1
Data Assimilation: General Background

1.1 Introduction

Data assimilation includes two main components: simulation model and data. The simulation model is defined as a mathematical/numerical system that can simulate an event or a process. In most typical settings the simulation model is a prediction model based on partial differential equations (PDEs) that often includes empirical parameters. Data are generally associated with observations made by a measuring instrument, although data could also imply a product obtained by processing observations. Using an example from meteorology, data include observations such as atmospheric temperature and satellite radiances. The goal of data assimilation is to combine the information from a simulation model and data in order to improve the knowledge of the system, described by the simulation model. Apparently, the formulation of data assimilation will depend on interpretation of the *knowledge of the system*. Before we attempt to clarify a possible interpretation, it is useful to further understand the simulation model and data.

In agreement with common applications in geosciences and engineering, we narrow our discussion to a dynamic-stochastic PDE-based prediction model. Prediction models are developed with the general idea of improving the prediction of various phenomena of interest. From the theory of PDEs it is known that various parameters can impact the result of PDE integration, such as initial conditions (ICs), model errors (MEs), and empirical model parameters (EMPs). It is widely recognized that our knowledge of these parameters is never perfect, implying uncertainty of these parameters and uncertainty of the prediction calculated using such uncertain parameters.

Since the ultimate goal of using prediction models is to produce an improved prediction, it is natural to prefer a prediction that is in some way optimal. Such a prediction should be reliable, implying a desire to have a very small uncertainty associated with prediction. Then, the question is: How can the prediction be improved? First, it is anticipated that by improving the mentioned parameters (ICs, MEs, EMPs) and reducing their uncertainty would result in a desirable prediction. One could also try to improve model equations by including missing physical processes, coupling relevant components, and/or improving spatiotemporal resolution (if the prediction model is discretized). However, the only way to improve prediction is to introduce new information about the model parameters or model equations. The new information could come from another model with superior performance, but the most common source of new information about the real world comes

https://doi.org/10.1017/9781108924238.003 Published online by Cambridge University Press

from observations. An additional source of information could be introduced from past model performances if it is believed that the prediction model has some skill. If the prediction model has no skill, then observations are the only source of information, and one has to rely on using purely statistical methods. If the prediction model has some skill, however, then it is possible to combine the information from observations and from past model performances and then rely on using data assimilation.

Note that all sources of information, from observations and from prediction models, are uncertain. We already suggested that imperfect knowledge of model parameters (ICs, MEs, and/or EMPs), as well as model equations, implies an imperfect prediction. Information from observation is also not perfect. There are instrument errors, transmission errors, local errors, as well as the so-called representativeness errors. The instrument error is associated with every measuring instrument and can vary depending on the accuracy of the instrument. The errors created during a transmission from observation site to central location may not be detected in some instances and will contribute to observation error. Local errors refer to unforeseen errors of the local observation site, such as artificial heat sources and the impact of local vegetation. The representativeness error is the error caused by model prediction that is not representative of the actual observation. This can refer to inadequate model resolution, volume-averaged model variable versus point observation, etc. Therefore, observations also have errors, i.e., uncertainties.

Given that the two main components of data assimilation, prediction model and data, are inherently uncertain, then the output of data assimilation, the knowledge of the system, is expected to be uncertain as well. Uncertainty can be measured in many different ways. One can think of uncertainty as a measure of the difference between an estimate and the truth, if the truth is known. Unfortunately, the true value of the field is rarely known, except in a controlled experiments such as an observation system simulation experiment (OSSE). The theory of probability offers a mathematically consistent, formal way of dealing with uncertainties, and is used in our approach to data assimilation. A comprehensive object that describes the probabilistic system is the probability density function (PDF). Therefore, one can think of the PDF as the actual knowledge of the system, implying that the ultimate goal of data assimilation is to estimate the PDF. As will be shown in Chapters 3, 7, 8, 9, and 12, estimating the PDF is quite a challenging problem in realistic high-dimensional applications of data assimilation, mostly limiting practical data assimilation to estimating the first PDF moment (e.g., mean) and eventually the second PDF moment (e.g., covariance), with only an occasional capability of estimating the higher-order PDF moments.

Another critical aspect of data assimilation is the processing of information. Both prior model realizations and data contain information that can potentially contribute to improving the state of knowledge. Shannon's information theory (Shannon and Weaver, 1949), also based on using the probabilistic approach, offers the mathematical formalism for quantifying and processing information. Although still not used to its maximum, this information theory is a very handy tool for data assimilation. Implied from the above discussion of the impact of model parameters, such as ICs, MEs, and EMPs, on the prediction made by the model, and the aspiration of data assimilation to improve prediction by modifying model parameters ICs, MEs, and/or EMPs, the control theory is also an important tool of

https://doi.org/10.1017/9781108924238.003 Published online by Cambridge University Press

data assimilation. The implied dynamic-stochastic characterization of a prediction model also implies the important role of statistics and possibly chaotic nonlinear dynamics in data assimilation. Given that data assimilation is typically multivariate and applied to vectors and matrices, it relies heavily on using linear algebra and functional analysis.

There are several other considerations that are important for data assimilation. Realistic physical phenomena and processes, and their relation to observed variables, are all inherently nonlinear. As such, the treatment of nonlinearity in data assimilation plays an important role in choosing the adequate control theory methods and limiting the utility of linear algebra. The dynamical aspect of prediction models, generally characterized by time-dependent phenomena, implies that prediction uncertainties have to be dynamical and time-dependent as well. Given the sensitivity of PDEs to the initial (and boundary) conditions, data assimilation has to provide dynamically balanced ICs that would not cause spurious perturbations in prediction. In the case of chaotic nonlinear dynamics, as most realistic dynamical systems are, data assimilation needs to capture and eventually remove the errors of growing and neutral modes from the ICs.

With all these components, probability theory, statistics, information theory, control theory, linear algebra, and functional analysis, make data assimilation very complex and challenging.

1.2 Historical Background

First attempts to address what we now call data assimilation could be traced to data fitting and regression analysis applied in astronomy, most notably by Legendre (1805) and Gauss (1809). In solving the problem Gauss assumed normally distributed errors and introduced the normal probability distribution. Around that time Laplace (1814) introduced the Bayesian approach by developing a mathematical system on inductive reasoning based on probability. Starting with these discoveries, and after a considerable development of mathematical tools and theories, the modern-age data assimilation was made possible.

Early methods for data assimilation were deterministic and essentially represented a function fitting to measurements. This included the interpolation methods with distance-based interpolation weights in order to determine the relative importance of observations, such as the objective analysis schemes of Bergthórsson and Döös (1955), Gilchrist and Cressman (1954), Cressman (1959), and Barnes (1964). While useful for operational numerical weather prediction (NWP) of that time, these methods did not explicitly include probabilistic considerations. Other deterministic methods include nudging data assimilation (Hoke and Anthes, 1976; Davies and Turner, 1977), sometimes also referred to as four-dimensional data assimilation (4DDA) or a dynamic relaxation method. Later developments of the method include a generalization to accept uncertainties (e.g., Zou et al., 1992). Nudging implies a change of the original dynamical equations of a prediction model to include a forcing term. The coefficients associated with the forcing are generally determined by fitting the model state closer to the observations. Although nudging has been improved to implicitly accept uncertainties, it does not rely on using the Bayesian approach and does not attempt to estimate PDF moments as probabilistic data assimilation does.

https://doi.org/10.1017/9781108924238.003 Published online by Cambridge University Press

Probably the first data assimilation method that is critically relevant for understanding modern-age data assimilation is the Kalman filter (KF) (Kalman and Bucy, 1961), initially developed for signal processing. It provides a mathematically consistent methodology based on probability and Bayesian principles that produces a minimum variance solution. The KF is also helpful in describing the role of dynamics in forecast error covariance, as well in model error covariance. Since the KF is defined for linear systems, it fully resolves the Gaussian PDF and in that sense represents a satisfactory solution to general probabilistic data assimilation problems. There are, however, major obstacles in making the KF a practical data assimilation method. For one, it is a linear filtering method and as such it cannot satisfactorily address nonlinearities in the prediction model and observations. Another major obstacle is the required matrix inversion, which becomes practically impossible to calculate in realistic high-dimensional applications. Strictly relying on the Gaussian PDF assumption is also a disadvantage of the KF, given that prediction model variables and observations could have non-Gaussian errors.

