networks do not adhere to this. The spatial distribution of observations varies considerably, on the scales of countries, and on smaller scales depending on population density. Contributing to the spatial-density variation is the fact that numerous special-purpose mesoscale networks exist, for example those that are established to meet air-quality and highway-maintenance needs. Buoy data are another type of near-surface measurement.
- *Commercial aircraft* – Onboard sensors measure wind speed and direction, temperature, pressure, and humidity. Some also measure turbulence intensity. Sloping profiles are provided at takeoff and landing, and a near-horizontal series of observations is available at cruising altitudes. Instrumented commuter aircraft, with shorter flight segments, generate a large number of vertical profiles in the lower troposphere. The reporting frequency varies, but observations are available at an interval of 60 seconds or less during ascent and descent, and approximately every 3 minutes at cruising altitude. Other aircraft sensor packages produce observations at specified pressure and horizontal-distance intervals. See Moninger *et al.* (2003) for additional information.
- *Doppler radar* – Measures the reflectivity from hydrometeors, and the radial wind speed relative to the radar. It scans a three-dimensional volume, and in modeling applications is used primarily for initialization of convective-scale models.
- *Doppler lidar* – Measures the radial wind speed relative to the lidar. It scans a three-dimensional volume on the convective scale, and is used primarily for initialization of mesogamma-scale and smaller-scale models of the boundary layer.
- *Wind profiler* – Upward-pointing radar that measures the horizontal wind vector in a column, with an hourly frequency. Often collocated with the wind profilers are Radio Acoustic Sounding Systems (RASS) for measuring temperature profiles.

Satellite-based measurement platforms include the following. There are many others that are described in the literature.

- QuikSCAT SeaWinds sea-surface winds from NASA are disseminated by the NOAA National Environmental Satellite, Data, and Information Service (NESDIS). The SeaWinds instrument on QuikSCAT is an active microwave radar that measures the backscatter from ocean-surface waves, and winds can be obtained in all conditions except for moderate to heavy rain. A function is used to relate the measured backscatter to the 10-m Above Sea Level (ASL) neutral-stability-equivalent winds. The QuikSCAT Level 3 gridded ocean wind vectors are available on an approximate 0.25° × 0.25° global grid with separate maps for the ascending and descending passes. The data are available for the period 2000 to the present. See Bourassa *et al.* (2003) and Hoffman and Leidner (2005) for additional information about data properties.
- Radio-occultation soundings of temperature and water vapor are obtained by using satellite-borne receivers to measure the phase delay of radio waves emitted from Global Positioning System (GPS) satellites, as the waves are occulted by Earth's atmosphere. These soundings are available globally, and can provide data where there are voids in other observation networks. Additional information can be found in Anthes *et al.* (2008).
- The Tropical Rainfall Measurement Mission (TRMM) product (Huffman *et al.* 2007) combines precipitation estimates from multiple satellites (retrievals from measurements in the microwave and infrared regions of the spectrum) as well as gauge-based analyses on a 0.25° × 0.25° grid that extends from 50° N to 50° S for the period from 1998 to the present. Latent-heating rates inferred from these rainfall analyses are used during the model-initialization process.
- The NASA Earth Observing System Terra and Aqua platforms have a MODerate-resolution Imaging Spectroradiometer (MODIS) sensor with visible, near-infrared, and infrared bands. The MODIS provides information on a suite of meteorological variables, including temperature and moisture. See Seemann *et al.* (2003) for an example of the retrieval of temperature and moisture.
- Special Sensor Microwave Imager (SSM/I) and Total Ozone Mapping Spectrometer (TOMS) provide data that have been used widely for model initialization. See Okamoto and Derber (2006), Goerss (2009), and Monobianco *et al.* (1994) for examples of the assimilation of SSM/I data.
- The Geostationary Operational Environmental Satellite (GOES) allows the calculation of hourly feature-track winds derived from infrared, visible, and water-vapor imagery (Gray *et al.* 1996, Nieman *et al.* 1997, Le Marshall *et al.* 1997). Also, estimates of precipitation based on the GOES Precipitation Index (GPI) can be used in diabatic initializations.
- The infrared Spinning Enhanced Visible and InfraRed Imager (SEVIRI) sensors on the geostationary Meteosat Second Generation satellites provide information about temperature and humidity (Di Giuseppe *et al.* 2009).

6.2.2 Observation-quality, -frequency, and -density variability

There is a great deal of space and time variation in the availability of observations used to initialize models. For example, *in-situ* measurements of the model-dependent variables are reported at time intervals (frequencies) that vary greatly, depending on the observation network. Figure 6.1 illustrates the spatial-density variability of near-surface observations, using North Africa as an example. The smaller number of population centers in the arid region is responsible for the paucity of observations there. In addition to the fact that the standard data-reporting interval varies among different observation networks, missing data are a frequent problem in some areas of the world. Such gaps in the data record can occur because meteorological observing stations are staffed only during the day, meteorological- and communications-equipment malfunctions occur, political unrests cause observing stations to be shut down for long periods, and late observations or communications-network delays cause data to arrive too late to be used to initialize a forecast. As an example of one of the more-continuous data records for West Africa, Fig. 6.2 shows the observed relative humidity for a surface station in Benin for a winter season. Clearly there are significant data gaps. Lastly, the suitability of instrument locations can vary considerably. Even though there are standards for the land-surface properties that should exist at an observation site, and for minimum distances between the observation site and physical obstructions, there is nevertheless considerable variability in the quality of the instrument

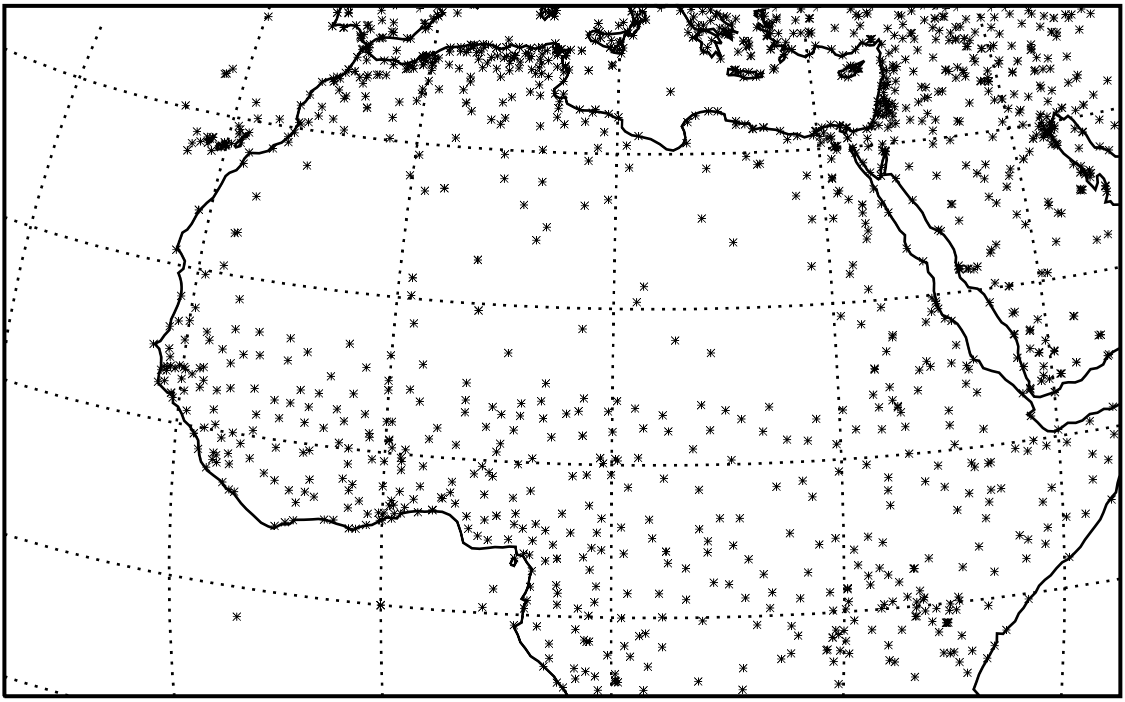


Fig. 6.1 The spatial distribution of near-surface, relative-humidity measurements in northern Africa.

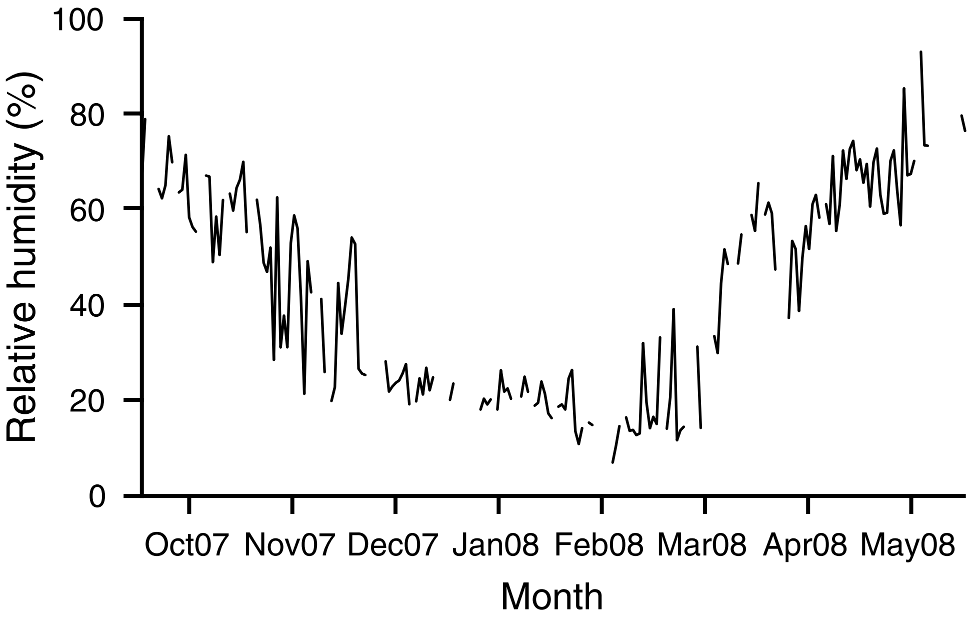


Fig. 6.2 A meteogram of the near-surface relative humidity at an observation location in Benin, illustrating the existence of gaps in the data reported during a winter season.

locations. For example, some measurement sites in cities are located on rooftops, where the thermal properties of the surface can be extreme, and where the winds are distorted by the structure. The next section describes specific approaches for ensuring the quality of observations used in a model.

6.2.3 Observation quality assurance and quality control

The term *Quality Assurance* (QA), in the context of meteorological observations, refers to the overall protocol that is employed to ensure the availability of quality observational data for use in NWP models and other applications. It is, in fact, a formal plan for accomplishing this goal, and typically might include the specifications for the instruments to be deployed, the instrument-siting requirements, the schedule for instrument calibration, the schedule for field inspection of instruments, and the routine numerical checks to be applied to the data (the *Quality-Control* (QC) process). See Shafer *et al.* (2000) for discussion of a complete QA procedure for a mesoscale network of sensors.

Historical (e.g., more than a day old) meteorological observations are publicly available from a variety of sources, some of which are mentioned in Section 10.10. In contrast, current observations for some nations are only available for a fee, which makes it difficult to establish real-time modeling systems for use in research or operations. Whether from historical observational-data archives, or from real-time observation networks, most observations have already undergone some quality checking. But, it is still important for those using the observations in a model initialization to perform checks of their own. A single grossly incorrect observation that erroneously passes the QC tests can have its negative influence spread over a large area in a model, and potentially damage the entire model solution.

There are a variety of causes for observations to be incorrect in some respect. A measurement may be of good quality, but the time, date, or geographic-location identifier can be wrong, resulting in the observation being applied at the wrong place or time. Or, the electronic transmission of the observation may have compromised it in a major or subtle way. The observation itself can have systematic and random errors. The systematic error is often related to incorrect instrument calibration. Another type of observation problem is referred to as representativeness error, which is discussed in more detail in Section 9.5.2. It results from the fact that an observation typically represents conditions at a point in space, and sometimes an average in time (e.g., winds), while the variables defined in model initial conditions represent grid-box-area averages, and apply at a specific time. Thus, the use of an observation to define conditions on a model grid can cause very-local properties to be spread over an erroneously large area. For example, if the observed wind and temperature associated with a convective outflow boundary are interpolated to a synoptic-scale-model's grid, and influence five to ten model grid boxes, an area of hundreds of square kilometers would be impacted by the small mesogamma-scale event.

Some simple, commonly employed QC tests include the following. More discussion about such checks can be found in Liljegren *et al.* (2009).

- *Limit tests* – Observations are compared with physical limits, sensor limits, and climatological limits. A physical limit, or constraint, for relative humidity would be that it cannot be less than 0% or much greater than 100%. And, wind speed cannot be less than zero. Similar absolute limits on the value of an observation can be defined in terms of the physical limits of a sensor or in terms of climatology (e.g., the minimum temperature ever observed at a station).
- *Temporal-consistency checks* – Successive observations of a variable define a rate of change, and this is compared with likely values. Because of rapid changes that can occur during convective weather, this check can be turned off when precipitation is occurring.
- *Spatial-consistency checks* – This is sometimes referred to as a buddy check, because observations are compared with horizontally or vertically adjacent data points. Or, an observation can be compared with an average calculated using a number of nearby observations. The resulting difference is compared with the historical maximum difference observed at that point, based on archived observations.

As will be seen later in this chapter, many modern data-assimilation systems merge observations with the most-recent gridded forecast that is valid at the observation time. Specifically, the gridded forecast is adjusted based on differences from the observations, and the result is used to initialize the next forecast cycle. But frequent large differences between the short forecasts and the observations at a particular location often result more from errors in the observations than from forecast errors. In effect, the volume of atmosphere that was initialized with accurate measurements is advected, during the short forecast, over the locations of new observations, and the statistical difference over a long period is used to judge observation quality. For example, Hollingsworth *et al.* (1986) describe how the operational ECMWF data-assimilation system can be used to monitor observation quality. This automated and economical approach to the QC process allows suspect instruments to be identified and corrective action taken, without routinely visiting and inspecting every instrument.

6.2.4 Other observation processing

Whether winds are observed and reported in terms of the individual components or as speed and direction, the measurements may need to be converted to the model wind components. This is because the model u that is defined to be parallel to the grid-point rows, and the model v that is parallel to the grid-point columns, generally differ from the geocentric u and v that are defined relative to latitude and longitude lines. For every vertical column of grid points (the same i, j coordinate), the mathematical transformation will be slightly different. This necessity may be most easy to accidentally overlook when the model coordinates are Cartesian, and the grid-point rows and columns are approximately oriented east–west and north–south.

Software that interpolates (analyzes) observations to a model grid operates in the framework of the model's horizontal coordinate system. Thus, because observation locations are typically defined in terms of latitude and longitude coordinates, there needs to be a transformation to the horizontal coordinates of the model, if it is x–y and not latitude–longitude based.

Lastly, the units of the observations may need to be transformed to those employed by the model. For example, wind speeds are often reported in knots, but models generally use the meter–kilogram–second (mks) system. And it is common for humidity observations to require conversion as well.

6.2.5 Metadata

Metadata (also called meta-knowledge) accompany the observations themselves, and provide information necessary for their use. Essential types of metadata include the file structure, data format (e.g., NetCDF), the variable (e.g., wind speed), the units (e.g., mks), and the time and three-dimensional-spatial coordinates of the observation. Optional, but useful, information includes the instrument type, the date of the most-recent calibration, and a photo of the instrument site and surroundings. The concept of metadata also applies to model-generated data as well, although the relevant information will obviously be different.

Conventions have been established for the format of metadata. For example, the NetCDF (Network Common Data Format) Climate and Forecast (CF) Metadata Convention is a well-documented standard for observational and forecast metadata, which is designed to promote the processing and sharing of files created with the NetCDF Application Programmer Interface [NetCDF API]. The CF conventions generalize and extend the convention of the Cooperative Ocean/Atmosphere Research Data Service, a NOAA/university cooperative group whose goal is the sharing and distribution of global atmospheric and oceanographic research data sets.

6.2.6 Targeted or adaptive observations

Economic and other constraints limit the number of observations that are made of the atmosphere, and thus it is reasonable to want to obtain observations from locations where they will have the largest positive impact on model-forecast accuracy, for a particular prevailing weather situation. Methods have been developed to satisfy this need, where the measurements are referred to as adaptive or targeted observations. However, it is clearly not economically feasible to deploy mobile observation platforms on a day-to-day basis. But, there are high-impact weather events, such as hurricanes or severe extratropical cyclones, for which special aircraft observations are made. If the aircraft can be routed so as to provide observations from locations for which the forecast skill is very sensitive to the accuracy of the initial conditions, the procedure can save lives. The routine use of targeted aircraft observations may become more common with the continued development of unmanned aerial vehicles.

Various strategies for observation targeting have been evaluated as part of the following field programs.

- Fronts and Atlantic Storm Tracks EXperiment (FASTEX; Emanuel and Langland 1998; Bergot 1999, 2001; Bishop and Toth 1999; Joly *et al.* 1999; and Bergot and Doerenbecher 2002)
- NORth Pacific EXperiment (NORPEX, Langland *et al.* 1999, Majumdar *et al.* 2002a)
- Atlantic THORPEX (The Hemispheric Observing-system Research and Predictability EXperiment) Observing System Test (Langland 2005)
- Annual US NWS Winter Storm Reconnaissance (WSR) programs (Szunyogh *et al.* 2000, 2002; Majumdar *et al.* 2002b)

The following notational framework for viewing the adaptive-observation problem is provided by Berliner *et al.* (1999), Majumdar *et al.* (2006), and others. Let $\mathbf{X}_i$, $\mathbf{X}_a$, and $\mathbf{X}_v$ represent n-dimensional vectors that define the state of the atmosphere at times t_i, t_a, and t_v, respectively, in terms of the grid-point values of variables or spectral coefficients. The initial time, t_i, is when the decision must be made, based on $\mathbf{X}_i$ information, about the types and locations of special observations to be collected at time t_a (the targeted observation time, and the analysis (initial) time of the operational forecast), where the objective is to optimize the statistical properties of a forecast $\mathbf{X}_v$ at the verification time t_v. Within the interval $t_a - t_i$, the observing platforms need to travel to the target locations so that observations can be made at t_a for use in initializing the forecast. The time interval $t_a - t_i$ is chosen based on logistical considerations associated with planning the surveillance mission, launching the aircraft, and getting the aircraft to the necessary locations to make the observations. The data set $\mathbf{X}_a$ is the result of assimilating standard observations and the special targeted observations, and is used as the initial conditions for the forecast.

A practical example of the above process is as follows. Assume that a 72-h forecast from a standard operational model run predicts very-heavy, flood-producing rainfall over New York City, associated with a coastal cyclone. At t_i it is decided to deploy dropsondes at t_a, 24 h in the future, at locations where they will have the greatest impact on the 48-h precipitation forecast over New York City. The best location for making the measurements (the target area) will depend on the variable whose forecast must be improved (rainfall) and the verification region (New York City). Most adaptive-observation strategies allow the association of the observation target area with a specific verification region in the model. An exception is the ensemble-spread method discussed below.

There are a number of approaches for defining the locations and types of observations that will have the greatest positive impact on the quality of forecasts. A few of these are summarized below. Discussions of other methods can be found in Palmer *et al.* (1998), Bishop *et al.* (2001), Aberson (2003), and other references cited in this section. Berliner *et al.* (1999) focused on a statistical framework for the adaptive-observation problem.

- *Ensemble variance/spread* – This is a simple approach, described by Aberson (2003), that can improve tropical-cyclone-track forecasts by locating supplemental observations in areas where the variance is largest among members of an ensemble prediction that is valid at the analysis time. In regions of large ensemble variance, it is assumed that there is also a large uncertainty in the wind analysis, implying the need for additional observations. Unfortunately, there is no way to propagate the uncertainty or error at the analysis time, t_a, into another region at the forecast verification time, t_v. Nevertheless, Aberson (2003) showed that observations made in areas with large ensemble variance improved tropical-cyclone-track forecasts more than did uniformly distributed observations.
- *Adjoint methods* – In Chapter 3 it was noted that the adjoint operator, which is based on a linear version of a nonlinear forecast model, produces sensitivity fields that indicate the quantitative impact on a particular aspect of the forecast of any small, but arbitrary, perturbation in initial conditions, boundary conditions, or model parameters. Thus, given a specific characteristic of the forecast for which the sensitivity will be calculated, for example the minimum pressure in a cyclonic storm, the area in the initial conditions to which the characteristic is most sensitive can be defined. Thus, the implication is that this region should be better measured. This kind of analysis is also discussed in Chapter 10 in relation to the design of sensitivity studies. Palmer *et al.* (1998), Pu *et al.* (1998), Bergot (1999), Buizza and Montani (1999), and Bergot and Doerenbecher (2002) describe the use of the adjoint method for targeting observations. Using the terminology described above, at t_i the forward linear version of the nonlinear forecast model is integrated from t_i to t_v. Then, the adjoint of the linear model is used to define the sensitivity of conditions at t_a to the forecast error at t_v. This sensitivity information defines the target area for observation platforms, which are deployed with sufficient time to reach the defined area and make the measurements at t_a. Shortly after time t_a, the operational model is initialized with the available data, and the forecast is performed. Figure 6.3 illustrates the process by which this method is applied. An issue with this approach is that a verification region must be defined, where error growth is to be

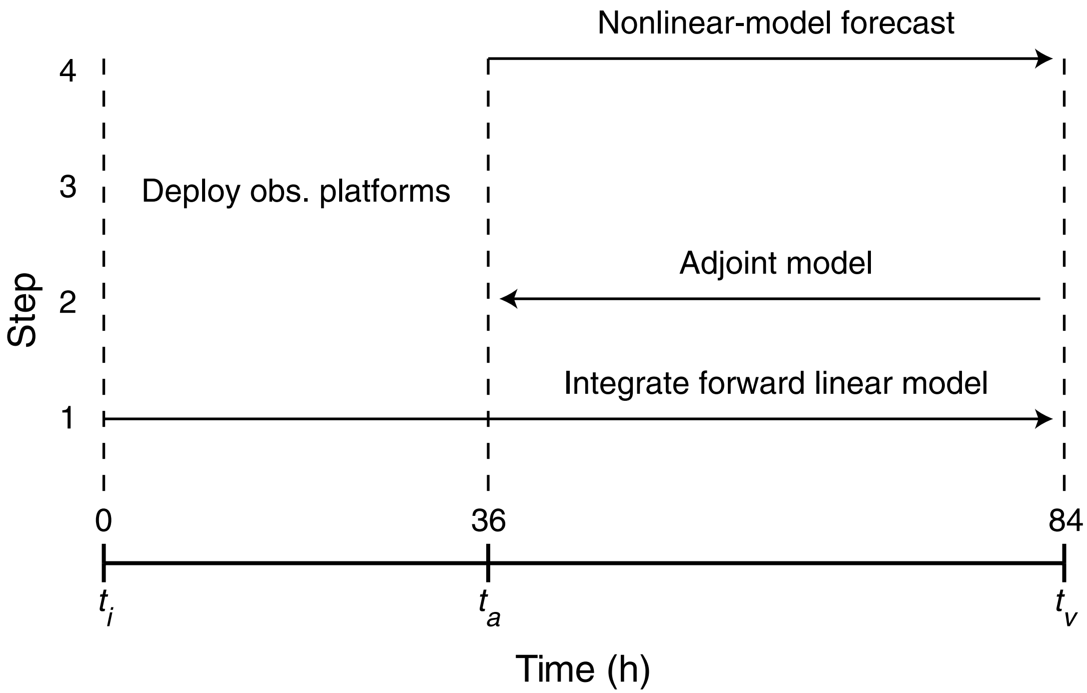


Fig. 6.3 Schematic of the process by which the adjoint method is used to define the region where special observations may be targeted to improve the skill of an operational forecast. At time t_i the forward linear model and the adjoint model are run (steps 1 and 2) to identify target areas for special observations, and the time when aircraft need to be deployed to a target area. The time t_a is when the special observations apply, and when the forecast is initiated. The numbers along the dashed line to the left define steps in the process.

minimized, and this is problematic for phenomena spanning a large area or for situations where there are multiple regions of interest.

- *Ensemble transform Kalman filter* – The ensemble transform technique (Bishop and Toth 1999; Szunyogh *et al.* 1999, 2000) and the subsequent ensemble transform Kalman filter (ETKF, Bishop *et al.* 2001, Majumdar *et al.* 2002a,b) employ information from ensemble forecasting systems to identify regions where sampling would lead to forecast improvements. Advantages of the ETKF technique compared with the adjoint method include the lack of a requirement for an adjoint of the model, its low computational cost, the fact that it is based on nonlinear (ensemble) forecasts, and the fact that it provides quantitative estimates of the reduction in forecast error (not simply sensitivity metrics).

Because all of the above methods employ a model, the obtained target locations for observations will depend on both the method and the model. Different models can produce quite different estimates of locations.

A commonly noted practical limitation of adaptive-observation methods is that aircraft observations, whether they are made from the aircraft itself or with dropsondes, can only measure a relatively small volume of atmosphere. Thus, even if the targeting region is calculated accurately, it is often not logistically possible to measure a sufficiently large area to adjust the position or amplitude of large-scale features such as fronts or baroclinic waves in the model initial conditions. This is especially problematic in a region that is otherwise a data void. A related issue is that data-assimilation systems are sometimes more appropriate for observations made over larger areas than are observable with a modest number of observing platforms. Therefore, the impact of targeted observations on forecast skill can depend on the data-assimilation scheme. For example, Bergot (2001) shows that targeted observations from 20 FASTEX cyclogenesis cases have a greater positive influence on forecast skill when used in a four-dimensional rather than a three-dimensional variational assimilation system.

Figure 6.4 provides an example of the impact of targeted observations on forecast skill. It is a scatter plot of the RMS 500 and 1000 hPa height errors for 30, 36, 42, and 48 h forecast lead times from the ECMWF global model for five FASTEX case studies. The model was run with and without the use of the targeted observations. Each point represents an average error for the verification region, and corresponds to a particular verification height, FASTEX case, and verification time. With a few exceptions, the errors were less when dropsonde data were used. Of course adding observations anywhere in the model domain might be expected to reduce forecast errors, so interpretation of these results in the context of the effectiveness of the targeting method needs to be done cautiously. See Montani *et al.* (1999) for information about the targeting method used.

6.2.7 Optimal siting of permanently located observations

In contrast to the targeted observations just described, conventional, permanent observation platforms are distributed geographically to allow convenient access for their maintenance. However, there are approaches that could be used to locate such fixed platforms so as to improve model initial conditions and therefore predictive skill. For example, if a LAM is being run primarily to forecast a specific type of severe-weather event in a particular area, such as wind shear in the vicinity of an airport, one of the above-described observation-targeting methods could be applied for a large number of historical cases, and the results used to define the best overall permanent locations for observations. Another approach that has been evaluated is called a field-coherence technique (Stauffer *et al*. 2000, Tanrikulu *et al.* 2000), which is based on a statistical analysis of the model-simulated atmospheric structure. The spatial and temporal coherence, as defined here, is a measure of the distance scale over which there is temporal consistency in the spatial structure within a variable field. Thus, the coherence indicates how well a measurement made at one location is able to serve as an estimate of the value of that field at another location at a given analysis time. The concept is that the larger the field coherence in a geographic area, the fewer measurement sites are needed to adequately resolve the dominant features of that field. Observing-system simulation experiments, discussed in Section 10.2, can also be employed to evaluate the relative benefits of different spatial distributions of observations.

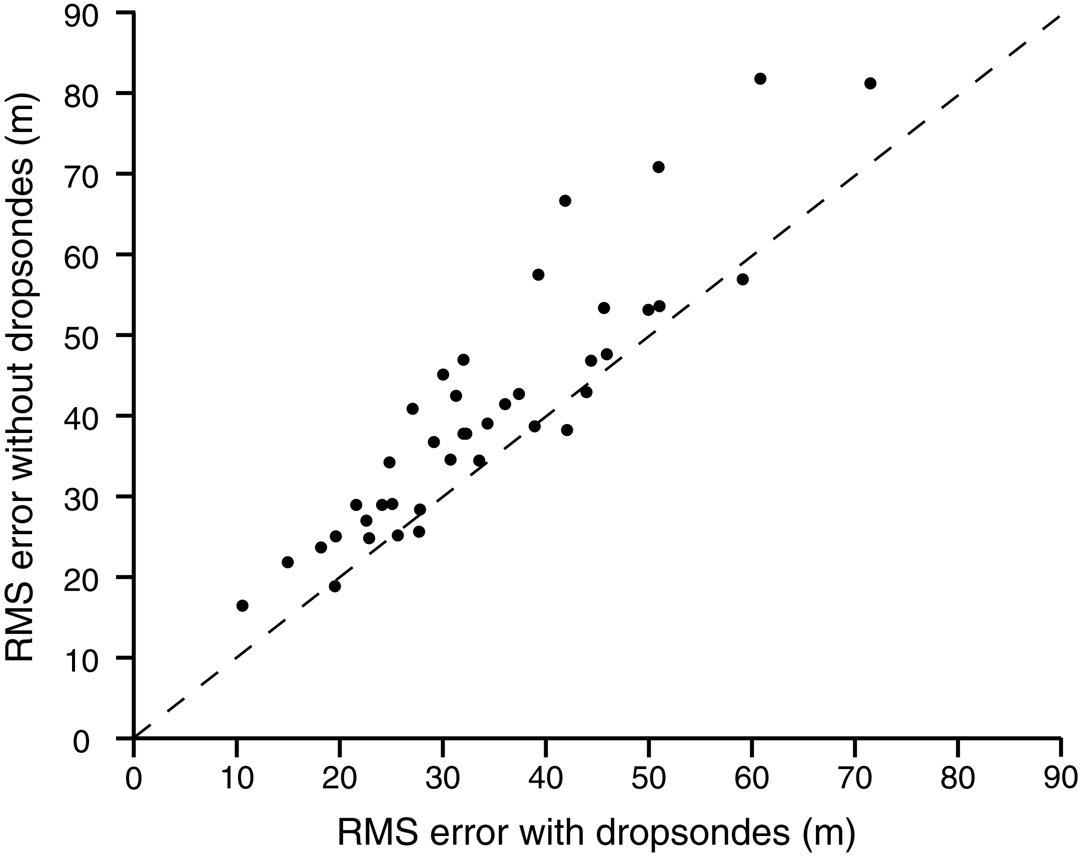


Fig. 6.4 Scatter plot of ECMWF global-model forecast errors of 500 hPa and 1000 hPa height, with and without the use of targeted observations. Each dot corresponds to RMS forecast errors in the verification region for one of five FASTEX cases, for four forecast lead times, and for two verification heights. From Montani *et al.* (1999).

6.3 Continuous versus intermittent data-assimilation methods

The processes of data assimilation and data analysis both have the objective of constructing a gridded data set that defines the state of a meteorological variable, and the terms are sometimes used interchangeably. That said, the use of the expression “data-assimilation system” generally means that a meteorological model is employed in the process. The overall purpose behind the use of the data-assimilation system can be the production of initial conditions for operational forecasts, or the construction of long-term reanalyses of the state of the atmosphere (see Chapter 16). The related expression “data-assimilation cycle” often encompasses the entire process of data quality control, the objective analysis, the initialization of the model (possible balancing), and the production of a short forecast to produce the next background field (Daley 1991). This section illustrates two general categories of data-assimilation systems, both of which involve the use of a model.

The qualities that are desirable in a computer-based objective-analysis process are well known by anyone who has constructed a manual, or subjective, analysis of observations. The following are traditional methods that have been used for decades in the manual analysis of observations.

- A first guess of the overall weather pattern is important. It provides the analyst with context for the observations, and can be based on the map constructed at the previous analysis time, a recent forecast, or personal knowledge of the typical regional weather patterns (the climatology).
- The variables should not be analyzed independently. For example, on large scales, areas with strong gradients in the height analysis are used to infer regions of high wind speeds when drawing isotachs.
- The overall weather patterns provide information that can be used in the interpolation between observation points. For example, when analyzing a jet maximum, isotachs are streaked out in the direction of the wind. And, at the analyzed position of fronts, isopleths of all variables reflect the transition in air-mass properties.
- The spatial density of observations is used in the analysis process. In areas where the observations are dense, the analysis is faithfully drawn to them, whereas in areas where the observations are sparse or nonexistent the analysis is based on knowledge of the background (climatology or the prior analysis). Also, an observation in a cluster of observations that is inconsistent with the rest is ignored, or given less weight, in the analysis.
- The smoothness of the analysis is made to be consistent with the density of the data and the known scales of the phenomena being analyzed.

6.3.1 Intermittent, or sequential, assimilation

Most operational data-assimilation systems use the intermittent, or sequential, approach. The general process is shown in Fig. 6.5. The cycle begins with an initial forecast. The next forecast in the cycle is initialized using a merger of observations (upper left) and a first-guess field (upper right). The latter is typically the output from the most-recent forecast, which is valid at the initial time of the current forecast. Observations that are made within a specific time window (±*n* minutes in the figure) that spans the initialization time are aggregated and used in the analysis. The prior forecast in this process is called the first guess, or the prior estimate, or the background field. This use of the forecast in the analysis process allows the model solution to better fill spatial observation gaps than would interpolation over large distances between observations. In addition, the model solution can develop circulations in response to local surface forcing, and this allows those signatures to be included in the initial conditions (see Section 6.4). Because the merger of the forecast and the observations involves the use of information from different times, the process is referred to as Four-Dimensional Data Assimilation (FDDA). The initial conditions are then used to initialize the forecast, which will need LBCs if the model is a LAM. For global models, new forecasts are typically initiated every 6 hours, whereas for regional models this can occur as frequently as hourly. In either case, model fields are extracted at

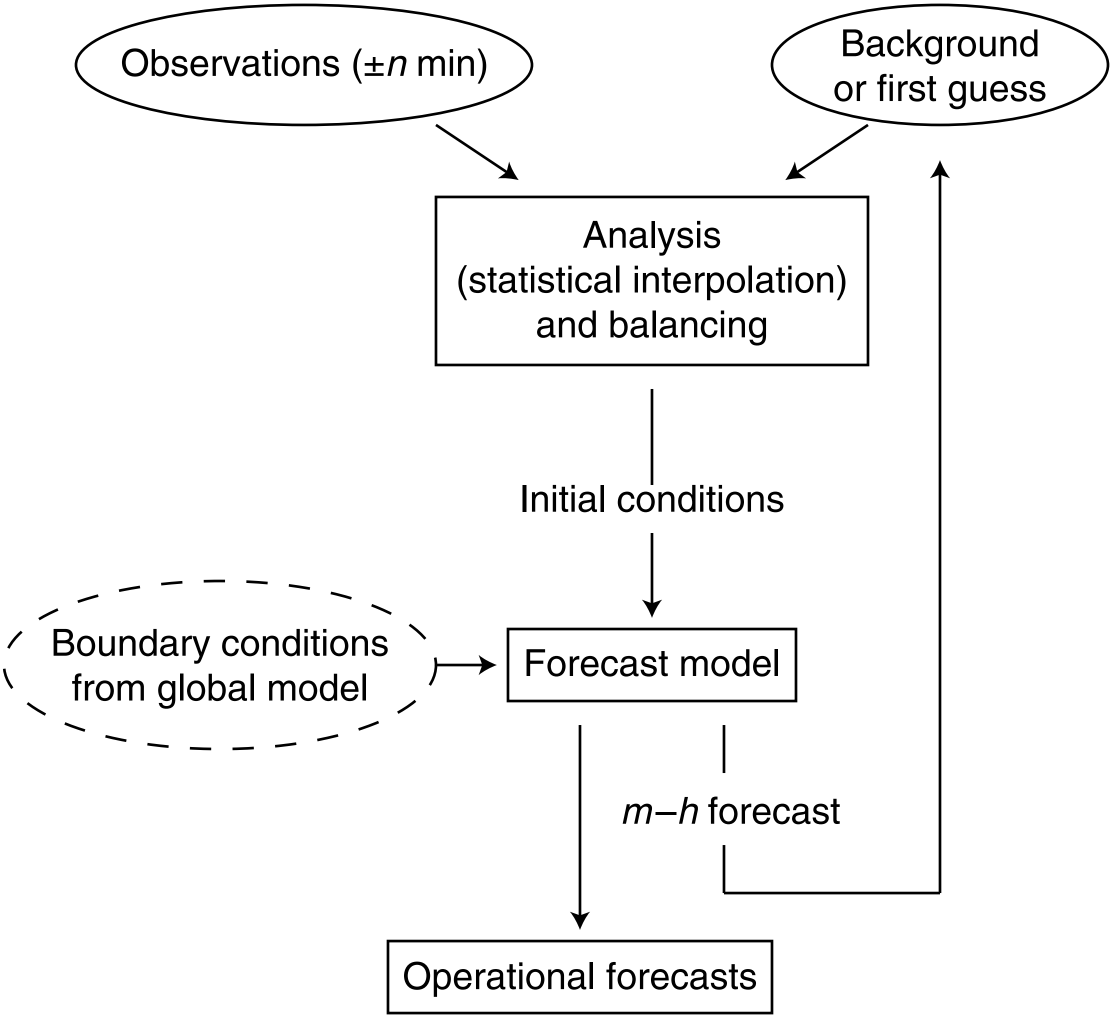

Observations (±*n* min)
Background or first guess
Analysis (statistical interpolation) and balancing
Initial conditions
Boundary conditions from global model
Forecast model
m–*h* forecast
Operational forecasts

Fig. 6.5 Schematic showing the process of intermittent, or sequential, data assimilation. See the text for discussion.

forecast time *m* to be used as the first guess for the next forecast cycle. This sequential approach to data assimilation serves as the basis for the optimal-interpolation, three-dimensional variational, ensemble Kalman filter, and other methods described later in this chapter. These represent particular approaches for accomplishing the process in the upper rectangular box of Fig. 6.5. See Fig. 6.6b for a different graphical depiction of this sequential-assimilation method.

6.3.2 Continuous assimilation

These continuous approaches involve the assimilation of observations at the times that they are made, rather than in batches, as with the sequential methods. Four-dimensional variational assimilation is a continuous-assimilation method, and is described later after background material on statistically optimal methods is presented. The other major continuous-assimilation method, Newtonian relaxation, is summarized in this section. Data assimilation by Newtonian relaxation (or nudging) is accomplished by adding nonphysical nudging terms to the model predictive equations. These terms force the model solution at each grid point to observations (observation, or station, nudging), or analyses of observations (analysis nudging), in proportion to the difference between the model solution and the observation or analysis. The following equation illustrates the form of the relaxation term in a prognostic equation, where f is any dependent variable, F represents all the physical-process terms, f_{obs} is the observed value of f interpolated to the grid point, and τ is a relaxation time scale. This relaxation-term weight can be separated into three components: the factor that determines the magnitude of the term relative to the physical terms in the equation (G), the function that defines the spatial and temporal influence of observations (W), and the observation-quality factor (ε). In finite-difference space, this equation applies at a particular grid point and at a particular time step:

$$\frac{\partial f}{\partial t} = F(f, \boldsymbol{x}, t) + \frac{f_{obs} - f}{\tau(f, \boldsymbol{x}, t)} = F(f, \boldsymbol{x}, t) + G(f) W(\boldsymbol{x}, t) \varepsilon(f, \boldsymbol{x})(f_{obs} - f) .$$

If the relaxation time scale is too small, the model solution will converge to the observation too quickly, and the other variables will not have sufficient time to dynamically adjust. If the time scale is too large, errors in the model solution will not be corrected by the observations.