The first practical method that incorporates the basic data assimilation setup with Bayesian and probabilistic assumptions is the optimal interpolation (OI) method of Gandin (1963), sometimes referred to as statistical interpolation. This is a minimum variance estimator and as such it can be related to the KF and other probabilistic data assimilation methods. A more detailed overview of OI can be found in Daley (1991) (see chapters 3, 4, and 5 therein). The OI method is very much a simplified version of the KF. The OI employs a linear observation operator, in early versions only the identity matrix. For nonlinear observations, such as satellite radiances, an inversion algorithm (i.e., retrieval) that produces a model variable from observations is required. The forecast error covariance is modeled and includes separate vertical and horizontal correlations. By construction the forecast error, covariance is homogeneous (i.e., all grid points are treated equally) and isotropic (all directions are treated equally). In addition, the covariance is stationary, being approximated by a correlation function with statistically estimated correlation parameters. Since it is related to the KF, OI can also produce an estimate of the posterior error covariance. However, such an estimate is not reliable since the input covariances and parameters are not accurate. The OI is also local, in the sense that only observations within a certain distance from the model point impact the analysis at that point. Although theoretically and practically an important step in probabilistic data assimilation development, when measured against our motivation to produce a reliable estimate of PDFs, OI leaves much to be desired. At best it can produce a meaningful estimate of the first PDF moment only, however with serious limitations related to preferred capabilities such as the nonlinearity of observation operators and dynamical structure of forecast error covariance.

Another fundamental development that led to current variational data assimilation (VAR) methods was the introduction of variational principles in data assimilation by Sasaki (1958). While at the time it was understood as a method for objective analysis based on least squares, the new method for the first time introduced variational formalism and minimization under the geostrophic constraint and also under the more general balance constraint between winds and geopotential. Then, in a trilogy of papers (Sasaki, 1970a, 1970b, 1970c) expanded

https://doi.org/10.1017/9781108924238.003 Published online by Cambridge University Press

the previous approach to include the time dependency of observations and established a basis for future development of four-dimensional variational data assimilation (4DVAR) methodology.

While the use of variational principles in data assimilation have been known since the early work of Sasaki (1958), it took almost 25 years before variational methodology had another push into the field of data assimilation, mostly because of the advancements of computers in NWP. Addressing the deficiencies of OI, most importantly the local character of the analysis, nonlinearity of observations, and to some extent the specification of forecast error covariance, variational methods for data assimilation were revived in the mid 1980s (e.g., Lewis and Derber, 1985; Le Dimet and Talagrand, 1986). The subtypes of variational methods include three-dimensional variational (3DVAR) (e.g., Parrish and Derber, 1992) and 4DVAR data assimilation (e.g., Navon et al., 1992; Županski, 1993; Courtier et al., 1994). They include a global minimization (i.e., over all model points) of the cost function that can incorporate nonlinear observations and solves the inversion problem using adjoint equations. The forecast error covariance is improved over OI as it includes complex cross-correlations with additional dynamical balance constraints, but the correlations are still modeled. On the positive side, the modeling of error covariance allows the covariance to be of full rank, meaning that all degrees of freedom (DOF) required for solving the analysis problem are included. The covariance is stationary, although in 4DVAR there is limited capability to introduce time dependence during the assimilation window. Also, variational methods primarily estimate the first moment of PDFs. Although it is possible to estimate the second PDF moment, especially in 4DVAR, there is no feedback of uncertainties from one data assimilation cycle to the next implying a limited use of Bayesian inference. The main advantage of variational methods is their capability to assimilate nonlinear observations, in particular the satellite radiances that now represent the major source of information in meteorology (e.g., Derber and Wu, 1998). By introducing 4DVAR the prediction model itself could be used as a constraint in optimization. The cost of applying VAR has increased compared to previous methods, but it can still be considered efficient since potentially costly matrix inversions are avoided. The variational methods are still used in practice.

Immediately following this development of variational methods, ensemble Kalman filtering (EnKF) methods have been introduced to data assimilation (Evensen, 1994; Houtekamer and Mitchell, 1998). The EnKF successfully addressed the problem of the nonlinear prediction model in the KF by introducing the Monte Carlo approach to the KF forecast step. At the same time the forecast error covariance is dynamic, and is therefore an improvement on the stationary and modeled error covariance used in variational methods. The most important impact of the EnKF was that a realistic data assimilation could be used to produce the first two moments of the PDF, the mean and the covariance, although still under a Gaussian PDF assumption. One of the issues of the EnKF is not being able to account for nonlinearity of observations, since the same linear KF analysis equation is used. More recently (e.g., Sakov et al., 2012), an iterative EnKF was introduced in a manner similar to the iterated KF to address the nonlinearity of observation operators. Implementing the EnKF requires the assimilation of perturbed observations, which results in the calculation of numerous analyses for each ensemble member, and therefore an increase in the cost. Square-root

https://doi.org/10.1017/9781108924238.003 Published online by Cambridge University Press

EnKF methods were introduced to reduce the computational cost, by directly calculating the mean of the analysis. Including a large number of ensembles required to resolve a realistic data assimilation problem proved to be practically impossible due to the computational cost and storage requirements. This prompted a need for covariance localization to increase the number of DOF of the low-rank ensemble covariance that could be feasible. This localization greatly helped the EnKF and related ensemble methods to remain of practical significance, although a modification of the dynamically based ensemble forecast error covariance via convolution with a prespecified localizing covariance matrix is required. Covariance localization implies that practical EnKF methods can be interpreted, in terms of forecast error covariance, as an intermediate approach between the full-rank EnKF (with all DOF) and the local OI method. The analysis solution in the EnKF with localization is essentially local since only observations within a certain distance can impact the analysis point.

Both EnKF and variational methods have practical and theoretical limitations. Variational methods have the capability of addressing nonlinearity through applying global numerical optimization. The EnKF is inherently designed to use the linear analysis equation of the KF. An alternative way of bridging this issue was introduced by the maximum likelihood ensemble filter (MLEF) (Županski, 2005), in which it was shown how the calculation of adjoint operators could be avoided by using nonlinear ensemble perturbations and applied in variational-like minimization of the cost function. As with other ensemble methods, MLEF includes the flow-dependent ensemble covariance and estimates the posterior uncertainty.

The implied limitation of error covariance representation in the practical EnKF due to an insufficient number of DOF and to some extent the nondynamical impact of covariance localization, even though the covariance is flow dependent, can result in an analysis that is not of the desired quality. The same could be said for variational methods, where the use of stationary and modeled error covariance is not sufficiently realistic and can produce unsatisfactory analysis. As a result, hybrid ensemble-variational methods that allow a combination of the flow-dependent ensemble, but low-rank covariance, and the stationary variational, but full-rank, error covariance were introduced (Lorenc, 2003; Buehner, 2005; Wang et al., 2007; Bonavita et al., 2012; Clayton et al., 2013).

Data assimilation can also be viewed as an application of Pontryagin's minimum principle (PMP) (e.g., Pontryagin et al., 1961; Lakshmivarahan et al., 2013) where a least squares fit of an idealized path to dynamics law follows from Hamiltonian mechanics. In this application of optimal control theory, the problem is posed as finding the best possible forcing for taking a dynamical system from one state to another, in the presence of dynamical constraints. This forcing is also related to accounting for ME in data assimilation. While the use of forcing reminds us of nudging, the PMP method is more general since it includes an optimization subject to dynamical constraints as well as uncertainties (Lakshmivarahan and Lewis, 2013). Similar to previous methods, it searches for optimal analysis that could be interpreted as the first PDF moment, but estimation of the posterior uncertainties is not an essential part of the method. It is possible to view 4DVAR as a special case of PMP.

The above historical overview also indicates the current status of practical data assimilation development. Other methods with stronger theoretical foundations have been introduced to data assimilation, such as particle filters (PFs) (e.g., van Leeuwen, 2009; Chorin et al., 2010), but they still have limitations for realistic high-dimensional applications.

https://doi.org/10.1017/9781108924238.003 Published online by Cambridge University Press

However, by directly calculating arbitrary PDFs through the Bayesian framework they have a theoretical advantage in accounting for nonlinearity and non-Gaussianity and therefore offer numerous possibilities for the future development of data assimilation.

1.3 Terminologies and Notation

Data assimilation consists of two major elements – a model of the dynamical system and a set of data (i.e., observations), and aims to procure optimal estimates of model states by combining model forecasts and observations. We represent the model states and the observations in terms of vectors, $\mathbf{x}$ and $\mathbf{y}^o$, respectively. The *true* state, $\mathbf{x}^t$, can never be obtained but can be estimated through an adequate estimation procedure. Such an estimate, made at a given time, is called the *analysis*, $\mathbf{x}^a$. The *estimate* is also denoted by $\hat{\mathbf{x}}$ and is interchangeably used with $\mathbf{x}^a$. The *background*, $\mathbf{x}^b$, is an a priori estimate of $\mathbf{x}^t$ before the analysis is conducted. For the notations in data assimilation, we generally follow Ide et al. (1997).

Data assimilation represents a process to obtain $\mathbf{x}^a$, as close to $\mathbf{x}^t$ as possible, by correcting $\mathbf{x}^b$ using a correction, $\Delta\mathbf{x}$. Mathematically, it is formulated, in its simplest form, as:

$$\mathbf{x}^a = \mathbf{x}^b + \Delta\mathbf{x}. \tag{1.1}$$

Note that $\Delta\mathbf{x}$ is a function of both $\mathbf{y}^o$ and $\mathbf{x}^b$, and it is called an *analysis increment*.

1.3.1 Observation Equation

A variety of observations, assembled in $\mathbf{y}^o$, are used for data assimilation (see Figure 1.1). As observations are much fewer than model states and are irregularly distributed, direct comparison between observations and model states is unfeasible. Thus, we define a nonlinear function, H, called an *observation operator*, that transforms the state vector from the state space, $\mathcal{R}^m$, to the observation vector in the observation space, $\mathcal{R}^n$. The observation is described in terms of the true state as:

$$\mathbf{y}^o = H\mathbf{x}^t + \boldsymbol{\varepsilon}^o, \tag{1.2}$$

where $\boldsymbol{\varepsilon}^o$ is the observation (measurement) error. Equation (1.2) is called the *observation equation* or the *observation model*.

1.3.2 Observation Error Statistics

We assume that the measurement error $\boldsymbol{\varepsilon}^o$ in (1.2) is random and independent, and hence have zero mean, i.e.,

$$mean(\varepsilon_i^o) = E(\varepsilon_i^o) = 0 \ \text{ for } \ i = 1, \ldots, n. \tag{1.3}$$

This implies that $\mathbf{y}^o$ in (1.2) depends only on $\mathbf{x}^t$ and all other variation in $\mathbf{y}^o$ is random. For the random errors, the variance and the covariance of the errors are

$$var(\varepsilon_i^o) = E(\varepsilon_i^o \varepsilon_i^o) = \sigma_i^2 \ \text{ for } \ i = 1, \ldots, n \tag{1.4}$$

https://doi.org/10.1017/9781108924238.003 Published online by Cambridge University Press

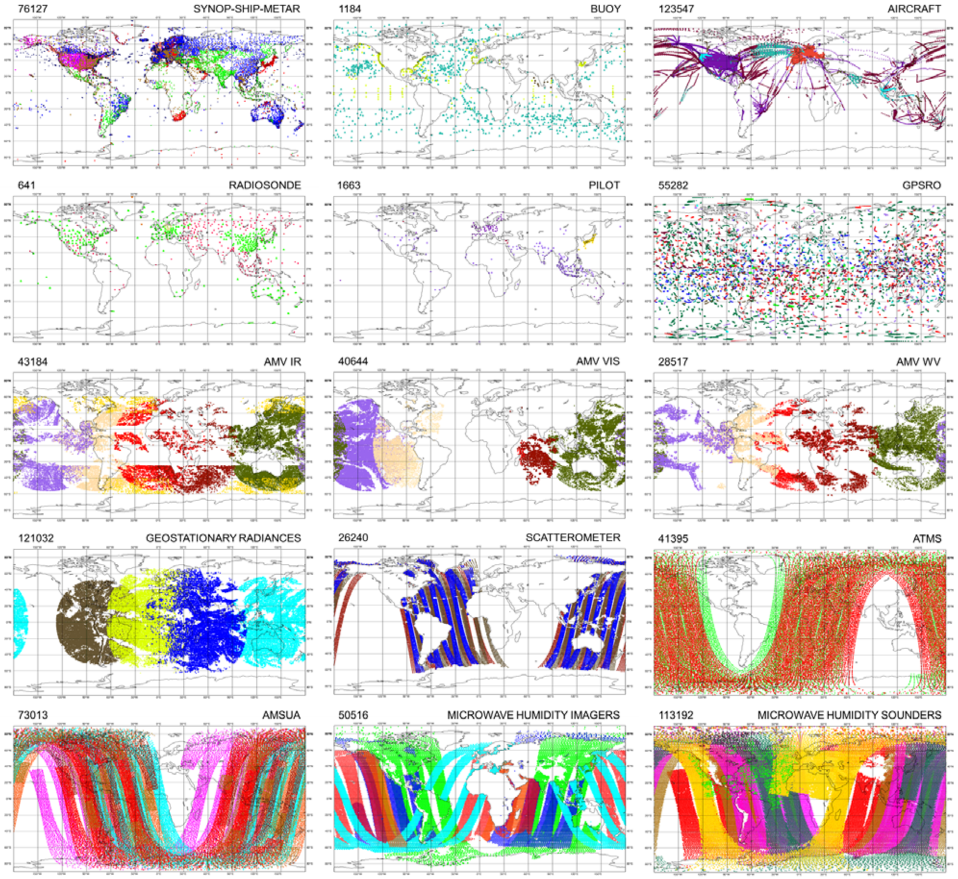


Figure 1.1 Various observation data at the global scale, available on 0000 UTC 18 August 2020, with the observation platforms (top-right corners) and the number of data used for data assimilation (top-left corners). The details of legends in the subfigures refer to the data coverage from the European Centre for Medium-Range Weather Forecasts (ECMWF, 2020). CC BY-NC-ND 4.0 License.

and

$$cov(\varepsilon_i^o, \varepsilon_j^o) = E(\varepsilon_i^o \varepsilon_j^o) = 0 \text{ for } i, j = 1, \ldots, n \text{ and } i \neq j, \tag{1.5}$$

respectively. Here, σ_i^2 is the squared standard deviation of ε_i^o and (1.4) assumes that the variance of $\boldsymbol{\varepsilon}^o$ is constant; thus, not dependent on $\mathbf{x}^t$. With zero covariances in (1.5), the variables in $\boldsymbol{\varepsilon}^o$ are uncorrelated with each other. By combining the three assumptions in (1.3)–(1.5), we have

$$\begin{aligned} mean(\boldsymbol{\varepsilon}^o) &= E(\boldsymbol{\varepsilon}^o) = \mathbf{0}, \\ cov(\boldsymbol{\varepsilon}^o) &= E(\boldsymbol{\varepsilon}^o(\boldsymbol{\varepsilon}^o)^T) = \sigma^2\mathbf{I} = \mathbf{R}. \end{aligned} \tag{1.6}$$

That is, the observation error covariance **R** is a diagonal matrix composed of σ_i^2 though **R** is in general a nondiagonal matrix.