This approach has several advantages. It is efficient computationally, it is robust, it allows the model to ingest data continuously rather than intermittently, the full model dynamics are part of the assimilation system so that analyses contain all locally forced mesoscale features, and it does not unduly complicate the structure of the model code. Studies using Newtonian relaxation include Stauffer and Seaman (1990, 1994), Stauffer *et al.* (1991), Fast (1995), Seaman *et al.* (1995), and Liu *et al.* (2006, 2008a). A finding of these studies is that analysis nudging may work better than intermittent assimilation on synoptic scales. Furthermore, Stauffer and Seaman (1994) and Seaman *et al.* (1995) showed that nudging toward observations was more successful on the mesoscale than nudging toward analyses. Leslie *et al.* (1998) found that the impact of observation-nudging was similar to that of assimilating the same data in a four-dimensional variational system (Section 6.11.1), with the former being practicable while the latter was too computationally expensive. Bao and Errico (1997) applied the adjoint method to illustrate the impact of the nudging terms and some limitations of the method.

Figure 6.6 is a schematic that compares the intermittent and the continuous assimilation processes. In both cases, the time increases from left to right. For the continuous assimilation (Fig. 6.6a), observations are ingested into the model at every time step, and forecasts are launched at whatever frequency is desired (6 h in this example). The intermittent-assimilation process (Fig. 6.6b) uses the same observations, except that they are aggregated temporally over some time interval to produce an objective analysis (Anal) that is combined with a first-guess field from a short forecast. The resulting gridded fields may undergo balancing (Initialization – Init), and are then used for the initial conditions of a

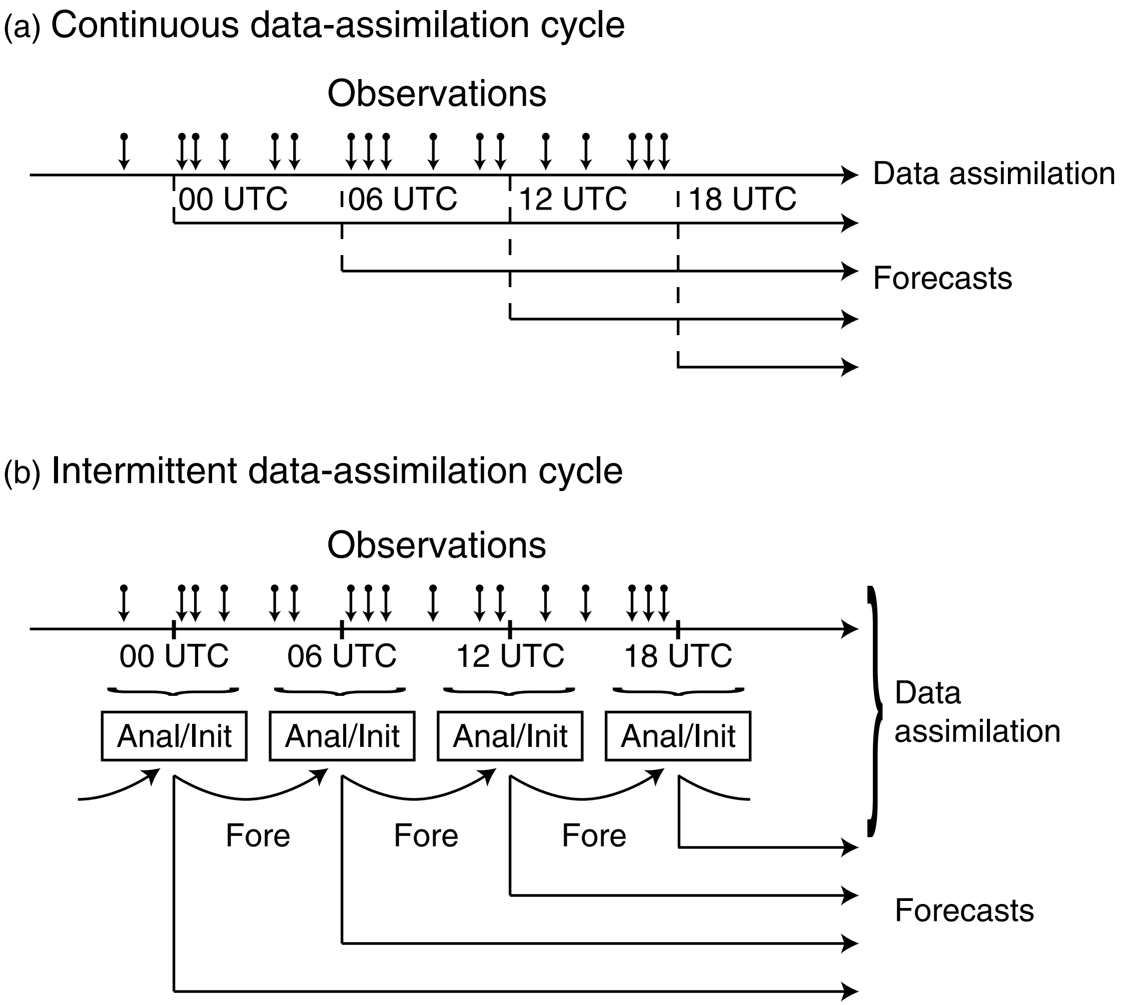


Fig. 6.6 Schematic showing the components of data-assimilation cycles for the intermittent and continuous methods. See the text for details.

forecast. Figure 6.6b shows the same cycling process as does Fig. 6.5, but emphasizes the distinctions with the continuous-assimilation method.

A negative aspect of this type of continuous data assimilation is encountered when relaxing a mesoscale-model solution toward a synoptic-scale analysis of observations. Specifically, the model will develop fine-scale atmospheric features in response to differential surface forcing, but relaxation terms will damp these features if they are not properly represented in the analysis. Consider a situation where the model develops a sea-breeze circulation and a coastal front in response to differential surface thermal forcing at a coastline. Given the typical density of two-dimensional and three-dimensional observations, and the resulting lack of spatial detail in an objective analysis based on these observations, an analysis will only represent large-scale features and not the mesoscale detail. Thus, relaxing the model solution toward this analysis will damage the solution. This issue also exists when relaxing the model solution towards observations (rather than gridded analyses of observations) because isotropic functions for spatially spreading the influence of the observations do not respect linear mesoscale features and will also damage fine-scale features in the model solution. To avoid this problem, a method called spectral nudging has been developed. Here, the evolving model solution is filtered so that only the larger-scale features are differenced with the analysis, to define the correction term. The concept of spectral nudging is also discussed in Chapter 16, relative to climate modeling, because it is sometimes used when regional models are employed to downscale from global-climate simulations. In particular, the solution of the regional climate model is spectrally filtered, and the large-scale fields in the regional model are nudged toward the global-model solution, thus avoiding a drift of the large-scale solution in the regional model.

6.3.3 Hybrid intermittent–continuous methods

Even the assimilation methods that are referred to as continuous are, strictly speaking, intermittent because the data are inserted at the time-step intervals. Thus, it is perhaps understandable that there is not an especially clear definition of the terminology. For example, as the time period between analyses decreases (Fig. 6.6b), it is easy to see that the intermittent method approaches the so-called continuous one. This is not a hypothetical point, because the cycle, or update, frequency is now hourly in some operational data-assimilation systems. In addition, other methods combine aspects of the continuous and intermittent approaches. For example, in the analysis-correction system described in Lorenc *et al.* (1991), batches of data within 6-h time intervals are analyzed at each model time step, and the results are inserted into the model solution at each time step, with greater weight given to the observations whose valid times are closer to the analysis time. And, Bloom *et al.* (1996) describe an incremental analysis updating method wherein an analysis based on statistical interpolation is conducted every 6 h, and the *analysis increments* (the difference between the analyzed value and the first guess) are used as a continuous forcing during a 6-h integration. Even though these two methods retain some aspects of the intermittent approach, the data impact the model simulation at every time step.

6.4 Model spinup

Now that a couple of different types of data-assimilation methods have been discussed, it is appropriate to introduce the concept of model *spinup*. Because of the typical lack of spatial density in the observing network, especially in terms of observation platforms that provide information in three space dimensions, observations cannot generally define sharp cross-frontal gradients, the correct wind-speed amplitude of upper-level or low-level jet maxima, the structure of thermally forced boundary-layer circulations, the waves or channeling associated with orography, and the small-scale vertical motions and humidity gradients associated with clouds and precipitation. And, because the observations are not adequate to define these features, a simple analysis of them is not going to suffice.

However, the model itself can provide information about the atmosphere, to supplement what is in the observations. For example, land-surface properties (e.g., terrain elevation, land–water boundaries) are known with a horizontal resolution that is orders of magnitude greater than the resolution of our information about the three-dimensional structure of the atmosphere. Thus, after the model integration is begun, the lower troposphere will respond to the dynamic and thermodynamic forcing from the landscape at the lower boundary, producing thermally and dynamically forced wind circulations, contrasts in the boundary-layer temperature and humidity fields at coastlines, etc. The model dynamics have added this structural information to what was defined in the initial conditions based on observations. In addition, during the early period of a model integration the deformation at fronts will increase poorly resolved gradients, nonlinear interactions among larger waves will generate finer scales in the spectrum, and ageostrophic circulations will strengthen, creating vertical motions that can produce the saturation necessary for the development of cloud and precipitation in the model. This post-initialization development of realistic three-dimensional features during the model integration is called spinup.

Even though the spinup process allows the generation in the model solution of features that are not observed, it is problematic because it occurs during the model forecast. Thus, the early period of the forecast – perhaps 12 h in duration – does not contain properly rendered, potentially important, atmospheric processes. For example, precipitation during the first half-day of a forecast may not be realistic. Thus, there has been a great emphasis on developing initialization procedures that produce model initial conditions that are spun up, or largely so. This has led to subjective terminology such as *cold starts* for initializations that contain no spun-up processes, *hot starts* for the use of initial conditions that are completely spun up, and *warm starts* for the use of partially spun-up initial conditions.

When reading about the various data-assimilation strategies that are described throughout the rest of this chapter, the reader should keep in mind the desirability of having reasonably well spun-up initial conditions. For example, in the context of the so-called intermittent (or sequential) and continuous assimilation methods described in the previous section, the sequential method could produce less-well-spun-up initial conditions if the influence of the observations is distributed in a way that smooths out the model-produced background field. The historical motivation for all dynamic-initialization methods that employ a model during a preforecast integration period has been the desire for spun-up initial conditions.

6.5 The statistical framework for data assimilation

6.5.1 Introduction, and illustration with scalar relationships

This section describes mathematical concepts that form the basis for many approaches to data assimilation. Data assimilation is an analysis method wherein information from observations is accumulated, over a period of time, into a model state. The observational information is carried forward in time by the model, which imposes dynamic consistency among the variables and spreads the information both spatially and among the variables. There are three components to the data-assimilation process: observations; background information about the state of the atmosphere, perhaps based on a previous analysis or a model forecast; and dynamic constraints, perhaps based on a model.

In the following discussion, the term “vector” will be used to refer to a group of elements that defines a state of the model atmosphere, either in the form of gridded values or spectral coefficients. For example, the vector $\boldsymbol{x}$ may be defined as $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)$. If this vector corresponds to the state of the atmosphere as defined in a grid-point model, the dimension n will be the number of grid points multiplied by the number of dependent variables.

In the above example, the column matrix that is a collection of numbers that defines the state of a model atmosphere is referred to as the *state vector*, $\boldsymbol{x}$. If this vector results from the use of an analysis system, it will disagree with observations because of errors in the analysis process, instrument error, and representativeness error that results from the finite spatial resolution of the analysis. The *true-state vector*, $\boldsymbol{x}_t$, represents the best-possible state that can be defined on the model grid. This is not the same as a perfect-state vector, which corresponds exactly to the atmospheric state, because of the unavoidable representativeness error. The gridded background field, which is the first-guess estimate of $\boldsymbol{x}_t$ before the analysis is conducted, is defined by the vector $\boldsymbol{x}_b$. Lastly, the analysis is represented as $\boldsymbol{x}_a$. The analysis problem is thus defined as finding a correction, $\delta\boldsymbol{x}$, such that

$$\boldsymbol{x}_a = \boldsymbol{x}_b + \delta\boldsymbol{x}$$

is as close as possible to $\boldsymbol{x}_t$.

The observations used in an analysis are collected into an *observation vector*, $\boldsymbol{y}$. In the analysis process, this observation vector needs to be compared with the state vector for the model-based first guess. Because each degree of freedom (the value of each variable defined at each grid point) in the state vector obviously does not have a corresponding observation (the observations being relatively few in number and irregularly located), for this comparison it is thus necessary to transform from model state space to observation space. This transformation is made by an *observation operator* (also called a *forward operator*) that is defined as $\boldsymbol{H}(\boldsymbol{x})$. In the simplest sense, it corresponds to interpolating state variables from grid points to observation points. It also can involve the transformation of a state variable to an observed variable. In the data-analysis process, differences between the observations and state vectors are calculated. The difference

$$\boldsymbol{y} - \boldsymbol{H}(\boldsymbol{x}_{\boldsymbol{b}})$$

is called the *innovation*, and the difference

$$\boldsymbol{y} - \boldsymbol{H}(\boldsymbol{x}_a)$$

is the *analysis residual*.

These concepts can be used in a simple illustration of least-squares estimation, which will lead to a general framework for data assimilation. Suppose we have two estimates, T_1 and T_2, of the true value of a scalar, say the temperature, T_t, at a point. In order to combine them optimally, we need statistical information about the errors, ε, of these estimates. Let

$$T_1 = T_t + \varepsilon_1 \text{, and} \tag{6.1}$$

$$T_2 = T_t + \varepsilon_2 , \tag{6.2}$$

where ε_i are unknowns. Let $E(X)$ be the expected value of measurement X, or the value that would be obtained by averaging many measurements. It is assumed that the instruments that measure T are unbiased. That is,

$$E(T_1 - T_t) = E(T_2 - T_t) = 0 \text{, or equivalently} \tag{6.3}$$

$$E(\varepsilon_1) = E(\varepsilon_2) = 0 . \tag{6.4}$$

And, we assume that we know the variances of the observational errors:

$$E(\varepsilon_1^2) = \sigma_1^2 \text{ and } E(\varepsilon_2^2) = \sigma_2^2 . \tag{6.5}$$

It is also assumed that the errors in the two observations are uncorrelated:

$$E(\varepsilon_1 \varepsilon_2) = 0 . \tag{6.6}$$

Equations 6.4–6.6 define the statistical information that we need about the two observations. Our objective is to linearly combine the two estimates of T in an optimal way, such that the result is the best least-squares estimate of T_t. Specifically, where T_a is the best (optimal) estimate of T_t, let

$$T_a = a_1 T_1 + a_2 T_2 \text{, and} \tag{6.7}$$

$$a_1 + a_2 = 1 . \tag{6.8}$$

The value of T_a will be the best estimate of T_t if the coefficients are chosen to minimize the mean-squared error of T_a. Specifically, using Eqs. 6.7 and 6.8,

$$\sigma_a^2 = E[(T_a - T_t)^2] = E[(a_1(T_1 - T_t) + a_2(T_2 - T_t))^2] . \tag{6.9}$$

Using the fact that $E(XY) = E(X)E(Y)$, for X and Y independent, and Eqs. 6.1, 6.2, 6.3, and 6.5, Eq. 6.9 becomes

$$\sigma_a^2 = a_1^2 \sigma_1^2 + a_2^2 \sigma_2^2 . \tag{6.10}$$

Given Eq. 6.8, and defining $a_2 = k$, Eq. 6.10 becomes

$$\sigma_a^2 = (1 - k)^2 \sigma_1^2 + k^2 \sigma_2^2 . \tag{6.11}$$

To find the value of k that corresponds to a minimum in the analysis variance, σ_a^2, differentiate Eq. 6.11 with respect to k, and set the expression to zero. This leads to

$$k = \frac{\sigma_1^2}{\sigma_1^2 + \sigma_2^2}. \tag{6.12}$$

Now, assume that our two sources of information, T_1 and T_2, are based on an observation and a background value. Eq. 6.7 becomes

$$T_a = kT_o + (1-k)T_b \tag{6.13}$$

and Eq. 6.12 becomes

$$k = \frac{\sigma_b^2}{\sigma_b^2 + \sigma_o^2}, \tag{6.14}$$

leading to

$$T_a = \frac{\sigma_b^2}{\sigma_b^2 + \sigma_o^2} T_o + \frac{\sigma_o^2}{\sigma_b^2 + \sigma_o^2} T_b. \tag{6.15}$$

For example, if the observation is very poorly known, σ_o is large and the analysis is weighted strongly toward T_b. Rearranging Eq. 6.13 yields

$$T_a = T_b + k(T_o - T_b). \tag{6.16}$$

Substitution of Eq. 6.14 into Eq. 6.11, and letting $\sigma_1 = \sigma_b$ and $\sigma_2 = \sigma_o$, yields

$$\sigma_a^2 = \frac{\sigma_b^2 \sigma_o^2}{\sigma_b^2 + \sigma_o^2} = \sigma_b^2(1-k). \tag{6.17}$$

This represents the uncertainty of the estimate σ_a in terms of the uncertainties of the observation and the background. Note that $\sigma_a^2 \le \sigma_o^2$ and $\sigma_a^2 \le \sigma_b^2$, meaning that the analysis variance is smaller than the variance of both sources of contributing information. Stated differently, using even a large-variance source of information will reduce the uncertainty in the analysis. Equation 6.17 can be rewritten as

$$\frac{1}{\sigma_a^2} = \frac{1}{\sigma_o^2} + \frac{1}{\sigma_b^2}. \tag{6.18}$$

The inverse of a variance is called the precision (the larger the variance, the lower the precision). Thus, the precision of the analysis is the sum of the precisions of the observation and the background.

Alternatively, instead of minimizing σ_a^2 in Eq. 6.11 to find an expression for the best estimate of T_a, a different approach can be used. For any T, the distance between T and T_b, and T and T_o can be measured by the following quadratic relationship:

$$J(T) = \frac{1}{2}(J_o(T) + J_b(T)) = \frac{1}{2}\left[\frac{(T-T_o)^2}{\sigma_o^2} + \frac{(T-T_b)^2}{\sigma_b^2}\right]. \qquad (6.19)$$

This function represents the square of the misfit of a variable (T) from each of the two sources of information, weighted by the precision of each of the estimators. It is often called a cost function or a penalty function. To define a value for T that corresponds to a minimum in the cost function, J is differentiated with respect to T, the resulting expression is set to zero, and it is confirmed that the extremum is, in fact, a minimum. The best estimate of T defined by this expression is T_a. The result is the same as defined in Eq. 6.15. Figure 6.7 illustrates graphically how the two penalty terms, J_o and J_b, in Eq. 6.19 are combined to produce the minimum in the analysis, T_a.

The above analysis involves the optimal combination of only two pieces of information, an observation and a background, or first-guess, value. That is, this has been posed as a simple scalar rather than a vector problem. It has also been assumed that these pieces of information are defined at the same location, and thus there has been no need for a forward operator to transform from model space to observation space. For application of these concepts in the framework of a model, the background state vector has a size in excess of 10^7 (for a grid-point model, the number of grid points times the number of dependent variables). And the observation vector has perhaps a size of 10^6. Fortunately, the above least-squares estimation methods have exactly the same form when applied to real multi-dimensional data-assimilation problems. The following summary of the most important points is based on Kalnay (2003).

- Equation 6.16 states that an analysis value is obtained by adding to the background (first guess), the innovation (the difference between the observation and first guess) multiplied by an optimal weight.
- The optimal weight, k, defined in Eq. 6.14, is the background error variance multiplied by the inverse of the total error variance (the sum of the background and observation error variances). The larger the background error variance, the larger is the correction to the background by the observation.
- Equation 6.18 states that the precision of the analysis is the sum of the precisions of the observation and the background.
- The rightmost part of Eq. 6.17 means that the error variance of the analysis is equal to the error variance of the background, reduced by a factor that is equal to one minus the optimal weight.

The application of the above least-squares methods to multi-dimensional and multi-variable problems is found in subsequent sections.