1.3.3 Observation Operator

Note that it is not only that the observation sites are not usually located at the grid points where model states are calculated, but also that the observation quantities often do not match the model variables. For instance, some remotely sensed observations do not have corresponding model states; thus, it is essential to convert the model state variables into observation variables. In practice, the operator H is performed in two steps:

1. *Interpolation*, say H^I, from the model grid points to the observation sites where the conversion will be performed for indirect observations or directly when the state variables are the same as the observation quantities.
2. *Conversion*, say H^C, of the model variables to the observables when the measurements are indirect, e.g., radiances measured by sensors onboard satellites. A radiative transfer model can serve as an H to calculate the radiance, using the whole state vector components, at a specific waveband (see, e.g., Figure 1.2).

The operation by $H : \mathcal{R}^m \to \mathcal{R}^n$ is composed of two mappings, that is, $H^I : \mathcal{R}^m \to \mathcal{R}^n$ and $H^C : \mathcal{R}^n \to \mathcal{R}^n$, giving

$$\hat{\mathbf{y}} = H\left(\hat{\mathbf{x}}\right) = H^C\left(\mathbf{x}^o\right) = H^C\left(H^I\left(\hat{\mathbf{x}}\right)\right) = H^C H^I\left(\hat{\mathbf{x}}\right). \tag{1.7}$$

This decomposition is illustrated in Figure 1.3. Here, $H^I(\hat{\mathbf{x}})$ interpolates an $m \times 1$ state estimate vector $\hat{\mathbf{x}}$ from the model space (i.e., grid points) to an $n \times 1$ vector $\mathbf{x}^o$ in the observation space. $H^C(\mathbf{x}^o)$ converts n state variables $\mathbf{x}^o$ into a set of n observations $(\hat{\mathbf{y}})$ – that is, observation estimate (or model equivalents of observations) – the radiance computed

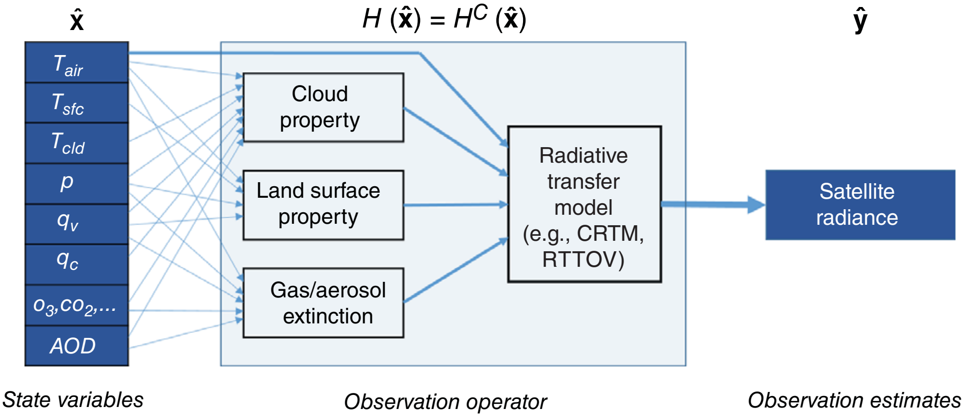


Figure 1.2 Composition of observation operator, H.

https://doi.org/10.1017/9781108924238.003 Published online by Cambridge University Press

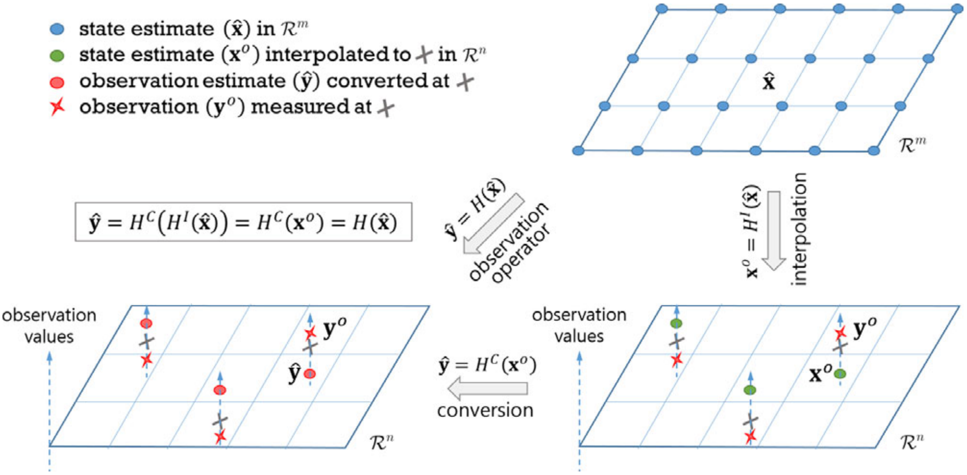

● state estimate ($\hat{\mathbf{x}}$) in $\mathcal{R}^m$
● state estimate ($\mathbf{x}^o$) interpolated to + in $\mathcal{R}^n$
● observation estimate ($\hat{\mathbf{y}}$) converted at +
observation ($\mathbf{y}^o$) measured at +
$\hat{\mathbf{x}}$
$\mathcal{R}^m$
$\hat{\mathbf{y}} = H^C(H^I(\hat{\mathbf{x}})) = H^C(\mathbf{x}^o) = H(\hat{\mathbf{x}})$
$\hat{\mathbf{y}} = H(\hat{\mathbf{x}})$
observation operator
$\mathbf{x}^o = H^I(\hat{\mathbf{x}})$
interpolation
observation values
$\mathbf{y}^o$
$\hat{\mathbf{y}}$
$\mathcal{R}^n$
$\hat{\mathbf{y}} = H^C(\mathbf{x}^o)$
conversion
observation values
$\mathbf{y}^o$
$\mathbf{x}^o$
$\mathcal{R}^n$

Figure 1.3 Composition of observation operator, H.

through a radiance transfer model using temperature, water vapor mixing ratio, cloud mixing ratio, etc. (see Figure 1.2). If the model states and the observation quantities were the same, $H(\hat{\mathbf{x}}) = H^I(\hat{\mathbf{x}})$: If the observation sites were exactly located at the grid points and the model states and observation quantities were different, $H(\hat{\mathbf{x}}) = H^C(\hat{\mathbf{x}})$.

The observation operator H is generally nonlinear; however, it is more convenient to use its linearized version, denoted by $\mathbf{H}$, in explaining the concept of data assimilation. For practical applications, in (1.7), we adopt linear operators, i.e., $H = \mathbf{H}$, for direct observations that match the model variables (wind, temperature, humidity, etc.): We employ the nonlinear operator H for indirect observations (satellite radiance, radar reflectivity, etc.).

Example 1.1 H for vertical sounding of moisture

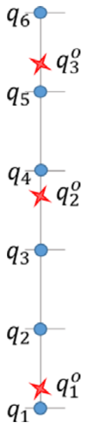
Assume that measurements of humidity (q) are made by a radiosonde at three levels and represented as

$$\mathbf{y}^o = (q_1^o, q_2^o, q_3^o)^T$$

while the model states are calculated at six levels and depicted as

$$\mathbf{x} = (q_1, \ldots, q_6)^T.$$

As in the figure, the levels of measurements, z_l^o ($l = 1,2,3$), are not necessarily the same as those of model states, z_k ($k = 1, \ldots, 6$). Because the observed quantity and the model state are the same, we can simply put $\mathbf{x}^o = \mathbf{x}$. We apply the observation operator $H = H^I$ to interpolate $\mathbf{x}^o$ to the observation space (i.e., levels l):



https://doi.org/10.1017/9781108924238.003 Published online by Cambridge University Press

$$H\left(\mathbf{x}^o\right) = \begin{pmatrix} q_1 + \frac{q_2-q_1}{z_2-z_1}\left(z_1^o - z_1\right) \\ q_3 + \frac{q_4-q_3}{z_4-z_3}\left(z_2^o - z_3\right) \\ q_5 + \frac{q_6-q_5}{z_6-z_5}\left(z_3^o - z_5\right) \end{pmatrix} = \begin{pmatrix} \beta_1 q_1 + \beta_2 q_2 \\ \beta_3 q_3 + \beta_4 q_4 \\ \beta_5 q_5 + \beta_6 q_6 \end{pmatrix} = \hat{\mathbf{y}}.$$