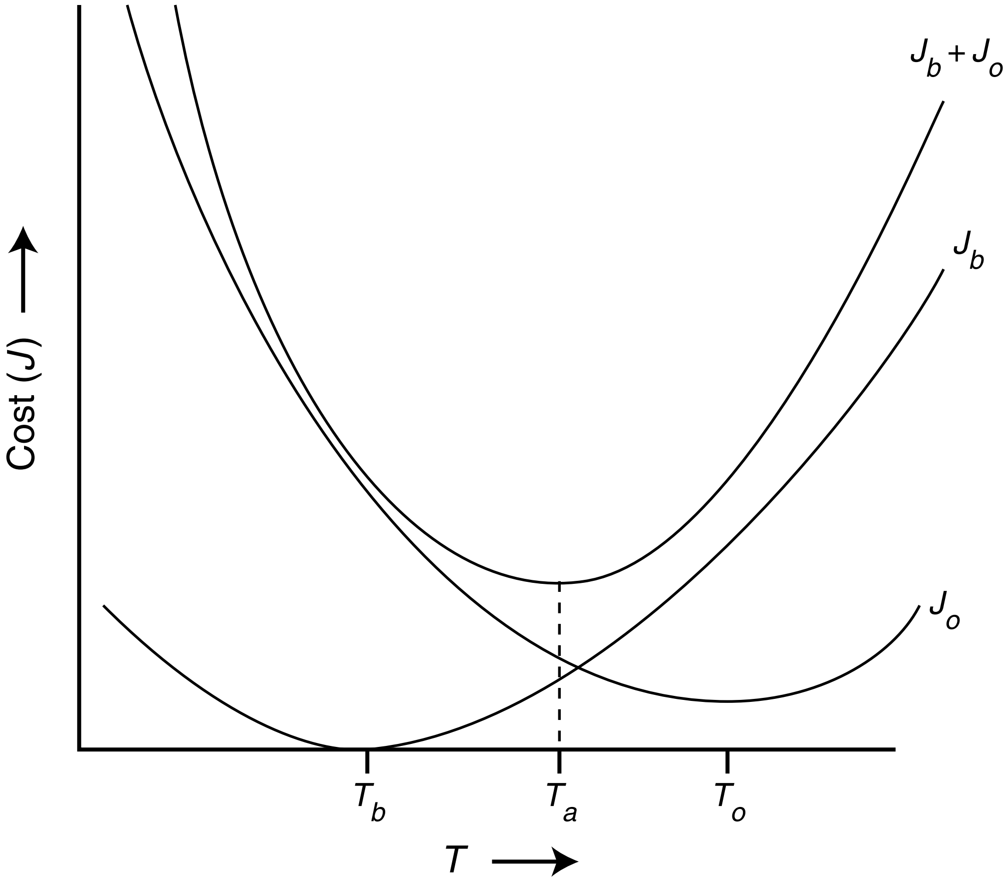


Fig. 6.7 Schematic showing how the two penalty terms, J_o and J_b, in Eq. 6.19, are combined to produce the minimum in the analysis error at T_a.

6.5.2 Statistical concepts for multi-dimensional problems

The following list summarizes the vectors and vector operators used in this and some of the following sections, as well as in the wider literature on this subject, and will serve as a reference for the discussion. The symbols are generally consistent with the unified notation proposed in Ide *et al.* (1997). State vectors $\boldsymbol{x}$ may be defined on a model grid (or they can define spectral coefficients). Depending upon the setup of the analysis, the unknown analysis vector $\boldsymbol{x}_a$ and the known background vector $\boldsymbol{x}_b$ can define the values of a single variable in a two-dimensional space – e.g., $\boldsymbol{T}_b(x, y)$. Or the vectors can represent the three-dimensional structure of a single variable – e.g., $\boldsymbol{T}_b(x, y, z)$. Or they can define all the variables in the three-dimensional space – e.g., $\boldsymbol{x_b} = (\mathrm{Psfc}_b(x, y), T_b(x, y, z), q_b(x, y, z), u_b(x, y, z), v_b(x, y, z)$, etc.). The model background and analysis fields are defined by column vectors ordered by grid point and by variable, where the vector length n is the product of the number of variables and the number of grid points.

State and observation vectors are defined as follows.

- $\boldsymbol{x}_t$ The true model state vector. As described in Section 6.5.1, it represents the best-possible state that can be defined on the model grid. It is not the same as a perfect-state vector, which corresponds exactly with the atmospheric state (perfect observations), because of the unavoidable representativeness error. The dimension is n.
- $\boldsymbol{x}_a$ The analysis model state vector. The dimension is n.
- $\boldsymbol{x}_f$ The forecast model state vector. The dimension is n.
- $\boldsymbol{x}_b$ The background model state vector. The dimension is n. If a model forecast is used to define this, as in sequential initialization, $\boldsymbol{x}_b = \boldsymbol{x}_f$.
- $\boldsymbol{y}$ Vector of observations. The dimension is p.

Error covariance matrices are defined as follows:

- **B** The covariance matrix of the background (or forecast) errors. The background-error matrix has dimensions $n \times n$. It is very important to have reasonable estimates for this matrix because it controls the influence function for the analysis increment, in terms of its magnitude and shape. Regarding the shape, it defines the spreading of information from an observation to the analysis grid. And regarding magnitude, when background errors are large, observations are given greater weight. If this error covariance matrix is relatively accurate, a better adjustment of the gridded background to the observations will result, and observations will be used more effectively. That is, the information in the innovation vector that is defined at an observation point will be translated by **B** into a spatially variable analysis increment that is applied at surrounding grid points in such a way as to minimize the analysis error. In a scalar system, the background error covariance is simply the variance, or the average squared departure from the mean, where

$$\mathrm{B} = \overline{(\varepsilon_b - \bar{\varepsilon}_b)^2}\,.$$

  In a multi-dimensional system,

$$\mathbf{B} = \overline{(\varepsilon_b - \bar{\varepsilon}_b)(\varepsilon_b - \bar{\varepsilon}_b)^T}\,,$$

  which is a square, symmetric matrix with variances along the diagonal. For a very simple three-dimensional system,

$$\mathbf{B} = \begin{bmatrix} var(e_1) & cov(e_1,e_2) & cov(e_1,e_3) \\ cov(e_1,e_2) & var(e_2) & cov(e_2,e_3) \\ cov(e_1,e_3) & cov(e_2,e_3) & var(e_3) \end{bmatrix}.$$

  The off-diagonal terms are cross-covariances between each pair of “variables” in the model, where, as noted earlier, the term variable here corresponds to the value of each physical dependent variable at each grid point. A variable pair can be the same model dependent variable at two different points, or it can be two different dependent variables. The number of variables, and the dimension of the matrix, is the product of the number of physical variables and the number of grid points. There are three approaches for estimating the covariance matrix.

1. *Precalculated error covariances* – Some data-assimilation methods use precalculated covariances that are based on (a) an average over many different observed states of the atmosphere, (b) theoretical considerations, or (c) model simulations. In any case, the statistics may be spatially and temporally homogeneous (i.e., not dependent on the specific meteorological regime or synoptic situation). Observation-based covariances are ideally calculated from a dense and homogeneous network of sensors with uncorrelated errors, where the innovation vector $[\boldsymbol{y}-\boldsymbol{H}(\boldsymbol{x}_b)]$ (observation minus forecast) is calculated for varying separations between the locations of $\boldsymbol{y}$ and $\boldsymbol{x}_b$. In contrast, the so-called NMC (US National Meteorological Center) method is based entirely on model simulations, and is discussed briefly in Section 6.8 on three-dimensional variational (3DVAR) analysis. Examples of two different assumed decreases in correlation with increasing distance between $\boldsymbol{y}$ and $\boldsymbol{x}_b$ are seen in Fig. 6.8. See Schlatter (1975), Hollingsworth and Lönnberg (1986), Lönnberg and Hollingsworth (1986), Thiebaux *et al.* (1986), Bartello and Mitchell (1992), Xu and Wei (2001, 2002), and Xu *et al.* (2001) for additional discussion of the calculation of the non-regime-dependent error-covariance matrix.
2. *Nonoptimal, anisotropic spatial weighting* – One class of such methods employs information about the orographic elevation to control the spread of the innovation vector at lower elevations in the model atmosphere. The logic here is that covariances should be smaller between points on opposite sides of a mountain ridge, so an observation on one side has a weaker effect on grid points on the opposite side. So, the distribution of the analysis increment is anisotropic, with smaller increments (adjustments based on observations) on the opposite side of a barrier from an observation. Lanzinger and Steinacker (1990) employ this approach for an Optimal Interpolation (OI, see Section 6.7) analysis in the area of the Alps mountains. And, Miller and Benjamin (1994) made use of the fact that variables at two points will be better correlated if their potential temperatures and elevations are similar. So, the effective distance between an observation and a grid point was made proportional to the differences in elevation and potential temperature between the two points. Similarly, Dévényi and Schlatter (1994) spread their OI observation increments along isentropic surfaces.
3. *Fully regime-dependent error covariances* – The above methods do not account for the existence of “errors of the day” (Kalnay *et al.* 1997), which are weather-dependent errors in the background (forecast) field that should greatly influence the way that observations are analyzed. Ignoring these day-to-day variations in the covariance statistics can lead to large analysis errors. Some advanced data-assimilation methods, described in Section 6.11, calculate flow-dependent background-error covariances that evolve during the assimilation process. For example, Fig. 6.21 in Section 6.11.3 illustrates the spatial variability of these covariances.

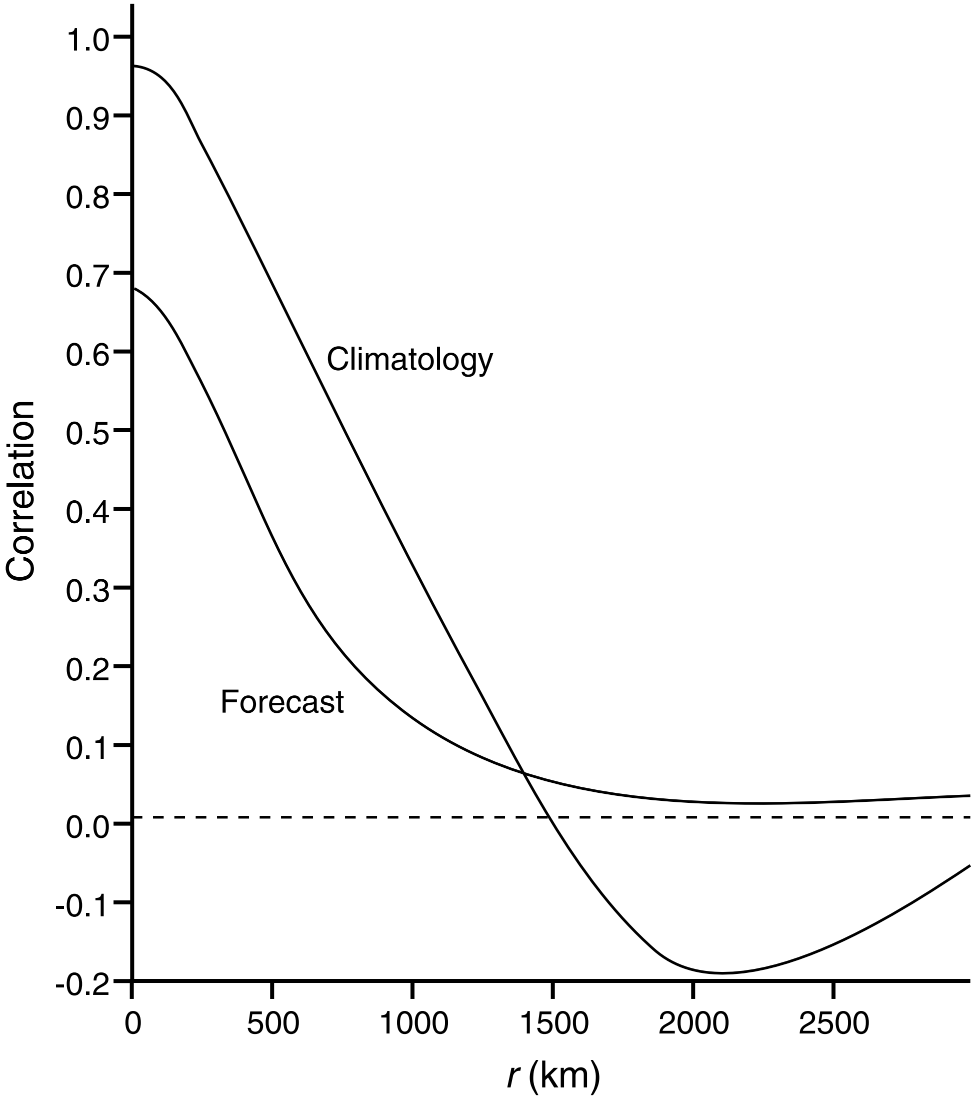


Fig. 6.8 Examples of two functions that define the correlation between observations and background values for 500-hPa geopotential height over North America. One curve is based on the use of climatology as the background (Schlatter 1975) and the other is based on the use of model forecasts (Lönnberg and Hollingsworth 1986). Adapted from Daley (1991).

Figure 6.9 illustrates the difference between the use of a typical isotropic covariance matrix to spread the influence of an observation, and the use of one that is regime dependent, for a two-dimensional (x,y) system. Shown are surfaces that define the value of the u velocity component on a grid. The background value (u_b) is spatially uniform (flat), and the observation y produces a positive innovation. In Fig. 6.9a, the analysis increment is distributed isotropically on the grid, around y, producing a conical impact

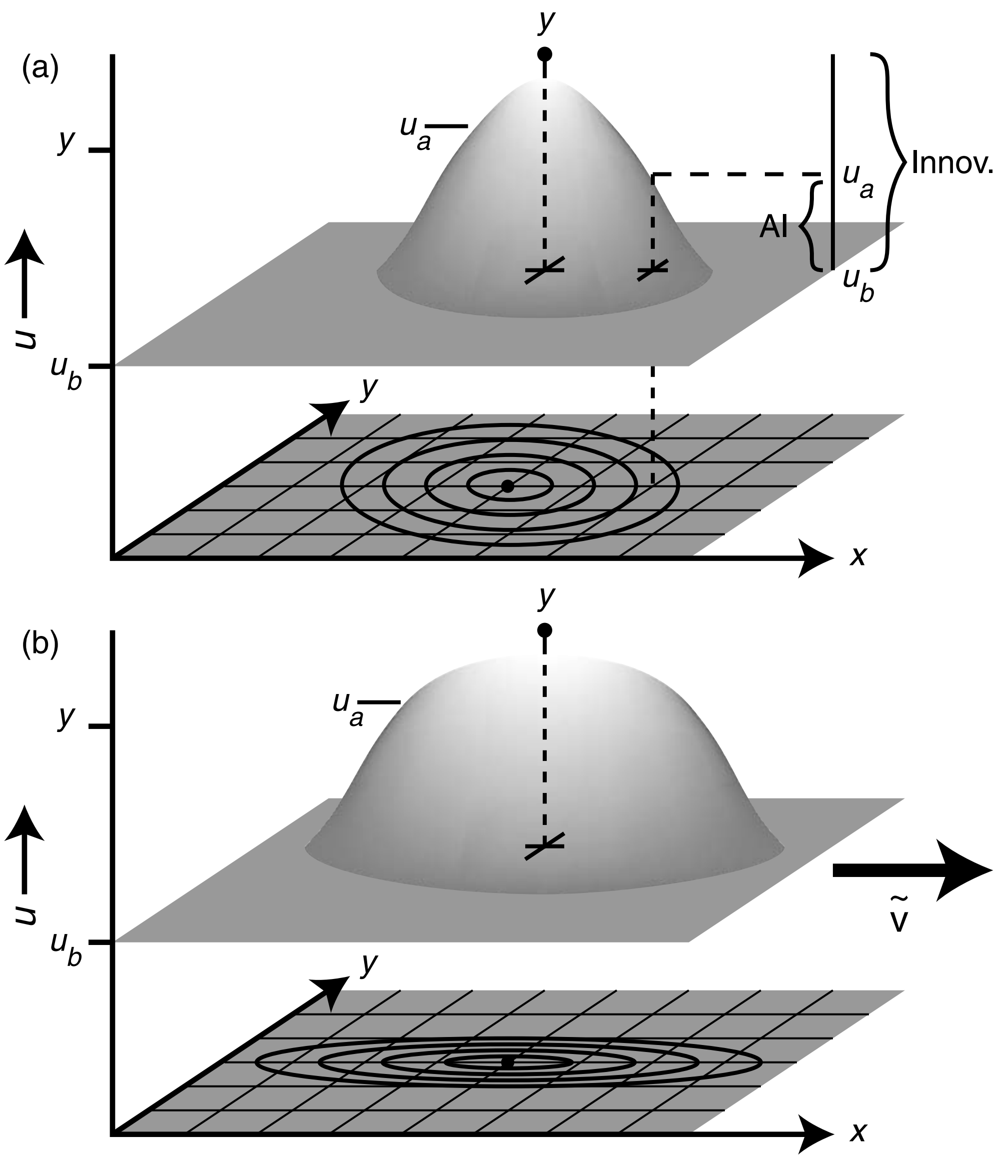


Fig. 6.9 Schematic illustrating the difference between the use of a typical isotropic covariance matrix to spread the influence of an observation (a), and the use of one that is regime dependent (b), for a two-dimensional (*x*, *y*) system. See the text for discussion.

of the observation (u_a). The shaded surface below shows isotachs for the resulting analysis on the grid. In Fig. 6.9b, the covariance matrix recognizes that, at this location, the wind speed pattern is streaked out in the direction of the total wind vector (shown), producing a nonisotropic, weather-regime-dependent distribution of the analysis increment and the isotachs.

- **R** The covariance matrix of the observation errors ($\varepsilon_o = \boldsymbol{y} - \boldsymbol{H}(\boldsymbol{x}_t)$). The dimensions are $p \times p$. Observation errors are often considered to be independent, especially when the observations are made by different instruments (e.g., in contrast to radiosonde profiles of observations). The variances are generally estimated based on knowledge of the instrument characteristics, which can be studied in the laboratory, even though representativeness errors and errors in the operator $\boldsymbol{H}$ can also be important. Most models of **R** are diagonal, or almost diagonal.
- **A** The covariance matrix of the analysis errors ($\boldsymbol{x}_a - \boldsymbol{x}_t$). The dimensions are $n \times n$.
- **Q** The covariance matrix of model forecast errors ($\boldsymbol{x}_f - \boldsymbol{x}_t$). The dimensions are $n \times n$.

Vector operators are defined as follows.

- $\boldsymbol{M}$ The model dynamic operator. For example, $\boldsymbol{x}_f(t + 1) = \boldsymbol{M}[\boldsymbol{x}_f(t)]$ refers to the fact that a model is used to advance the forecast value of vector $\boldsymbol{x}$ from time (t) to time ($t + 1$). Dimensions are from n to n.
- $\boldsymbol{H}$ The observation operator. This is also known as the forward operator. Dimensions are from n to p because the transformation is from model state space (n) to observation space (p). Imagine interpolating a variable from model grid points to the location of an observation.

The following defines the general problem of finding an optimal analysis, $\boldsymbol{x}_a$, of a set of model variables, given a background field, $\boldsymbol{x}_b$, available at a two- or three-dimensional set of grid points, and a set of observations $\boldsymbol{y}$ available at irregular locations $\boldsymbol{r}$ (see Fig. 6.10). Analogous to Eq. 6.16, which pertains to a scalar problem, the following relationship applies to a full multi-dimensional problem, where the vectors and vector operators have just been defined:

$$\boldsymbol{x}_a = \boldsymbol{x}_b + \mathbf{K}(\boldsymbol{y} - \boldsymbol{H}[\boldsymbol{x}_b])\text{, where} \tag{6.20}$$

$$\mathbf{K} = \mathbf{BH}^{\mathrm{T}}(\mathbf{HBH}^{\mathrm{T}} + \mathbf{R})^{-1}. \tag{6.21}$$

As before, the variable **K** is a weight matrix of the analysis. Exactly as in Eq. 6.16, the innovation is multiplied by an optimal weight, and this defines the analysis increment, $\boldsymbol{x}_a - \boldsymbol{x}_b$. The gain matrix is obtained by multiplying the background error covariance in the observation space and the inverse of the total error covariance (the sum of the background and the observation error covariances). The larger the magnitude of the elements of $\mathbf{BH}^{\mathrm{T}}$, corresponding to an observation and an analysis variable at a grid point, the larger the weight with which the innovation vector is applied at that grid point. Regarding the inverse term $(\mathbf{HBH}^{\mathrm{T}} + \mathbf{R})^{-1}$, the larger the uncertainty in the observation, the smaller the observation increment will be weighted in the analysis. The vector $\boldsymbol{x}_a$ is the optimal least-squares estimate. Most of the references listed at the end of the chapter, e.g., Kalnay (2003), can be consulted for a derivation of the gain matrix in Eq. 6.21.

6.6 Successive-correction methods

One of the first procedures to be used for interpolating observations to a grid is called the *Successive-Correction* (SC) method (Bergthorsson and Doos 1955, Cressman 1959). Variations of this approach are in use today because it is simple and robust. As mentioned above, a first-guess field represents a best estimate of the variable defined on the grid, and it is corrected using successive adjustments in which the observations influence surrounding grid-point values. The process is defined by the following expression:

$$x_i{}^{n+1} = x_i{}^{n} + \frac{\sum_{k=1}^{K_i^n} w_{ik}^n (y_k - H(x_k^n))}{\sum_{k=1}^{K_i^n} w_{ik}^n + \varepsilon^2}, \tag{6.22}$$

where x_i^n is the n-th iteration estimation at grid point i, y_k is the k-th observation of x surrounding the grid point i, $H(x_k^n)$ is the value of the n-th estimate of x interpolated from the surrounding grid points to the observation point k, and ε^2 is an estimate of the ratio of the observation-error variance to the first-guess error variance. The weights can be formulated in various ways. In Cressman (1959) they are defined as

$$w_{ik}^n = \frac{R_n^2 - r_{ik}^2}{R_n^2 + r_{ik}^2} \quad \text{for} \quad r_{ik}^2 \le R_n^2 \tag{6.23}$$

$$w_{ik}^n = 0 \quad \text{for} \quad r_{ik}^2 > R_n^2,$$

where r_{ik}^2 is the square of the distance between an observation point k and a grid point i. Figure 6.10 illustrates an influence region defined on a field of regularly spaced grid points and irregularly distributed observations.

For the initial iteration, x_i^0 in Eq. 6.22 is the value of the first guess. For each grid point, i, in the first-guess field, each of the K observations that is within the first radius of influence, R_0, of the grid point is used to adjust the first guess. The adjustment for each observation is weighted based on its distance from the grid point and on the difference between the observed value and the value of the first guess interpolated to the observation point. The difference between the first guess and the observation is, as before, called the innovation, and its weighted distribution around each observation point is isotropic. This process is repeated with successively smaller radii, given the constraint that there should be at least a few observations within each area of influence. Thus, the first guess is adjusted on the broader scales based on the effects of a large number of observations, and a progressively smaller number of nearby observations is reused in subsequent iterations to account for local effects.

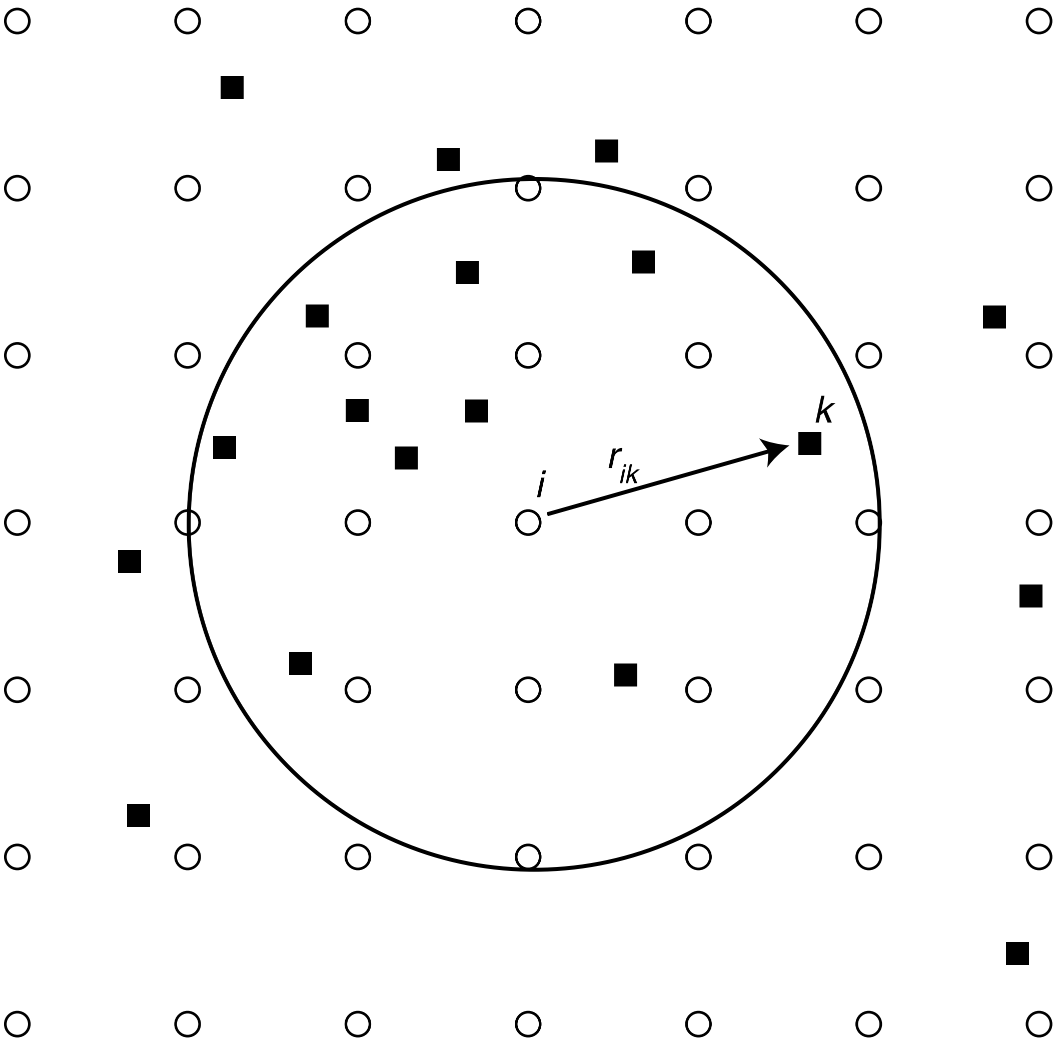


Fig. 6.10 Schematic of regularly spaced grid points (open circles) and irregularly distributed observations (black squares). Also shown is a circular influence function around grid point *i*, with a displacement vector ***r*** between that grid point and observation *k*.

If ε^2 is defined to be zero, the implication is that the observations are perfect. This means that the procedure will faithfully analyze to the observations, to the point of resulting in a bulls-eye pattern in the isopleths that encircle a bad grid-point value. By using a realistic value for ε^2, the adjustment to the observation in Eq. 6.22 is smaller and the first guess is given more weight. With $\varepsilon^2 = 0$, the impact of a single bad observation can be reduced by not using small radii of influence, so that multiple observations influence each grid point. Figure 6.11 shows a one-dimensional schematic of the generation of an analysis through the correction of a background field within an influence region of observations.

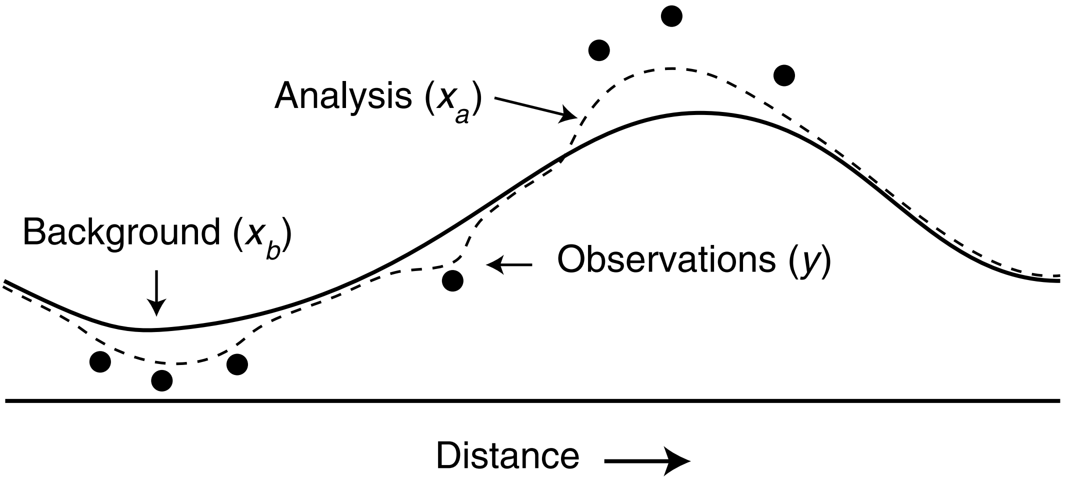


Fig. 6.11 Schematic showing the generation of an analysis (dashed line) through the correction of a background field (solid line) within an influence region of observations (black circles). The background and analysis are coincident outside the influence of the observations.

Another version of this method was developed by Barnes (1964, 1978), where one of the advantages is that no independent first guess is needed. In fact, the first guess is defined by a weighted sum of the observations within a radius of influence:

$$x_i^0 = \sum_{k=1}^{K_i} w_{ik}^0 y_k .$$

This has advantages on small scales where there may be no operational model to produce a background estimate. The Barnes formulation is similar to that in Eq. 6.22, except that ε^2 is assumed to be zero because we have no first guess (i.e., the error is large). The weights are given by

$$w_{ik}^n = e^{-r_{ik}^2/2R_n^2},$$

where the radii of influence

$$R_{n+1}^2 = \gamma R_n^2$$

are reduced by a constant fraction (γ) for each iteration. Additional discussion of the SC method, and examples of recent applications, can be found in Daley (1991), Barnes (1994a,b), and Garcia-Pintado *et al.* (2009).

The above use of an isotropic distribution of the analysis increment around each observation point obviously makes this method easy to implement, but it is an unnecessary approximation. Alternatives are based on the aforementioned idea that the spatial influence of observations can be related to prevailing meteorological structures. For example, observations should probably not be used to influence grid points on the opposite side of a front. And, in the atmosphere, scalar variables will be spread out more in the along-stream rather than the cross-stream direction. Regarding the latter point, an elliptical weighting function whose aspect ratio is proportional to the wind speed is an option (Benjamin and Seaman 1985). If the flow is curved, the semi-major axis of the ellipse can be curved accordingly. Such a weighting pattern has been called a banana function, for obvious reasons. And, Stauffer and Seaman (1994) adjusted the weight of an observation at a low-level grid point based on the surface-elevation difference between the two points. Obviously these weights are not optimal in any statistical sense. See Otte *et al.* (2001) for a description of other methods of weighting observations based on prevailing meteorological structures. And, see Bratseth (1986) for a discussion of how the SC method can be formulated so that it is equivalent with the statistical optimal-interpolation method (see next section).

6.7 Statistical interpolation (optimal interpolation)

Statistical interpolation is sometimes referred to as optimal interpolation. Equations 6.20 and 6.21 serve as the basis for OI. The unknown analysis and known background can be two-dimensional fields of a single variable, or three-dimensional fields for all the model dependent variables. The benefit of the OI method and three-dimensional variational (3DVAR) assimilation described below, relative to the SC method above, is that the spatial distribution of analysis increments is defined by the background-error covariance matrix that is based on archived model solutions or climatology (observations). In contrast, with the SC method the weight is often isotropic and somewhat arbitrary, only depending on distance from the observation. A fundamental computational-cost-saving concept in OI is that for each model variable (again, a model variable is one dependent variable at one grid point), only a few nearby observations are considered important in determining the analysis increment. These observations are selected based on empirical criteria, where it is assumed that distant observations would have small background error covariances $\mathbf{BH}^{\mathrm{T}}$.

The OI approach has been most often applied in intermittent-analysis schemes, such as depicted in Figure 6.6b. That is, the OI is used for the “Anal” at the beginning of each cycle. The model is integrated from the time of one analysis to the time of the next one. This provides the background vector $\boldsymbol{x}_b$, while all the observations that are available in the analysis time window are used to build the vector $\boldsymbol{y}$. Methods of defining $\mathbf{B}$ for OI approaches are described in Thiebaux and Pedder (1987) and Hollingsworth and Lönnberg (1986), and involve differencing the short forecast and the radiosonde observations (see the discussion of the background-error covariance matrix in Section 6.5.2). An advantage of OI is the simplicity with which it can be implemented, and the modest cost if the necessary assumptions can be made (e.g., observation selection).

6.8 Three-dimensional variational analysis

It was shown in Section 6.5.1 that there is a correspondence between two methods for the optimal analysis of a scalar: (1) minimizing the analysis error variance (by finding the optimal weights through a least-squares approach) and (2) using a variational approach (finding the analysis that minimizes a cost function that is a measure of the distance of the analysis to both the background and the observation). This correspondence also holds true for analyses of multi-dimensional fields, as described in the previous section for OI.

Lorenc (1986) showed the formal equivalence of the approach used in OI (where an optimal gain matrix **K** is found that minimizes the analysis-error covariance matrix) and a particular variational-assimilation problem. The latter is used in 3DVAR analysis, and corresponds to finding an optimal analysis field, $\boldsymbol{x}_a$, that minimizes a cost function, such that the cost function is defined as the sum of (1) the distance between $\boldsymbol{x}$ and $\boldsymbol{x}_b$, weighted by the inverse of the background-error covariance and (2) the distance to the observation $\boldsymbol{y}$ weighted by the inverse of the observation-error covariance. Mathematically, the cost function is

$$J(\boldsymbol{x}) = J_b(\boldsymbol{x}) + J_o(\boldsymbol{x}) = \frac{1}{2}(\boldsymbol{x}-\boldsymbol{x}_b)^T\boldsymbol{B}^{-1}(\boldsymbol{x}-\boldsymbol{x}_b) + \frac{1}{2}(\boldsymbol{H}(\boldsymbol{x})-\boldsymbol{y})^T\boldsymbol{R}^{-1}(\boldsymbol{H}(\boldsymbol{x})-\boldsymbol{y}), \tag{6.24}$$

and the gradient with respect to $\boldsymbol{x}$ is

$$\nabla J(\boldsymbol{x}) = 2\boldsymbol{B}^{-1}(\boldsymbol{x}-\boldsymbol{x}_b) - 2\boldsymbol{H}(\boldsymbol{x})^T\boldsymbol{R}^{-1}(\boldsymbol{y}-\boldsymbol{H}(\boldsymbol{x})).$$

Note the parallel between Eqs. 6.24 and 6.19. The control variable, the variable with respect to which the cost function is minimized, is the state vector, $\boldsymbol{x}$. The minimum of the cost function can be found analytically (e.g., Kalnay 2003), but in practice it is far less computationally demanding to estimate it iteratively by performing multiple evaluations of both equations. The minimum is approached by using a minimization, or descent, algorithm such as the conjugate-gradient or quasi-Newton methods. Only a small number of iterations are used, to produce an approximate minimum. Figure 6.12 illustrates the minimization process for a two-variable model space. The quadratic cost function has the shape of a paraboloid. The initial point in the iteration is generally taken to be the background value, $\boldsymbol{x}_b$, and the final point is $\boldsymbol{x}_a$, the approximate location of the minimum of J. Each step of the iteration moves the estimate down the gradient of the cost function.

Despite their formal equivalence, 3DVAR has a few advantages relative to OI. They are listed below, and are discussed more extensively in Kalnay (2003).

- In 3DVAR, there is no selection of only a limited number of observations that are within an influence region of a grid point. All observations are used simultaneously, which leads to a smoother analysis.
- The forecast- or background-error covariance, $\boldsymbol{B}$, is defined using fewer assumptions in 3DVAR. In particular, the so-called NMC (now NCEP) method (Parrish and Derber 1992)

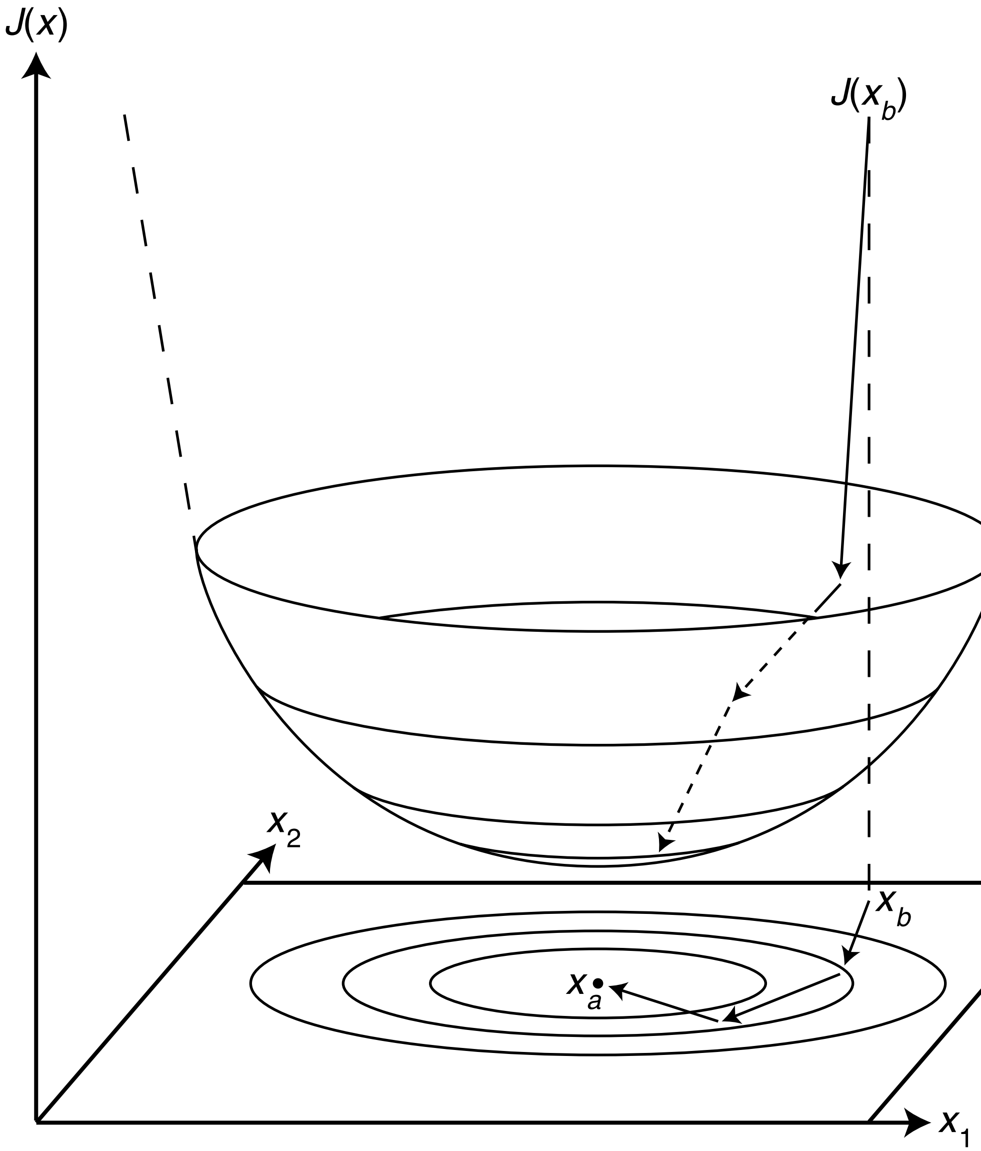


Fig. 6.12 Schematic showing the variational cost-function minimization for a two-variable (x_1, x_2) model space. See the text for details.

is generally employed. This does not depend on measurements, as with OI (Hollingsworth and Lönnberg 1986, Thiebaux and Pedder 1987) but rather the error covariance is based on the average (over perhaps 50 instances) difference between forecasts valid at the same time. Even though any time lag and lead time can be used, the following example is based on 24-h and 48-h forecasts.

$$\boldsymbol{B} \approx \alpha E\{[\boldsymbol{x}_f(48h) - \boldsymbol{x}_f(24h)][\boldsymbol{x}_f(48h) - \boldsymbol{x}_f(24h)]^T\}\,.$$

Even though this is the covariance of the forecast differences, which is only a surrogate for the background- or forecast-error covariance, it has been shown to produce better results than the methods used in OI, where forecasts and observations are employed. Parrish and Derber (1992) and Rabier *et al.* (1998) point out that the radiosonde network is not sufficiently dense to properly estimate structures. Nevertheless, the covariances in 3DVAR are typically isotropic and climatological (e.g., Fig. 6.8) – i.e., they are not situation (case, regime) dependent – which is a major disadvantage.