This is a set of linear equations; thus, we can define a linear observation operator $\mathbf{H} \equiv \frac{\partial H}{\partial \mathbf{x}}$, given by

$$\mathbf{H} = \begin{pmatrix} \beta_1 & \beta_2 & 0 & 0 & 0 & 0 \\ 0 & 0 & \beta_3 & \beta_4 & 0 & 0 \\ 0 & 0 & 0 & 0 & \beta_5 & \beta_6 \end{pmatrix}.$$

Example 1.2 *H* for the radar reflectivity factor

The following relation, derived from the Marshall–Palmer distribution of raindrop size without considering ice phases, was used as an observation operator for the radar reflectivity factor (e.g., Sun and Crook, 1997; Xiao et al., 2007; Sugimoto et al., 2009):

$$Z = 43.1 + 17.5 \log\left(\rho q_r\right), \tag{1.8}$$

where Z is the reflectivity factor (in dBZ), ρ is the air density (in kg m^{-3}), and q_r is the rainwater mixing ratio (in g kg^{-1}).

By considering two ice phases – the snow and hail mixing ratios (q_s and q_h, respectively) – Gao (2017) devised the reflectivity observation operator as:

$$Z = 10 \log Z_e. \tag{1.9}$$

Here, the equivalent radar reflectivity factor Z_e is given by

$$Z_e = \begin{cases} Z(q_r) & \text{for} \quad T_b \geq 5°\text{C} \\ \alpha Z(q_r) + (1-\alpha)\left[Z(q_s) + Z(q_h)\right] & \text{for} \quad -5°\text{C} < T_b < 5°\text{C} \\ Z(q_s) + Z(q_h) & \text{for} \quad T_b \leq -5°\text{C}, \end{cases} \tag{1.10}$$

where T_b is the background temperature, and α varies linearly between 0 at $T_b = -5°$C and 1 at $T_b = 5°$C, for different components of the reflectivity factors as the following:

Phase	Reflectivity factor	References/Conditions
Rain	$Z(q_r) = 3.63 \times 10^9 (\rho q_r)^{1.75}$	Smith Jr. et al. (1975)
Snow (dry)	$Z(q_s) = 9.80 \times 10^8 (\rho q_s)^{1.75}$	$T_b < 0°$C
Snow (wet)	$Z(q_s) = 4.26 \times 10^{11} (\rho q_s)^{1.75}$	$T_b > 0°$C
Hail	$Z(q_h) = 4.33 \times 10^{10} (\rho q_h)^{1.75}$	Lin et al. (1983); Gilmore et al. (2004)

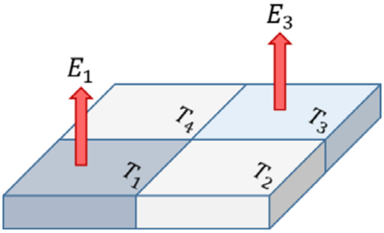
https://doi.org/10.1017/9781108924238.003 Published online by Cambridge University Press

Practice 1.1 H for irradiance

Assume that a satellite measures irradiance, E, emitted from an atmospheric layer. In a numerical model, E_i, from the ith grid box, can be simply calculated by the Stefan–Boltzmann law as:

$$E_i = \sigma T_i^4,$$

where σ is the Stefan–Boltzmann constant and T_i is the layer temperature in that grid box. Develop the observation operator H and its linearized operator $\mathbf{H}$ ($\equiv \partial H / \partial \mathbf{x}$), for the following figure, which has measurements of E from grid boxes 1 and 3 only.



1.3.4 Background Field

Observations are generally much fewer than model states that are assigned to 3D grid points per time step: the total number of conventional observations is of the order of 10^4 while that of grid-point variables to be calculated in an NWP model is of the order of 10^7 (Kalnay, 2003). This makes data assimilation an *underdetermined* problem. Furthermore, observations are distributed irregularly over the globe, e.g., there exists much poorer data over the southern hemisphere than the northern hemisphere.

Therefore, additional information on each grid point – called a *background* or *first guess* and denoted by $\mathbf{x}^b$ – is necessary to create ICs for numerical prediction. In these days, operational centers generate $\mathbf{x}^b$ using a short-range forecast (e.g., a 6-h forecast for a global model) out of an analysis cycle (e.g., a 6-h cycle in the global data assimilation system), as shown in Kalnay (2003).

1.3.5 Analysis Equation

The goal of data assimilation is to find the analysis ($\mathbf{x}^a$) by correcting the background ($\mathbf{x}^b$) using the observation ($\mathbf{y}^o$), which is formulated as

$$\mathbf{x}^a = \mathbf{x}^b + \mathbf{W}\left[\mathbf{y}^o - H\left(\mathbf{x}^b\right)\right]. \tag{1.11}$$

https://doi.org/10.1017/9781108924238.003 Published online by Cambridge University Press

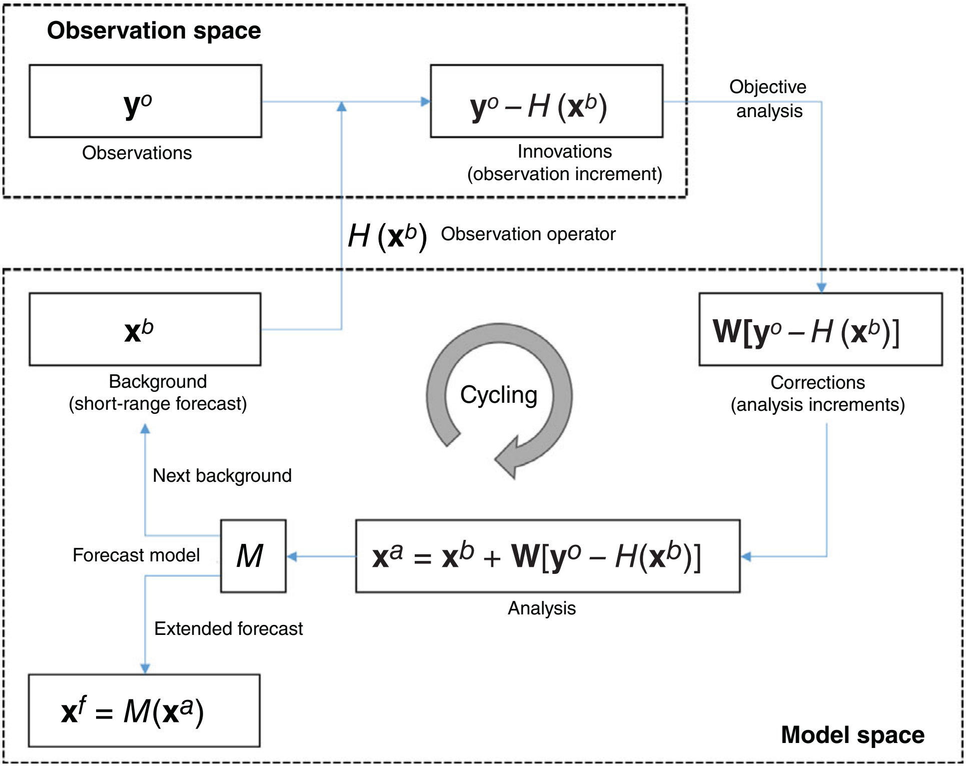


Figure 1.4 Basic concept of data assimilation.

Here, $\mathbf{y}^o - H(\mathbf{x}^b)$ is called the *observation increment* or *innovation*, representing the difference between the observations and the model states. The term $\mathbf{W}\left[\mathbf{y}^o - H(\mathbf{x}^b)\right]$ is called the *analysis increment*, where $\mathbf{W}$ is a weight – or *gain* – represented by the error characteristics (e.g., the background and observation error covariances).