- Additional constraints can be added to the cost function, such as those related to dynamical-balance relationships. For example, Parrish and Derber (1992) employed an additional penalty term in Eq. 6.24, forcing the analysis increments to approximately satisfy the balance equation. In contrast, it was often found necessary to follow an OI analysis with a Nonlinear Normal-Mode Initialization (NNMI, see Section 6.10.3). Importantly, with the implementation of 3DVAR it became unnecessary to perform a separate balancing, or initialization, step in the analysis cycle (cf., Fig. 6.6b).
- Prior to the availability of 3DVAR, satellite radiances had to be processed through a retrieval algorithm that generated values of a model dependent variable that would be assimilated. But with 3DVAR, the radiances themselves can be assimilated directly.

6.9 Diabatic-initialization methods

Diabatic initialization involves the use of observations of precipitation and other variables to produce estimates of the four-dimensional distribution of latent-heating rate, and the associated humidity and divergence fields, in a model during the initialization process. This has two motivations. One is that it employs precipitation observations in a model initialization, and the other is that it helps produce reasonable vertical-motion and moisture fields at the initial time of a forecast. The latter goal is motivated by the fact that, typically, initial conditions do not have realistic vertical motions and humidities, and there is the resulting spinup period during which the model must internally develop such precipitation-scale circulations and humidity fields.

Some early diabatic-initialization methods simply inserted estimated latent-heating profiles into the model grid columns during a preforecast dynamic-initialization period. The column-total latent heat was based on satellite-, rain-gauge-, or radar-estimated rain rates, and the vertical distribution of the heating was typically defined to be consistent with the model's parameterizations (Fiorino and Warner 1981, Danard 1985, Ninomiya and Kurihara 1987, Wang and Warner 1988, Monobianco *et al.* 1994). Sometimes, additional observations are assimilated during the preforecast period using Newtonian relaxation. Figure 6.13 shows a schematic of this method (a), and of other methods to be described shortly. The black bars on the time axis of Fig. 6.13a indicate the insertion of latent-heating-rate information at appropriate grid points. Also shown is the simultaneous use of Newtonian relaxation to assimilate other observations during the dynamic-initialization period. A related approach, referred to as latent-heat nudging, is described in Jones and Macpherson (1997), and is applied in Leuenberger and Rossa (2007) to a

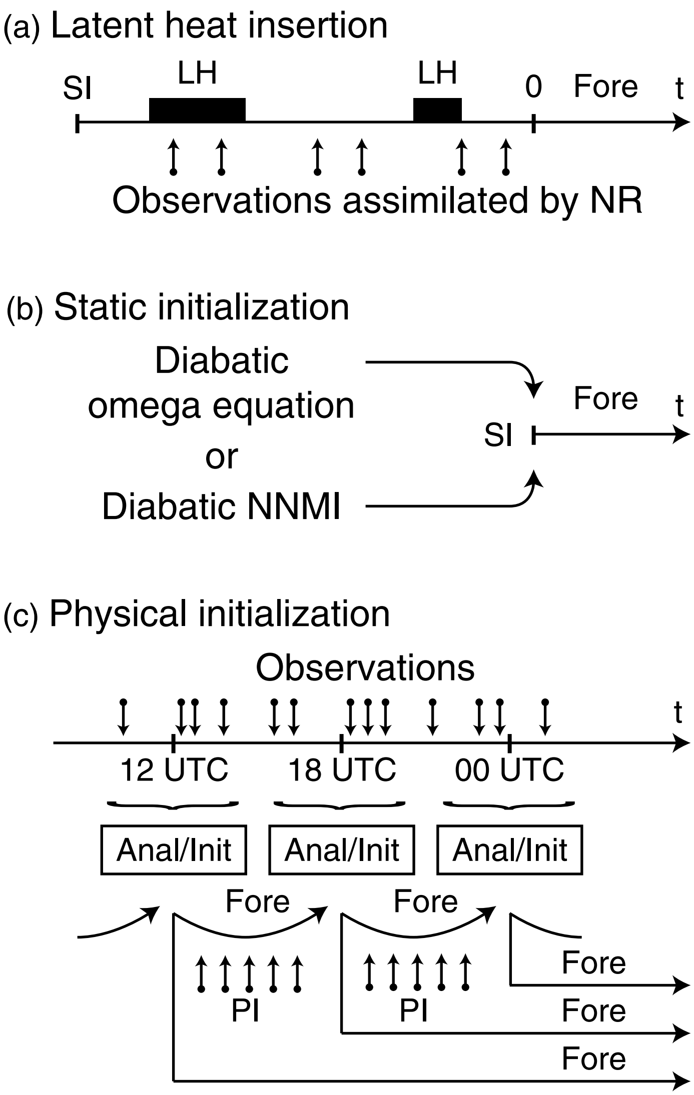

(a) Latent heat insertion
SI
LH
LH
0
Fore
t
Observations assimilated by NR
(b) Static initialization
Diabatic omega equation or Diabatic NNMI
SI
Fore
t
(c) Physical initialization
Observations
t
12 UTC
18 UTC
00 UTC
Anal/Init
Anal/Init
Anal/Init
Fore
Fore
PI
PI
Fore
Fore
Fore

Fig. 6.13 Schematics of three different types of diabatic initializations: (a) insertion of latent-heating (LH) profiles based on observed rain rates during a preforecast dynamic-initialization period, with possible simultaneous Newtonian relaxation (NR) using other observations; (b) the use of a diabatic nonlinear normal-mode initialization (NNMI) or a diabatic omega equation to incorporate precipitation processes in a static initialization (SI); and (c) a physical initialization where information about model dependent variables obtained from a physical initialization (PI, see text) is assimilated in an intermittent-assimilation cycle to improve the first-guess field.

mesogamma-scale rainfall simulation. Application of the method involves correcting the model's latent heating at each time step based on the ratio of the observed and model-simulated surface precipitation. Another approach is the use of a diabatic omega equation to define the vertical motion and divergent component of the wind in a static initialization

(Fig. 6.13b, Tarbell *et al.* 1981, Salmon and Warner 1986). Turpeinen *et al.* (1990) and Raymond *et al.* (1995) include a summary of these early studies.

Simultaneous with these efforts was the development of a diabatic version of the NNMI method (Wergen 1988) described in Section 6.10.3. This process incorporated estimated latent-heating rates into a static initialization (Fig. 6.13b), where the objective was also to provide initial conditions that did not require significant spinup of the precipitation processes during the early period of the forecast. Examples of the use of NNMI for diabatic initialization can be found in Puri (1987), Heckley *et al.* (1990), Turpeinen (1990), Turpeinen *et al.* (1990), and Kasahara *et al.* (1996).

The so-called physical-initialization method, described in Krishnamurti *et al.* (1991), also enables the use of rainfall-rate estimates and other observations during a preforecast integration to provide model initial conditions that contain spun-up vertical motion, horizontal divergence, and humidity fields. The details of the method can vary, but a common feature is that reverse algorithms are employed for relationships in the model that involve the moisture variables: the parameterizations for convection and outgoing longwave radiation (OLR), and similarity theory. For example, convective parameterizations provide the convective rain rate as a function of grid-resolved variables such as the vertical moisture profile. "Reversing" the convective-parameterization algorithm allows observations of the rain rate to be translated into an estimate of grid-resolved model dependent variables, which can be ingested into model initial conditions using Newtonian relaxation or they can replace simulated variables during a preforecast integration (Fig. 6.13c). Following the conceptual explanation in Treadon (1996), consider the equation $y = f(x)$ that represents a simple model parameterization relationship, where x is a grid-resolved variable, and y is an observed, parameterized variable. For example, let y be OLR, and x the variables that are used in its parameterization (e.g., temperature, specific humidity, etc.). Thus, based on measurements of OLR, the reverse algorithm provides estimates of the large-scale forcing, x. This estimate can be assimilated, or it can be fed back into the forward algorithm to provide an improved value after iteration. Physical initialization methods have been applied most often in the tropics, where the scarcity of conventional observations increases the dependence of the forecast on a reasonable first guess. Krishnamurti *et al.* (1994, 2007) and Shin and Krishnamurti (1999) show that this method leads to a significant improvement in the skill of forecasts of tropical rainfall, global cloudiness, land-surface hydrology, and tropical cyclones. The method has been tested with the Florida State University spectral model, the US Navy's NOGAPS model (Van Tuyl 1996), and the NCEP Global Data Assimilation System (Treadon 1996). A recent application of the method is described in Milan *et al.* (2008). Physical initialization has also been employed for ensemble prediction (Chaves *et al.* 2005, Ross and Krishnamurti 2005) as well as with LAMs (Li and Lai 2004, Nunes and Cocke 2004).

The above methods are not as likely to be effective for midlatitude precipitation events that are associated with large-scale forcing. For example, using observed precipitation-related fields to develop vertical motion and latent-heating rates in the initial conditions for a case of frontal precipitation is not going to be effective if the front, or the cyclone itself, is in the wrong location in the model solution. In this case, the lack of correct large-scale forcing for the specified precipitation in the model will cause the precipitation to dissipate during the early period of a forecast. Similarly, precipitation imposed in the solution through diabatic initialization will not persist in situations of air-mass convection if the large-scale environment is not sufficiently unstable. In contrast, for convection within air masses with realistic stability, and for tropical cyclones, the methods will be more likely to improve predictive skill.

Other methods, such as three- and four-dimensional variational data assimilation (discussed elsewhere), can also be used to assimilate observations that represent precipitation-scale processes in model initial conditions.

6.10 Dynamical balance in the initial conditions

This section focusses on the need to have a realistic dynamic balance in model initial conditions, the consequences of not having that balance, and methods that have been employed to ensure that a reasonable balance exists. The first section reviews the relevance to initialization of the concept of geostrophic adjustment, the second describes how integrating a model for a period prior to the initialization can contribute to an improved balance, and the third briefly summarizes the use of diagnostic relationships for achieving a balance.

6.10.1 Geostrophic-adjustment concepts, and relevance to initialization

The atmosphere on synoptic and planetary scales is in approximate geostrophic balance. Thus, if the mass and wind fields on these scales are significantly inconsistent with each other in the model initial conditions, inertia–gravity waves will cause those fields to adjust toward balance during the early period of the integration. Of course, the large-scale mass and wind fields in the real atmosphere are always in a state of continuous imbalance and adjustment. But, inertia–gravity waves that result from poorly defined initial conditions are not physically based, can be of large amplitude, and are potentially problematic in a model solution. This adjustment process is relevant to numerical modeling for a few reasons. First, if the inertia–gravity waves that effect the adjustment are of significant amplitude, they can mask the true meteorological features in the model solution until the waves are damped or propagate away from the area of interest. This could mean that the model solution will not be useful for the first 12–24 h after initialization. Also, the resulting waves can have troublesome interactions with the LBCs of LAMs. Lastly, the geostrophic-adjustment process can cause good-quality observations in one field to compensate for poor-quality observations in the other, or poor-quality observations in one field can negate the value of good-quality observations in the other.

To better understand the geostrophic-adjustment process, consider the simple geopotential-height pattern in Fig. 6.14. A parcel of air at location 1 is moving with a speed that is in geostrophic balance with the local pressure gradient. As the parcel moves toward location 2, the local pressure gradient becomes greater and the motion is subgeostrophic. The fluid system can respond in two ways to regain a balance: Either the parcel's speed can

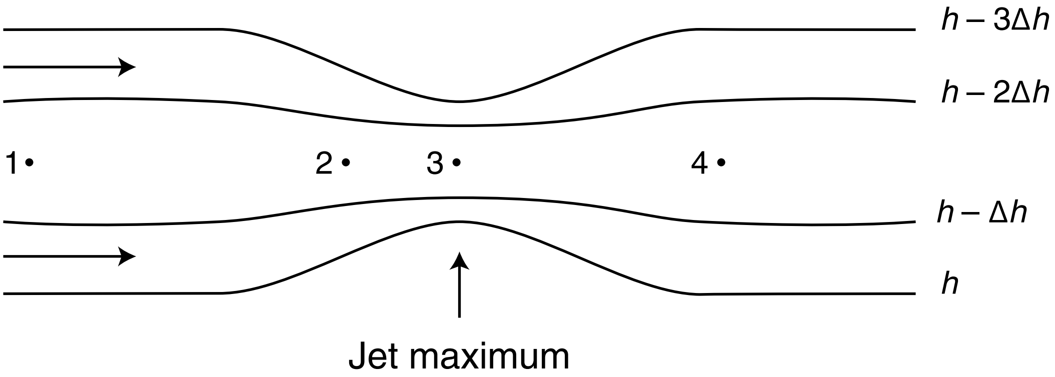


Fig. 6.14 Schematic of a jet maximum in the height field, where the flow is from left to right. The numbers represent locations of air parcels, where the discussion focusses on whether the wind or the mass (height) fields change as the parcels move to the right.

increase as it moves into the geostrophic jet maximum, or the pressure-gradient maximum can move downstream, causing the pressure-gradient at location 2 to weaken. Of course a similar question could be posed about the response of the system as parcels move farther downstream, away from the pressure-gradient maximum (from location 3 to 4). To resolve this, consider the ways in which the mass and wind fields can adjust in such situations. The shallow-fluid equations (see Chapter 2) represent a simple framework for addressing this, which nevertheless contains all the relevant dynamics. Two of the admissible wave solutions are gravity waves and inertia waves. Both mechanisms operate simultaneously to reconcile an imbalance. The inertia waves modify the winds, and the gravity waves modify the mass field (the fluid depth, in the shallow-fluid system). As an indicator of how much of the adjustment results from changes in each of the mass field and momentum field, consider the periods of these waves. For inertia waves $T_i = 2\pi/f$, and for gravity waves $T_g = L/\sqrt{gH}$, where L is the length of the gravity wave (defined by the horizontal scale of the imbalance) and H is the depth of the imbalance (the vertical scale). Given that both types of waves simultaneously act to adjust the atmosphere toward the geostrophic state, the wave mode with the shortest period accomplishes most of the adjustment. To define the condition where there is equal adjustment from both types of waves, the expressions for the two periods can be equated. Solving for the wavelength yields

$$L_R = \frac{2\pi\sqrt{gH}}{f},$$

where this length scale is the Rossby radius of deformation for the shallow-fluid system. For wavelengths shorter than this value, redistribution of the mass field through gravity waves is responsible for most of the adjustment, whereas for longer wavelengths, modification of the windfield by the inertia waves accomplishes most of the adjustment. The value of L for deep (tropospheric) midlatitude adjustments is ~15 000–18 000 km. So, for planetary scales and very-long synoptic scales, the winds adjust to the mass field when there is an imbalance. That is, the mass field does not change much, so we say that it dominates the adjustment. This is consistent with the fact that large-scale weather maps are analyzed in terms of geopotential height, and the winds are sometimes inferred from that field. For much smaller scales, the winds change little when there is an imbalance because the gravity waves act quickly to adjust the mass. At these scales, we say that the winds dominate the adjustment process. A related important point is that the short periods of the gravity waves on small scales means that adjustments take place quickly, relative to the situation on larger scales where the time scale is that of the inertial period (~17 h at 40° latitude). In tropical latitudes (small *f*), for an imbalance of a given length scale, the windfield changes less than in higher latitudes. For shallow adjustments, for example those that are limited to the boundary layer or lower troposphere, the mass field is more dominant than with deeper adjustments.

As mentioned earlier, the geostrophic-adjustment process has many implications for model initialization and data assimilation. First, the model initial conditions need to be in a realistic state of balance, or the resulting excited gravity waves can mask the meteorological pressure field. As an example, Fig. 6.15a shows the evolution of the surface pressure at a point, after a well-balanced and a poorly balanced initialization of a LAM. For the poorly balanced initial conditions, the amplitude of the gravity waves created by the adjustment decreases with time, because the waves have propagated away from the area and possibly been damped by the model, but the sea-level pressure prediction for the first 6–12 h would have been relatively unusable by a forecaster. Also shown in the figure is the

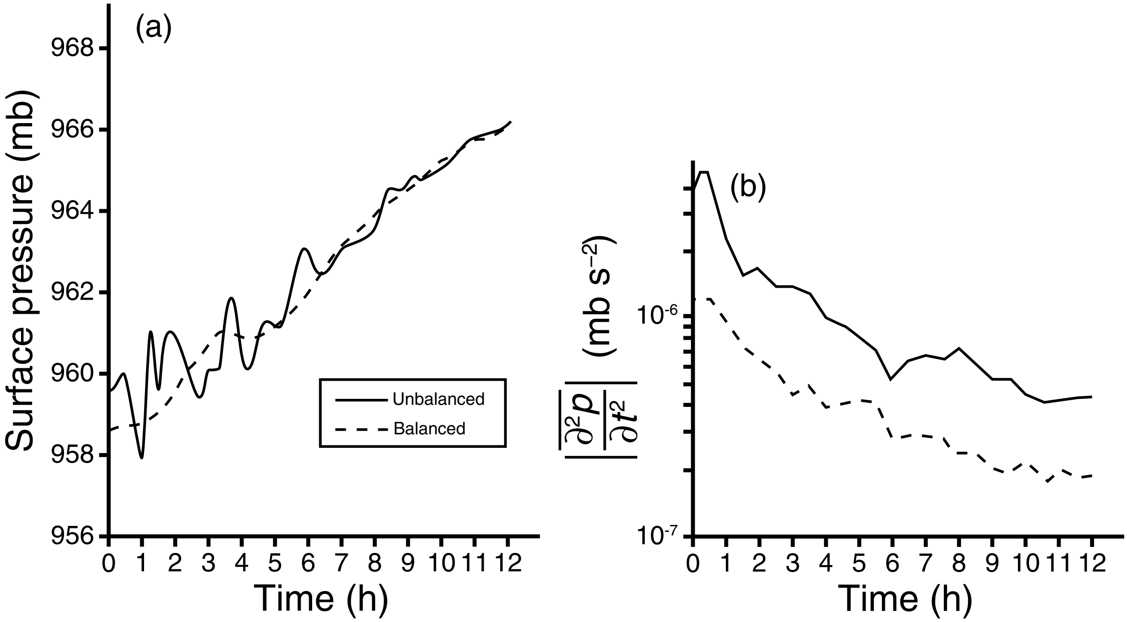


Fig. 6.15 Example of the model-simulated surface pressure at a grid point during the first 12 h of a LAM simulation, based on well-balanced and poorly balanced initial conditions (a). Also shown (b) is the computational-domain average of the absolute value of the second time derivative of the surface pressure, a measure of the intensity of inertia–gravity wave activity, for two LAM initializations with different degrees of initial imbalances. Part (b) is adapted from Tarbell *et al.* (1981).

computational-domain average of the absolute value of the second time derivative of the surface pressure, a measure of the intensity of inertia–gravity-wave noise, for two initializations with different degrees of initial imbalances. In both cases, the inertia–gravity-wave intensity decreases by a factor of 5–10 during the first 12 h of the integration, but there is still a residual benefit after 12 h of the better-balanced initial conditions. Ballish *et al.* (1992) show similar plots for a global model, where the unbalanced initial conditions led to high-frequency aphysical surface-pressure oscillations during the first 12 h, with a change of over 5 hPa in 2 h. However, in that study the use of one type of NNMI (see Section 6.10.3 below) to produce a balance filtered the gravity waves too effectively, such that the real semi-diurnal tidal oscillations were removed during the first 24 h of the simulation. An alternative NNMI approach removed the large-amplitude spurious waves, but retained the tidal oscillations.