Figure 1.4 describes the basic concept of data assimilation in the framework of NWP. Note that we are dealing with two separate spaces – the observation space and the model space – which are linked to each other via the observation operator H. The model prediction starts from $\mathbf{x}^b$, obtained from a short-range forecast (e.g., 6-h forecast in a global prediction system and 1-h forecast in a regional prediction system); when observations are available, $\mathbf{x}^b$ is transformed to the observation space through $H(\mathbf{x}^b)$ to calculate an innovation $\mathbf{y}^o - H(\mathbf{x}^b)$. Using the innovation, a correction term $\mathbf{W}\left[\mathbf{y}^o - H(\mathbf{x}^b)\right]$ is obtained in the model space through the objective analysis and added to $\mathbf{x}^b$ to get $\mathbf{x}^a$; then, a forward propagator M (i.e., a forecast model), operating on $\mathbf{x}^a$, produces an extended forecast and a new background for the next cycle of data assimilation. An algorithmic view of data assimilation and numerical prediction is depicted in Algorithm 1.1.

https://doi.org/10.1017/9781108924238.003 Published online by Cambridge University Press

Algorithm 1.1 General data assimilation with cycling

```
/* index $k$ denotes cycle number                                              */
/* $\mathcal{R}^m$ denotes model space; $\mathcal{R}^n$, observation space       */
***Initiation***: $\mathbf{x}^b$                                 ! Provide a background $\mathbf{x}^b$ at $k = 1$
**repeat**                                                     ! Loop for cycle $k = 1$ to $kmax$
    ***Analysis*** at cycle $k$                                  ! Procedure to obtain analysis $\mathbf{x}^a$
        ***Transformation***: $\mathbf{x}^b \longrightarrow H(\mathbf{x}^b)$              ! Operate $H: \mathcal{R}^m \to \mathcal{R}^n$
        ***Innovation***: $\mathbf{y}^o - H\left(\mathbf{x}^b\right)$                      ! Calculate innovation at $\mathcal{R}^n$
        ***Correction***: $\mathbf{W}\left[\mathbf{y}^o - H\left(\mathbf{x}^b\right)\right]$                 ! Calculate correction at $\mathcal{R}^m$
        ***Analysis***: $\mathbf{x}^a = \mathbf{x}^b - \mathbf{W}\left[\mathbf{y}^o - H\left(\mathbf{x}^b\right)\right]$      ! Obtain analysis at $\mathcal{R}^m$
    ***Forecast*** at cycle $k$                                  ! New background and forecast
        ***Background***: $\mathbf{x}^b = M(\mathbf{x}^a)\big|_{t_b}$          ! Short-range forecast (e.g., $t_b = 6$ h)
        ***Forecast***: $\mathbf{x}^f = M(\mathbf{x}^a)\big|_{t_f}$            ! Extended forecast (e.g., $t_f = 48$ h)
**until** *end of cycling*
```

1.4 Basic Estimation Problem

1.4.1 Least Squares Estimation

The least-squares approach was invented in the 1800s independently by Carl Friedrich Gauss and Adrien-Marie Legendre for calculating planetary motion. It constitutes the foundation of modern data assimilation (Sorenson, 1970; Kalnay, 2003; Lewis et al., 2006) through the core concept of estimating the unknown parameters by minimizing the squared differences between the model and the data.

From the observation model (1.2), we can express $\mathbf{y}^o$ in terms of the state estimate $\hat{\mathbf{x}}$, rather than the true state $\mathbf{x}^t$, and a linear observation operator $\mathbf{H}$ as

$$\mathbf{y}^o = \mathbf{H}\hat{\mathbf{x}} + \boldsymbol{\varepsilon}^r = \hat{\mathbf{y}} + \boldsymbol{\varepsilon}^r, \tag{1.12}$$

where $\boldsymbol{\varepsilon}^r$ is called the residual error – the difference between the *true* measurement $\mathbf{y}^o$ and the *estimated* measurement $\hat{\mathbf{y}} = \mathbf{H}\hat{\mathbf{x}}$. Equation (1.12) is also called the linear *regression* model (see Colloquy 1.1).

In the least-squares estimation, we seek to find a specific $\hat{\mathbf{x}}$ that minimizes a functional $J = J(\hat{\mathbf{x}})$, defined as the sum of squared residual errors:

$$J = (\boldsymbol{\varepsilon}^r)^T \boldsymbol{\varepsilon}^r = \|\boldsymbol{\varepsilon}^r\|_2^2 = \left(\mathbf{y}^o - \mathbf{H}\hat{\mathbf{x}}\right)^T \left(\mathbf{y}^o - \mathbf{H}\hat{\mathbf{x}}\right), \tag{1.13}$$

where $\|\boldsymbol{\varepsilon}^r\|_2 = \left((\varepsilon_1^r)^2 + \cdots + (\varepsilon_n^r)^2\right)^{1/2}$ represents the Euclidean or L^2 norm. Minimization of the quadratic function J should satisfy the following requirements, for the *gradient* $\nabla_{\hat{\mathbf{x}}} J$ and the *Hessian* $\nabla^2_{\hat{\mathbf{x}}} J$:

$$\nabla_{\hat{\mathbf{x}}} J = \frac{\partial J}{\partial \hat{\mathbf{x}}} = -2\mathbf{H}^T\mathbf{y}^o + 2\left(\mathbf{H}^T\mathbf{H}\right)\hat{\mathbf{x}} = 0 \tag{1.14}$$

and

$$\nabla^2_{\hat{\mathbf{x}}} J = \frac{\partial^2 J}{\partial \hat{\mathbf{x}} \partial \hat{\mathbf{x}}^T} = 2\left(\mathbf{H}^T\mathbf{H}\right) \text{ is positive definite.} \tag{1.15}$$

Through the necessary condition (1.14), we can calculate the minimizer $\hat{\mathbf{x}}$ – the optimal estimate that minimizes J – as the solution of the *normal* equation

$$\left(\mathbf{H}^T\mathbf{H}\right)\hat{\mathbf{x}} = \mathbf{H}^T\mathbf{y}^o. \tag{1.16}$$

When $\mathbf{H}^T\mathbf{H}$ is square $(m \times m)$ and *nonsingluar* – i.e., having its inverse, nonzero determinant, and $\mathbf{H}$ with full rank of m – we get the solution of (1.16) as

$$\hat{\mathbf{x}} = \left(\mathbf{H}^T\mathbf{H}\right)^{-1}\mathbf{H}^T\mathbf{y}^o. \tag{1.17}$$

Here, $\left(\mathbf{H}^T\mathbf{H}\right)^{-1}\mathbf{H}^T \equiv \mathbf{H}'$ is called the *generalized inverse* or *pseudoinverse* of $\mathbf{H}$. The sufficient condition (1.15) implies that, for any nonzero norm of $\mathbf{q}$ (i.e., for all $\|\mathbf{q}\| > 0$),

$$\mathbf{q}^T\left(\mathbf{H}^T\mathbf{H}\right)\mathbf{q} = (\mathbf{H}\mathbf{q})^T(\mathbf{H}\mathbf{q}) = \|\mathbf{H}\mathbf{q}\|_2^2 > 0, \tag{1.18}$$

which will hold only when $\mathbf{H}\mathbf{q} \neq 0$ for $\mathbf{q} \neq 0$: This is true when the columns of $\mathbf{H}$ are linearly independent (Lewis et al., 2006). Furthermore, the rank of $\mathbf{H}$ is m (i.e., full). These confirm that $\hat{\mathbf{x}}$ in (1.17) is the minimizer of J only when $\mathbf{H}^T\mathbf{H}$ is *positive definite*.