Knowledge of the adjustment process can also inform decisions about the variables that need to be better observed. For example, because of the dominance of the windfield during the adjustment in the tropics (even though the atmosphere is less geostrophic in those latitudes), the model initialization should emphasize the use of wind observations. Information from pressure observations will be lost in the adjustment. Similarly, at all latitudes, on the scales where the winds dominate the adjustment, assimilated mass-field information can be viewed as redundant.

The concept of adjustment can be viewed in the context of the errors in the initial fields. If the observations and the analysis system are perfect, the generated model initial conditions will be in perfect balance relative to the model equations, and there will be no artificial adjustment. If only the winds have errors, and the winds dominate the adjustment, error will be induced in the mass field during the adjustment process, and vice versa. This means that there should be an interconsistency or compatibility of initial-condition mass-field and windfield errors on the scale of the motion being studied. This concept of initial-condition-error interconsistency for large-scale motions was discussed extensively as part of the Global Atmospheric Research Program (Jastrow and Halem 1970). In the context of data assimilation, the idea is that assimilated observations should have error consistency. The problem of error inconsistency exists because of the exchange of error-related energy through dynamic adjustment. A convenient way of demonstrating the transfer of errors among model variables is through the use of a stochastic-dynamic model (Fleming 1971a,b). Here, the statistical moments of the dependent variables are explicitly predicted with the model equations. Figure 6.16 illustrates a simulation with a one-dimensional (x), spectral, stochastic-dynamic, shallow-fluid model, where the smallest resolved wave had a length of 1250 km. In this experiment, there was no initial error in the u component, but a spatially uniform standard error of 23 m was defined in the initial height field. The plot shows the temporal evolution of the grid-average second moment of the fluid depth for different spatially uniform initial errors in v. When there was no initial error in u or v, the h error decreased with time because the perfectly known v component caused an adjustment of h toward a compatible value. For larger initial v errors, there was a greater consistency in the initial mass and wind errors, and there was less net change in $\overline{\sigma}_h$. These results are independent of the first moments that are specified for the variables.

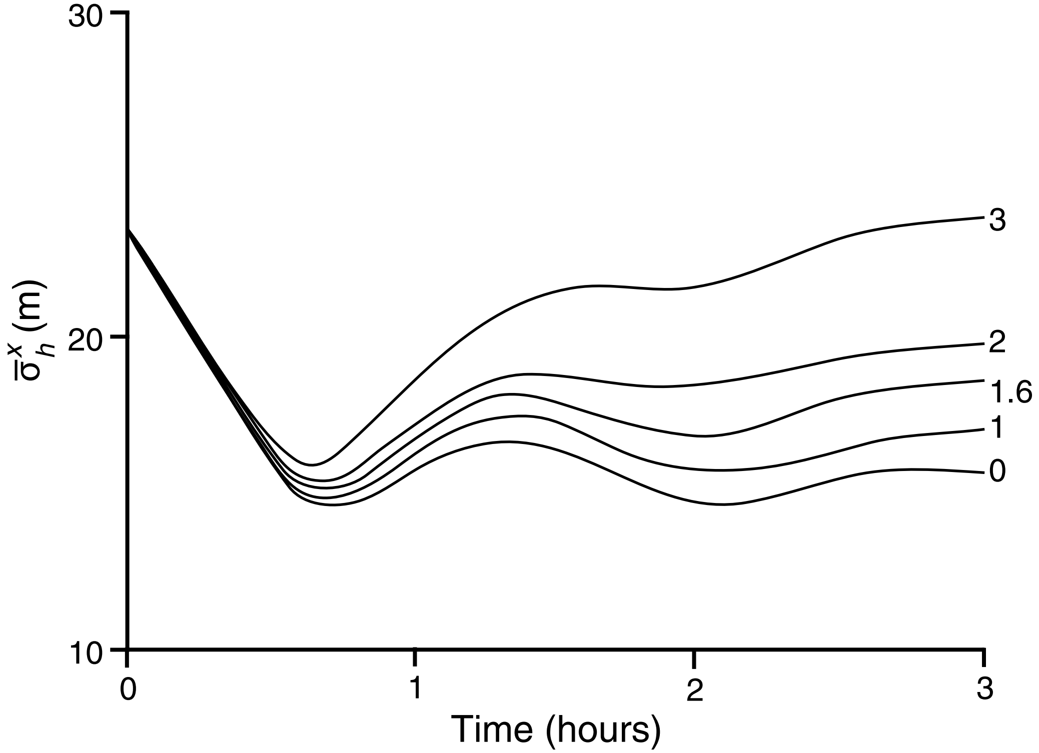


Fig. 6.16 Adjustment of the uncertainty in the fluid depth in a one-dimensional (x), spectral, stochastic-dynamic, shallow-fluid model, where the smallest resolved wave had a length of 1250 km. A spatially uniform standard error of 23 m was defined in the initial height field. The different curves are the simulated height-field errors resulting from the use of different initial standard errors (m s^{-1}) in the v wind component (curve labels on the right).

6.10.2 Preforecast integration of the model equations for achieving a dynamic balance

The objective of preforecast integrations has been mentioned before in the context of allowing ageostrophic circulations, related to precipitation and other processes, to spin up before the initial time of a forecast. Observations can also be assimilated during this period. In addition, initial imbalances can be reconciled during this preforecast integration, if there is some way to eliminate the inertia–gravity waves that are produced. Figure 6.13a is an example of a preforecast integration during which observations are assimilated, ageostrophic circulations spin up, and initial imbalances are reconciled. Another approach is to utilize a damping differencing scheme, such as the Euler-backward method described in Chapter 3, during a backward–forward integration of the model equations (e.g., Nitta and Hovermale 1969). A version of the forecast model that includes only the reversible processes is used; e.g., precipitation and explicit diffusion terms are removed. As the reversible model is integrated backward and forward by one or more time steps, inertia–gravity waves are generated by the adjustment process, and they are damped by the differencing scheme. At the end of each forward–backward cycle, the better observed of the wind or mass field can be recovered. This effectively forces the more-poorly observed field to adjust to the better-observed one. See Figure 6.17 for a schematic of the process. For additional discussion of this type of initialization method, see Fox-Rabinovitz and Gross (1993), Fox-Rabinovitz (1996), and Kalnay (2003).

6.10.3 The use of diagnostic relationships for achieving a balance

It is important to recognize that the use of a hypothetical method that provides a balance between the gridded initial wind and mass fields that is perfect with respect to the real atmosphere, will nevertheless result in the generation of gravity-wave noise in the model. This is because the model obviously represents a numerical approximation of the real atmosphere, so it has its own unique balance that is a function of the numerical methods and their associated truncation error. Thus, any initialization method, whether it is static or dynamic, should employ equations that use numerical approximations that are similar to those in the forecast model. Otherwise, the truncation error inconsistency between the forecast model and the initialization method will be a source of inertia–gravity wave noise in the model solution.

Early models used simple large-scale balance relationships to define somewhat compatible mass and wind fields in the initial conditions. The earliest and most primitive

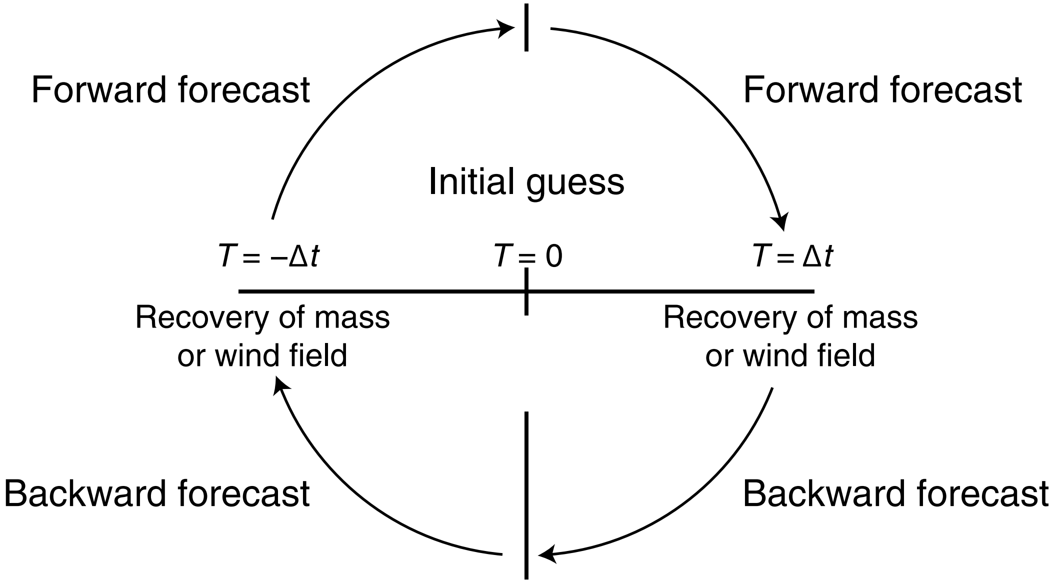


Fig. 6.17 Schematic of a model-initialization method that utilizes a backward–forward integration, with a reversible version of the model, to permit the geostrophic-adjustment process to be completed before the initial time of a forecast. The integration is performed with a damping differencing scheme. Adapted from Nitta and Hovermale (1969).

approach was to use the geostrophic relationship. More-complete balances were achieved through other diagnostic equations such as a combination of a divergent balance equation (see Holton 2004) and a vertical–velocity equation (e.g., Tarbell *et al.* 1981).

Another method of diagnosing a balanced set of initial conditions is the previously mentioned NNMI, introduced by Machenhauer (1977) and Baer and Tribbia (1977). It is applied after the analysis step (e.g., that uses OI). As the name implies, it requires the determination of a model's "normal modes" (i.e., solutions of a linearized version of the model equations) as a first step, and then the high-frequency (inertia–gravity waves) and low-frequency (quasi-geostrophic) components of the model input data are separated. The high-frequency modes are assumed to have no meteorological significance, and are removed. This method has been used in the initialization of many operational modeling systems.

Kalnay (2003) summarizes a few shortcomings of the standard NNMI. One is that physically meaningful fast modes are removed with the rest. And, the importance of diabatic processes in the tropics led to the need for a diabatic NNMI (Wergen 1988). Ballish *et al.* (1992) describe a so-called incremental NNMI procedure in which the process is applied to analysis increments rather than to the complete analysis fields. This procedure substantially reduces the aforementioned problems with NNMI, as well as others. See Daley (1991) for additional information about NNMI.

6.11 Advanced data-assimilation methods

With the OI, 3DVAR, and SC sequential methods of data assimilation, the observations employed are made at or near the time of the analysis, and the process is repeated at regular intervals that are defined by the update cycle (e.g., every 6, 12, 24 h). In each case, the model is used to propagate the observations and background information from one analysis to the next. Unfortunately, many types of observations are not available at regular intervals. These so-called asynoptic observations are abundant, and are provided by satellites, aircraft, radars, etc. But, unless they are available near one of the standard initialization times, they are not very useful with sequential methods. Of course it is possible to perform some sort of temporal adjustment for off-time observations, but this is not a very satisfying process. And, with the above methods the background-error covariance matrix remains the same throughout a simulation, as if the forecast errors were statistically stationary. The solution is thus to employ a data-assimilation method that can use observations that are available at any arbitrary time, and one for which the background-error covariance evolves during the forecast.

6.11.1 Four-dimensional variational initialization

The method of four-dimensional variational (4DVAR) assimilation is a generalization of 3DVAR, to allow inclusion of observations that are distributed in time within an interval (t_0, t_n), where the subscript 0 denotes observations at the initial time of the assimilation period and the subscript n corresponds to the time of the last (n-th) observation that is assimilated. The following cost function must be minimized, where the control variable is $\boldsymbol{x}(t_0)$, the model state vector at the initial time of the forecast:

$$J[\boldsymbol{x}(t_0)] = \frac{1}{2}[\boldsymbol{x}(t_0) - \boldsymbol{x}_b(t_0)]^T \boldsymbol{B}_0^{-1}[\boldsymbol{x}(t_0) - \boldsymbol{x}_b(t_0)] + \frac{1}{2}\sum_{i=0}^{n}[\boldsymbol{H}(\boldsymbol{x}_i) - \boldsymbol{y}_i]\boldsymbol{R}_i^{-1}[\boldsymbol{H}(\boldsymbol{x}_i) - \boldsymbol{y}_i]. \tag{6.25}$$

The summation is over the number of observations, n. The zero-th observation applies at the beginning of the assimilation window. Note that, if observations are only available at a single time t_0, the cost function is the same as that used for 3DVAR (Eq. 6.24). That is, observations for $t > 0$ appear as additional penalty terms. This minimization problem is subject to the strong constraint that the sequence of model states, $\boldsymbol{x}_i$, represents solutions to the model equations. That is

$$\forall i, \quad \boldsymbol{x}_i = \boldsymbol{M}_{0 \to i}(\boldsymbol{x}),$$

where $\boldsymbol{M}_{0 \to i}$ is a model forecast operator that is applied from $t = 0$ to the time of the last observation. The 4DVAR process is illustrated in Fig. 6.18, where the ordinate is the model state vector $\boldsymbol{x}$ and the abscissa is time. With observations distributed throughout the assimilation period, the objective of the 4DVAR process is to estimate the state vector $\boldsymbol{x}_a$ that produces a model solution $\boldsymbol{M}$ that minimizes the cost function that has terms that (1) represent the distance to the background (the previous forecast) at the beginning of the

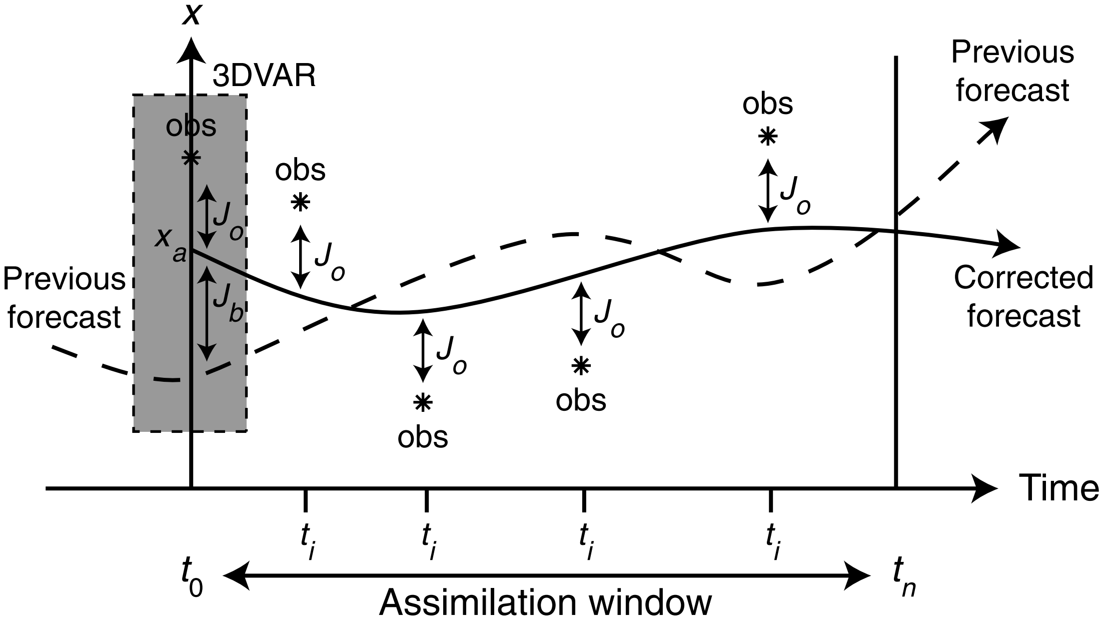


Fig. 6.18 Schematic illustrating the 4DVAR data-assimilation process. See the text for details.

interval and (2) are based on the observational increment computed with respect to the model solution at the time of the observations. That is, the process defines an initial condition that produces a forecast that best fits the observations within the assimilation window. In practice, if a forecast model is run with a 6-h update cycle, the assimilation window would extend from the time of the previous initialization to the time of the current one.

To solve the minimization problem, we must differentiate Eq. 6.25 with respect to the control variable, and solve for $\boldsymbol{x}$ at the minimum. For $\boldsymbol{J} = \boldsymbol{J}_b + \boldsymbol{J}_o$, the differentiation of $\boldsymbol{J}_b$ is identical to the process for 3DVAR. However, the evaluation of $\boldsymbol{J}_o$ and $\nabla\boldsymbol{J}_o$ requires a model integration from $t_0 \to t_n$, as well as an integration of the adjoint model defined as the transpose of the model operator $\mathbf{M}_i$ ($\mathbf{M}_i^T$). Note that $\mathbf{M}$ is the tangent linear version of $\boldsymbol{M}$. Constructing this adjoint model from the forecast model is a complex process, and, furthermore, the adjoint model must be maintained (bug fixes, improvements) in parallel with the forecast model. See Kalnay (2003) for mathematical details and additional references.

6.11.2 Extended Kalman filtering

A specific implementation of the least-squares analysis method is called the Extended Kalman Filter (EKF, Ghil and Malanotte-Rizzolli 1991, Bouttier 1994, Kalnay 2003, Hamill 2006). Equations 6.26a–e below summarize the process. As in OI, the EKF is based on the least-squares analysis method applied in the framework of sequential data assimilation, where each background is produced by a forecast that is initiated from the previous analysis. However, now the background error-covariance matrix is time dependent. The background (i.e., forecast) and analysis error-covariance matrices are now represented as $\mathbf{P}_f$ and $\mathbf{P}_a$, respectively. Compare the constant background error covariance $\mathbf{B}$ in the weight matrix used for OI (Eq. 6.21) with the time-dependent $\mathbf{P}_f(t)$ below in the gain matrix defined by Eq. 6.26c.

$$\text{State forecast } \boldsymbol{x}_f(t+1) = \boldsymbol{M}_{t\to t+1}(\boldsymbol{x}_a(t)) \tag{6.26a}$$

$$\text{Error-covariance forecast } \mathbf{P}_f(t+1) = \mathbf{M}_{t\to t+1}\mathbf{P}_a(t)\mathbf{M}^{\mathrm{T}}_{t\to t+1} + \mathbf{Q}(t) \tag{6.26b}$$

$$\text{Kalman-gain computation } \mathbf{K}(t) = \mathbf{P}_f(t)\mathbf{H}^{\mathrm{T}}(t)[\mathbf{H}(t)\mathbf{P}_f(t)\mathbf{H}^{\mathrm{T}}(t) + \mathbf{R}(t)]^{-1} \tag{6.26c}$$

$$\text{State analysis } \boldsymbol{x}_a(t) = \boldsymbol{x}_f(t) + \mathbf{K}(t)[\boldsymbol{y}(t) - \boldsymbol{H}(t)\boldsymbol{x}_f(t)] \tag{6.26d}$$

$$\text{Error covariance of the analysis } \mathbf{P}_a(t) = [\mathbf{I} - \mathbf{K}(t)\boldsymbol{H}(t)]\mathbf{P}_f(t) \tag{6.26e}$$

Similar to Eqs. 6.20 and 6.21, Eqs. 6.26c and 6.26d estimate the optimal analysis state $\boldsymbol{x}_a(t)$ by correcting the background $\boldsymbol{x}_f(t)$ based on the observation increment $\boldsymbol{y}(t) - \boldsymbol{H}(t)\boldsymbol{x}_f(t)$ that is weighted by the Kalman-gain matrix $\mathbf{K}(t)$. As noted before, the purpose of $\mathbf{K}$ is to spatially distribute the influence of the observation increment in order to correct the background at grid points in the vicinity of the observation. And, $\boldsymbol{H}$ is the forward operator that maps the state to the observations. The matrix $\mathbf{H}$ is the Jacobian matrix of $\boldsymbol{H}$, such that $\mathbf{H} = \partial\boldsymbol{H}/\partial\mathbf{x}$. Equation 6.26e updates the background error covariance to reflect the reduced uncertainty that results from the assimilation of the observations. Here, **I** is the identity matrix. Equations 6.26a and 6.26b propagate the state vector and error-covariance vector forward to the time when observations are next available. The matrix $\boldsymbol{M}$ is the nonlinear model-forecast operator that integrates forward in time from the analysis vector $\boldsymbol{x}_a(t)$ (initial conditions) to the time of the next analysis, where the model forecast $\boldsymbol{x}_f(t+1)$ will become the background. The matrix **M** is the Jacobian matrix of $\boldsymbol{M}$, where $\mathbf{M} = \partial \boldsymbol{M}/\partial \mathbf{x}$ and $\mathbf{M}^{\mathrm{T}}$ is its adjoint. The matrix **Q** is the covariance of model errors that accumulate during the update interval. Figure 6.19 shows a schematic of how Eqs. 6.26 are solved. See Kalnay (2003) and Hamill (2006) for the assumptions that are involved in the use of this method, and LeDimet and Talagrand (1986) and Lacarra and Talagrand (1988) for additional information about the mathematics.