Practice 1.2 Least squares cost function, J

Show that the least squares cost function, J, satisfies the following equality:

$$\begin{aligned} J &= \left(\mathbf{y}^o - \mathbf{H}\hat{\mathbf{x}}\right)^T\left(\mathbf{y}^o - \mathbf{H}\hat{\mathbf{x}}\right) \\ &= (\mathbf{y}^o)^T\mathbf{y}^o - 2(\mathbf{y}^o)^T\mathbf{H}\hat{\mathbf{x}} + \hat{\mathbf{x}}^T\left(\mathbf{H}^T\mathbf{H}\right)\hat{\mathbf{x}}. \end{aligned}$$

Then, derive $\nabla_{\hat{\mathbf{x}}} J$ and $\nabla^2_{\hat{\mathbf{x}}} J$.

Colloquy 1.1

Linear regression

Linear regression – finding a line that best fits a set of data in the least-squares sense – is regarded as a useful paradigm for more complex inverse problems (e.g., atmospheric/oceanic data assimilation) (see Thacker, 1992). Assume that we use a set of data (X_i, Y_i), with a total of N pairs, to develop a function (i.e., regression model) relating the dependent variable Y (or predictand) to the independent variable X (or predictor) as in the following form:

$$Y_i = \hat{Y}_i + \varepsilon_i = aX_i + b + \varepsilon_i, \quad i = 1, \ldots, N, \tag{1.19}$$

https://doi.org/10.1017/9781108924238.003 Published online by Cambridge University Press

where a and b are the slope and the intercept, respectively, of the regression line $\hat{Y}_i$ and ε_i is the residual (or error). Finding the best fit implies estimating the values of a and b that minimize the sum of squares between the observations (Y_i) and and the model solutions $(\hat{Y}_i)$:

$$SSE = \sum_i \varepsilon_i^2 = \sum_i \left(Y_i - \hat{Y}_i\right)^2 = \sum_i (Y_i - (aX_i + b))^2, \tag{1.20}$$

where SSE stands for the error sum of squares. The following normal equations are derived by taking the derivatives of the SSE with respect to a and b and setting them to 0:

$$\frac{\partial(SSE)}{\partial a} = 0 = \sum_i 2X_i\left((aX_i + b) - Y_i\right) = 2a\sum_i X_i^2 + 2b\sum_i X_i - 2\sum_i Y_iX_i$$

$$\frac{\partial(SSE)}{\partial b} = 0 = \sum_i 2\left((aX_i + b) - Y_i\right) = 2Nb + 2a\sum_i X_i - 2\sum_i Y_i,$$

producing the least squares estimates of a and b as

$$a = \frac{\sum_i \left(Y_i - \overline{Y}\right)\left(X_i - \overline{X}\right)}{\sum_i \left(X_i - \overline{X}\right)^2}; \; b = -a\overline{X} + \overline{Y},$$

where $\overline{X}$ and $\overline{Y}$ denote the means of X and Y, respectively (see the figure in Colloquy 1.2).

With multiple predictors (i.e., X_{ik}, $k = 1, \ldots, K$, for given Y_i), we can construct a multiple linear regression, $\hat{Y}_i = b + \sum_k X_{ik}a_k$, represented in vector form as:

$$\mathbf{Y} = \mathbf{Xa} + \boldsymbol{\varepsilon},$$

where b is included in the vector $\mathbf{a}$. Then, the SSE (1.20) becomes the least squares cost function (1.13) as

$$SSE = \boldsymbol{\varepsilon}^T\boldsymbol{\varepsilon} = (\mathbf{Y} - \mathbf{Xa})^T(\mathbf{Y} - \mathbf{Xa}). \tag{1.21}$$

We can obtain the optimal parameter a that minimizes the SSE through the normal equation

$$\frac{\partial(SSE)}{\partial \mathbf{a}} = \mathbf{X}^T(\mathbf{Y} - \mathbf{Xa}) = \mathbf{0}, \tag{1.22}$$

producing

$$\mathbf{a} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}, \tag{1.23}$$

which is equivalent to the least squares estimate (1.17).

https://doi.org/10.1017/9781108924238.003 Published online by Cambridge University Press

Colloquy 1.2

Goodness-of-fit

We can assess the goodness-of-fit by defining the total sum of squares (SST):

$$SST = \sum_i \left(Y_i - \overline{Y}\right)^2, \tag{1.24}$$

which measures the prediction error without using regression. In contrast, the SSE, defined in (1.20), reflects the prediction error using the least squares regression. How much prediction error is reduced by using the regression?

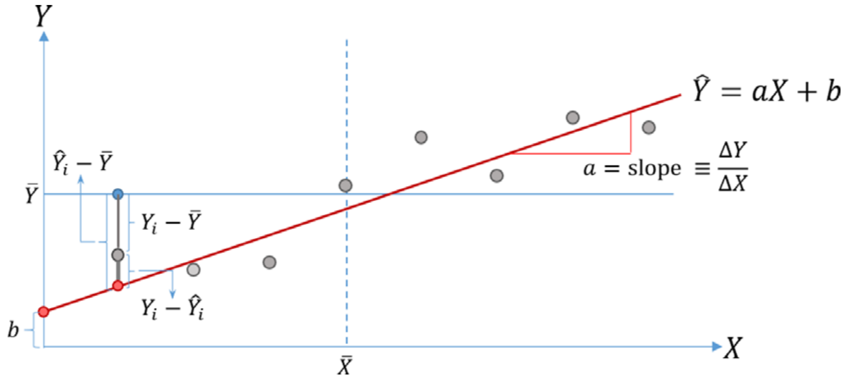
By combining the SSE and SST, we can further define a measure – the *coefficient of determination* or R^2 – to test the goodness of the regression model:

$$R^2 = \frac{SST - SSE}{SST} = 1 - \frac{SSE}{SST}, \tag{1.25}$$

which measures the proportion of variation in the predictand (i.e., dependent variable) that has been explained by the regression model. For instance,

$$R^2 = \frac{50.18 - 17.76}{50.18} \approx 0.6461$$

means that 64.61% of the variance in Y can be explained by the variance in X: The remaining variation in Y may be due to random variability. It further tells us that the overall sum of squares of 50.18 without regression is reduced down to 17.76 by empoying the least squares regression. Note that $R^2 = 1$ implies the perfect linear fit; thus, the higher the R^2 value is, the better the regression model fits the data. See the following figure for interpreting the linear regression model in a given scatter plot and the errors involved in the linear regression.



Practice 1.3 Linear regression: Global warming vs sea level rise

Shown in the table are 11-year average values, centered at the specified years, of the change in global mean temperature (ΔT) and the change in global mean sea level (ΔSL). For example, the values in 1895 are averaged over 1890–1990.

1. Plot a scatter diagram with ΔT on the x-axis and ΔSL on the y-axis. Construct a linear least squares regression model, i.e., calculate a and b in (1.19), and draw the regression line.
2. Estimate the corresponding values of ΔSL for the values of $\Delta T = 0.75$ and 1.5°C.
3. Discuss the accuracy of the regression model in terms of R^2 in (1.25).