A major benefit of the EKF relative to 3DVAR is that the forecast, or background, error-covariance matrix is explicitly advanced using the model itself, evolving during the forecast. A few of the distinctions between the EKF and 4DVAR methods are summarized as follows.

- The EKF explicitly evolves the covariance matrix, whereas the covariance evolution in 4DVAR is implicit.
- Unlike the EKF, 4DVAR assimilation is based on the assumption that the model is perfect (i.e., $\mathbf{Q} = 0$).

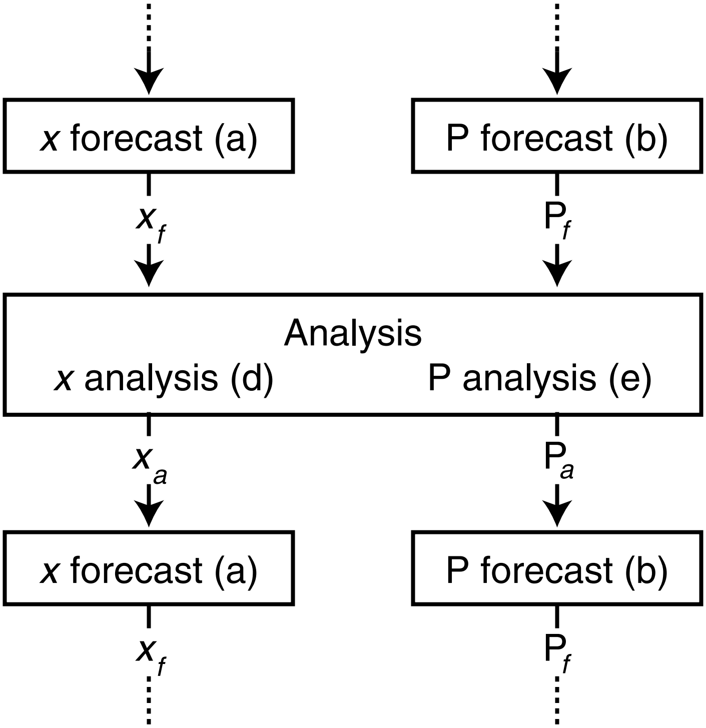


Fig. 6.19 Schematic showing the organization of the computations for the solution of Eqs. 6.26 in an EKF assimilation. The letters refer to the specific equations in the Eqs. 6.26 series.

- The 4DVAR method can be used operationally in NWP because it is computationally much cheaper than the EKF. In contrast, with current computing resources the EKF is prohibitively expensive to use for all but the very smallest modeling systems.
- The 4DVAR method simultaneously uses all the observations within the update interval, whereas the EKF is a sequential method that assimilates observations that are grouped at the update times. The former situation is preferable.

An additional difficulty with the EKF is the fact that the accuracy of the assimilation depends greatly on the quality of the determination of **Q**, but this is especially difficult to estimate.

6.11.3 Ensemble Kalman filtering

Ensemble-based data-assimilation methods are sequential in the sense that there is an ensemble of parallel short forecast and analysis steps. With the Ensemble Kalman Filter (EnKF) method, an ensemble of analyses is produced for a given time using (1) backgrounds that are produced by an ensemble of forecasts and (2) observations that have been perturbed by the addition of random noise that is drawn from a distribution of observation errors. The next ensemble of backgrounds is produced by running a short forecast from each of the members of the ensemble of analyses. The analyses are produced using the Kalman filter method described earlier, where each background is updated with a slightly different realization of the observations because of the addition of errors. For each cycle, the ensemble of backgrounds provides an estimate of the covariance matrix of the background error, and the ensemble of analyses allows the calculation of the covariance matrix of the analysis error. Specifically,

$$\bar{\boldsymbol{x}}_f = \frac{1}{K}\sum_{i=1}^{K}\boldsymbol{x}_{f,i} \tag{6.27}$$

defines the ensemble mean of the forecast state vector, where K is the number of ensemble members, and the following represents the covariance of the sample of $\boldsymbol{x}_f$,

$$\hat{\mathbf{P}}_f = \frac{1}{K-1}\sum_{i=1}^{K}(\boldsymbol{x}_{f,i}-\bar{\boldsymbol{x}}_f)(\boldsymbol{x}_{f,i}-\bar{\boldsymbol{x}}_f)^T, \tag{6.28}$$

where $\hat{\mathbf{P}}_f$ is an estimate of $\mathbf{P}_f$ from a finite ensemble. Similar to the state analysis in Eq. 6.26d,

$$\boldsymbol{x}_a(t) = \boldsymbol{x}_f(t) + \hat{\mathbf{K}}(t)[\boldsymbol{y}_i(t) - \boldsymbol{H}(t)\boldsymbol{x}_f(t)],$$

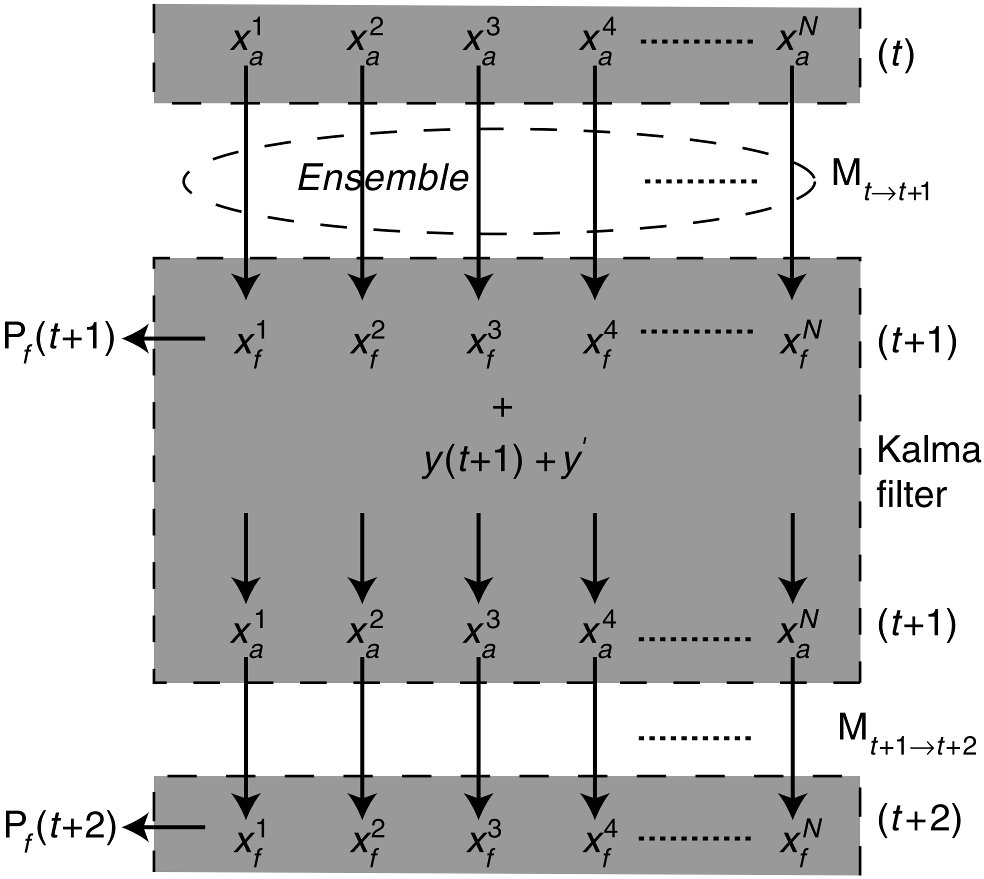
where $\boldsymbol{y}_i = \boldsymbol{y} + \boldsymbol{y}_i'$ are perturbed observations, and the Kalman gain matrix (Eq. 6.26c) is now

$$\hat{\mathbf{K}}(t) = \hat{\mathbf{P}}_f(t)\mathbf{H}^{\mathrm{T}}(t)[\mathbf{H}(t)\hat{\mathbf{P}}_f(t)\mathbf{H}^{\mathrm{T}}(t) + \mathbf{R}(t)]^{-1}. \tag{6.29}$$

Methods for simplifying and parallelizing the application of the EnKF are summarized in Hamill (2006), and include efficient ways of obtaining the Kalman gain without explicitly computing the background-error covariance matrix.

Figure 6.20 shows a schematic of the EnKF process, where time progresses downward from the top. An ensemble of analyses serves as the initial conditions for an ensemble of forecasts, where the forecast length is consistent with what is typically used for sequential initialization methods – e.g., 6 h. The ensemble of forecasts of $\boldsymbol{x}$ at $t+1$ is used to calculate the forecast, or background, error covariance using Eqs. 6.27 and 6.28. The resulting vector is then used in the Kalman gain calculation in Eq. 6.29, and this gain matrix is applied to the analysis increment to obtain the optimal correction to $\boldsymbol{x}_a$ in order to obtain $\boldsymbol{x}_f$. After another ensemble forecast, the Kalman filter is applied again, and the process continues.

The EnKF method unifies data assimilation and ensemble forecasting. Using a Monte Carlo approach, the ensemble of forecasts provides a sample of the relationship between observations and state variables, from which the forecast-error covariance can be calculated. The statistics derived from this sample serve as the basis for an ensemble of analysis from which the next ensemble of forecasts is initiated. One of the strengths of the EnKF is



x_a^1 x_a^2 x_a^3 x_a^4 .... x_a^N
(t)
Ensemble ....
$M_{t\to t+1}$
$P_f(t+1)$
x_f^1 x_f^2 x_f^3 x_f^4 .... x_f^N
(t+1)
+
$y(t+1) + y'$
Kalman filter
x_a^1 x_a^2 x_a^3 x_a^4 .... x_a^N
(t+1)
....
$M_{t+1\to t+2}$
$P_f(t+2)$
x_f^1 x_f^2 x_f^3 x_f^4 .... x_f^N
(t+2)

Fig. 6.20 Schematic showing the basis of the EnKF data-assimilation method. See the text for discussion.

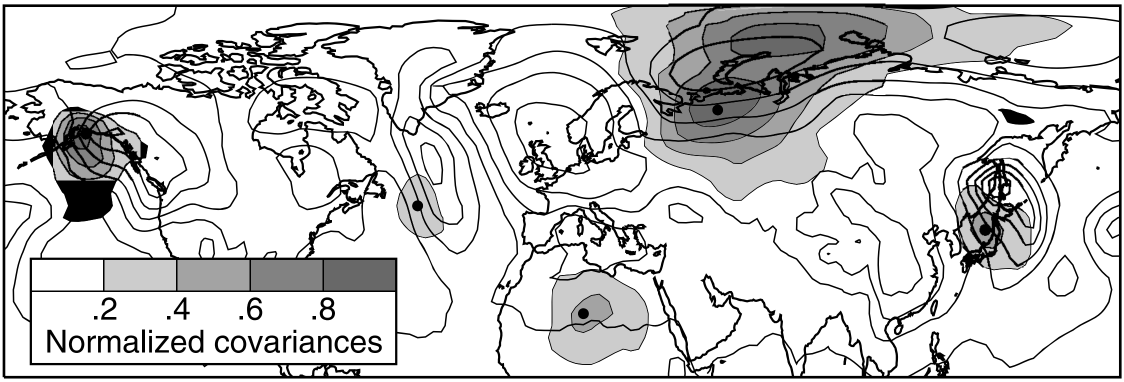


Fig. 6.21 An example of the background-error covariance (gray shading) of sea-level pressure (solid lines) around the five indicated observation points (black dots). Adapted from Hamill (2006), based on experiments conducted by Whitaker *et al.* (2004).

the fact that the background-error covariance varies with location and time. Figure 6.21 shows an example of the covariance field for a particular case from a global-data-assimilation system. The covariance patterns are shown relative to five different observation locations in the Northern Hemisphere, based on a 100-member ensemble. The background-error magnitudes and patterns vary from one region to another, with greater values surrounding the grid points in northern Russia and to the south of Alaska. The pattern around the former location is especially complex.

The 4DVAR and EnKF methods are equivalently demanding computationally, but the EnKF approach has the advantage that adjoint and tangent-linear models are not required. And, where ensemble simulations are used operationally for reasons aside from data assimilation, the added computational burden associated with the EnKF approach to data assimilation is not especially great. See Lorenc (2003) for a further comparison of these two methods. Additional information about the EnKF method can be found in Burgers *et al.* (1998), Houtekamer and Mitchell (1998, 1999, 2001), Hamill and Snyder (2000), Keppenne (2000), Mitchell and Houtekamer (2000), Hamill *et al.* (2001), Heemink *et al.* (2001), Keppenne and Rienecker (2002), Mitchell *et al.* (2002), Anderson (2003), Evensen (2003, 2007), Lorenc (2003), Snyder and Zhang (2003), Houtekamer *et al.* (2005), Hamill (2006), and Zheng (2009). For recent practical tests of the EnKF method in realistic settings, see Fujita *et al.* (2007), Bonavita *et al.* (2008), Meng and Zhang (2008), Torn and Hakim (2008), and Houtekamer *et al.* (2009).

6.12 Hybrid data-assimilation methods

There are quite a few hybrid data-assimilation approaches, drawn from the aforementioned methods, that have had historically distinct development paths. For example, Hamill and Snyder (2000) suggest a hybrid of the EnKF and 3DVAR methods where the background-error covariance is a linear combination of the typically constant, isotropic, and homogeneous 3DVAR covariance (see Section 6.8) and the variable EnKF covariance. Specifically,

$$\mathbf{P}_f^{hybrid} = (1-\alpha)\mathbf{P}_f + \alpha\mathbf{B},$$

where α is a tunable parameter that varies from 0.0 to 1.0. The objective here is to compensate for the relatively small sample of ensemble members that are used to calculate $\mathbf{P}_f$ by also incorporating information that is represented in **B**. Hamill and Snyder (2000) obtained the best results for $0.1 < \alpha < 0.4$.

Another hybrid method combines Newtonian relaxation and the adjoint of 4DVAR, the two approaches that are able to assimilate asynoptic observations at the actual measurement time. Recall that the adjoint equations compute the gradient of a cost function with respect to a control variable. In applications of the adjoint method for model initialization, such as in the 3DVAR and 4DVAR methods, the control variable is the model initial state. For model-parameter estimation, a vector of model parameters is the control variable. Examples of this hybrid approach include Zou *et al.* (1992) and Stauffer and Bao (1993), who employed the adjoint method for an optimization of analysis-nudging coefficients, which were the control variables.

Lastly, the 4DVAR and EnKF methods have been combined by Zhang *et al.* (2009). Using a somewhat idealized experimental setting, the EnKF-4DVAR system outperformed both the EnKF and 4DVAR methods.

6.13 Initialization with idealized conditions

For a variety of reasons, it is useful to be able to perform model simulations based on idealized (or synthetic) initial conditions. Even though this is discussed in Chapter 10, in the context of experimental designs of modeling studies, it will be mentioned here as well. The term idealized means that the initial conditions are not based on observations, but rather on a conceptual model of an atmospheric state. In general, this state is described by an analytic function that represents the dependent variables, which are defined at grid points. The general motivation for the use of synthetic initial conditions is that a single process or phenomenon can be isolated, in contrast to real-data initializations where there exist the inevitable complexities of the real atmosphere with many processes and scales. Specific purposes for using synthetic initial conditions include the following.

- *Instruction* – It is straightforward to demonstrate the effects of model numerics on a known solution. An example is a comparison of the correct phase speed with that produced by a given model solver. Or, simple experiments can be performed, for example of the geostrophic-adjustment process.
- *Testing a model for the existence of code errors* – There are some simple phenomena for which analytic solutions exist, and the model solutions can be compared with them to

assess model performance. Or a new model configuration can be tested against the solution that results from an older well-tested version of the model.

- *Dynamic-solver evaluation* – The same simple case can be run for many space and time scales.

Some modeling systems include software that allows the user to run a variety of preconstructed, idealized test cases. For example, the WRF system includes the following cases: flow over a bell-shaped mountain, a two-dimensional squall line, a three-dimensional supercell thunderstorm, a three-dimensional baroclinic wave, a two-dimensional gravity wave, a three-dimensional large-eddy-simulation case, and a two-dimensional sea breeze. See Chuang and Sousounis (2000) for an example of how idealized initial conditions can be implemented in a LAM.

SUGGESTED GENERAL REFERENCES FOR FURTHER READING

Cohn, S. E. (1997). An introduction to estimation theory. *J. Meteor. Soc. Japan*, **75**, 257–288.

Daley, R. (1991). *Atmospheric Data Analysis*. Cambridge, UK: Cambridge University Press.

Daley, R. (1997). Atmospheric data assimilation. *J. Meteor. Soc. Japan*, **75**, 319–329.

Evensen, G. (2007). *Data Assimilation: The Ensemble Kalman Filter*. Berlin, Germany: Springer.

Hamill, T. M. (2006). Ensemble-based atmospheric data assimilation. In *Predictability of Weather and Climate*, T. Palmer and R. Hagedorn (eds.). Cambridge, UK: Cambridge University Press.

Kalnay, E. (2003). *Atmospheric Modeling, Data Assimilation and Predictability*. Cambridge, UK: Cambridge University Press.

Lewis, J. M., A. S. Lakshmivarahan, and S. K. Dhall (2006). *Dynamic Data Assimilation: A Least Squares Approach*. Cambridge, UK: Cambridge University Press.

Talagrand, O. (1997). Assimilation of observations, an introduction. *J. Meteor. Soc. Japan*, **75**, Special Issue 1B, 191–209.

PROBLEMS AND EXERCISES

1. What processes can damp inertia–gravity waves that are generated by a model initialization?
2. What variables, in addition to the second time derivative of the surface pressure, might be used to diagnose the intensity of inertia–gravity waves associated with imbalances in the initial conditions?
3. Describe the relationship between the need for spectral nudging and the domain-size, in relaxation-based continuous data-assimilation systems.
4. Explain how a bad meteorological observation can have its negative influence spread over a large area of the computational domain.
5. Define what is meant by the expression Monte Carlo approach.

6. Conceptually and mathematically relate the SC analysis method (Section 6.6) to the statistical approaches to data assimilation (Section 6.5).
7. Explain possible ways in which near-surface observations can be used to infer model initial conditions for the boundary layer and lower troposphere. How could vertical mixing during the early part of the model forecast cause the value of the observations to be lost?
8. Derive Eqs. 6.9 and 6.10.