Year	ΔT (°C)	ΔSL (m)
1885	−0.205	0.004
1895	−0.211	0.017
1905	−0.315	0.030
1915	−0.315	0.043
1925	−0.225	0.048
1935	−0.102	0.060
1945	0.018	0.080
1955	−0.044	0.102
1965	−0.023	0.115
1975	0.056	0.133
1985	0.250	0.150
1995	0.388	0.168
2005	0.602	0.198
2008	0.650	0.211

1.4.2 Weighted Least Squares Estimation

In operational data assimilation, many observations – of different variables (temperature, humidity, pressure, wind, etc.) and from different platforms (radiosonde, radar, satellite, etc.) – are used. Previously we have acquired $\hat{\mathbf{x}}$ by minimizing J (1.13) that assumed equal emphasis on all observations. However, we may have higher reliance (i.e., weight) on some measurements than others. This leads to the weighted least squares estimation, which defines the cost function as

$$J = (\boldsymbol{\varepsilon}^r)^T \mathbf{W} \boldsymbol{\varepsilon}^r = \|\boldsymbol{\varepsilon}^r\|^2_{\mathbf{W}} = \left(\mathbf{y}^o - \mathbf{H}\hat{\mathbf{x}}\right)^T \mathbf{W} \left(\mathbf{y}^o - \mathbf{H}\hat{\mathbf{x}}\right), \tag{1.26}$$

where $\mathbf{W}$ is a symmetric weight matrix. To obtain $\hat{\mathbf{x}}$ that minimizes (1.26), we should have the following requirements:

$$\nabla_{\hat{\mathbf{x}}} J = \frac{\partial J}{\partial \hat{\mathbf{x}}} = -2\mathbf{H}^T \mathbf{W} \mathbf{y}^o + 2\left(\mathbf{H}^T \mathbf{W} \mathbf{H}\right) \hat{\mathbf{x}} = \mathbf{0} \tag{1.27}$$

and

$$\nabla^2_{\hat{\mathbf{x}}} J = \frac{\partial^2 J}{\partial \hat{\mathbf{x}} \partial \hat{\mathbf{x}}^T} = 2\left(\mathbf{H}^T \mathbf{W} \mathbf{H}\right) \text{ is positive definite.} \tag{1.28}$$

Then, we obtain the weighted least squares estimate (WLSE) $\hat{\mathbf{x}}$ from (1.27) as:

$$\hat{\mathbf{x}} = \left(\mathbf{H}^T \mathbf{W} \mathbf{H}\right)^{-1} \mathbf{H}^T \mathbf{W} \mathbf{y}^o, \tag{1.29}$$

where $\mathbf{W}$ is positive definite from (1.28).

We now define the least squares cost function by normalizing the squared errors with the observation error covariance $\mathbf{R}$ in (1.6) or by putting the optimal weight matrix $\mathbf{W} = \mathbf{R}^{-1}$ from (1.26):

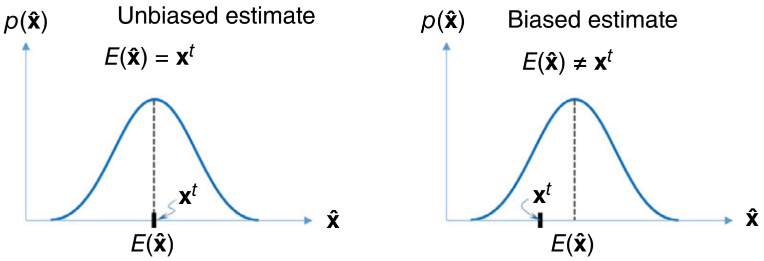


Figure 1.5 Unbiased vs biased estimate $\hat{\mathbf{x}}$ in terms of the PDF of the estimate $p(\hat{\mathbf{x}})$.

$$J = (\boldsymbol{\varepsilon}^r)^T \mathbf{R}^{-1} \boldsymbol{\varepsilon}^r = \left(\mathbf{y}^o - \mathbf{H}\hat{\mathbf{x}}\right)^T \mathbf{R}^{-1} \left(\mathbf{y}^o - \mathbf{H}\hat{\mathbf{x}}\right). \tag{1.30}$$

By minimizing J, we obtain

$$\hat{\mathbf{x}} = \left(\mathbf{H}^T \mathbf{R}^{-1} \mathbf{H}\right)^{-1} \mathbf{H}^T \mathbf{R}^{-1} \mathbf{y}^o. \tag{1.31}$$

Note that (1.31) becomes (1.17) when the measurement errors are uncorrelated (i.e., $\mathbf{R}$ is a diagonal matrix) and all errors have equal variance (i.e., $\mathbf{R} = \sigma^2 \mathbf{I}$) (Gelb, 1974).

1.4.3 Best Linear Unbiased Estimate

Given the observation model (1.2), where $\boldsymbol{\varepsilon}^o$ has zero mean and covariance matrix $E(\boldsymbol{\varepsilon}^o (\boldsymbol{\varepsilon}^o)^T) = \mathbf{R}$, as in (1.6), the *best linear unbiased estimate* (BLUE) of the true state $\mathbf{x}^t$, based on data $\mathbf{y}^o$, should satisfy the following conditions: 1) to be *linear*; 2) to be *unbiased*; and 3) to have the *minimum variance* among all unbiased linear estimates.

The observation model becomes linear by taking $H = \mathbf{H}$, where $\mathbf{H}$ is the linearized version of H, i.e.,

$$\mathbf{y}^o = \mathbf{H}\mathbf{x}^t + \boldsymbol{\varepsilon}^o, \tag{1.32}$$

and by assuming the estimate is a linear function of the data, in the form of $\hat{\mathbf{x}} = \mathbf{z}^T \mathbf{y}^o$, as in (1.17) and (1.29). The *bias* of the estimate is defined as $E(\hat{\mathbf{x}}) - \mathbf{x}^t$: if $E(\hat{\mathbf{x}}) = \mathbf{x}^t$, then $\hat{\mathbf{x}}$ is an *unbiased* estimate of $\mathbf{x}^t$. Note that the zero-mean condition (1.3) also implies that $\hat{\mathbf{x}}$ is unbiased (do Practice 1.4). Figure 1.5 provides a graphical interpretation of the unbiased and biased estimate in terms of the PDF, which will be explained in more detail in Chapter 2.

Practice 1.4 Unbiased estimate $\hat{\mathbf{x}}$

Using Eqs. (1.17), (1.3), and (1.32), show that $E(\hat{\mathbf{x}}) = \mathbf{x}^t$.

We now discuss the minimum variance condition for BLUE. Consider a linear unbiased estimate of $\mathbf{x}^t$:

$$\hat{\mathbf{x}} = \mathbf{z}^T \mathbf{y}^o. \tag{1.33}$$

https://doi.org/10.1017/9781108924238.003 Published online by Cambridge University Press

Noting that $\hat{\mathbf{x}}$ is unbiased,

$$\begin{aligned} E(\hat{\mathbf{x}}) &= \mathbf{x}^t \\ &= E\left(\mathbf{z}^T\mathbf{y}^o\right) = E\left(\mathbf{z}^T(\mathbf{H}\hat{\mathbf{x}} + \boldsymbol{\varepsilon}^r)\right) \\ &= \mathbf{z}^T\mathbf{H}\mathbf{x}^t; \end{aligned} \tag{1.34}$$

thus, $\mathbf{z}^T\mathbf{H} = \mathbf{I}$. The covariance is given by (do Practice 1.5)

$$\begin{aligned} cov(\hat{\mathbf{x}}) &= E\left((\hat{\mathbf{x}} - \mathbf{x}^t)(\hat{\mathbf{x}} - \mathbf{x}^t)^T\right) \\ &= \mathbf{z}^T\mathbf{R}\mathbf{z}. \end{aligned} \tag{1.35}$$

For BLUE, based on the Gauss–Markov Theorem (see Colloquy 1.3),

$$\mathbf{z}_{BLUE}^T = \left(\mathbf{H}^T\mathbf{R}^{-1}\mathbf{H}\right)^{-1}\mathbf{H}^T\mathbf{R}^{-1}, \tag{1.36}$$

following (1.31) and (1.43) in Colloquy 1.3. We also note that $\mathbf{z}_{BLUE}^T\mathbf{H} = \mathbf{I}$, and hence $E(\hat{\mathbf{x}}_{BLUE}) = \mathbf{x}^t$. The difference between $\mathbf{z}_{BLUE}^T$ and $\mathbf{z}^T$ is

$$\begin{aligned} \mathbf{z}_{BLUE}^T - \mathbf{z}^T &= \mathbf{d} \\ &= \left(\mathbf{H}^T\mathbf{R}^{-1}\mathbf{H}\right)^{-1}\mathbf{H}^T\mathbf{R}^{-1} - \mathbf{z}^T, \end{aligned} \tag{1.37}$$

giving $\mathbf{dH} = 0$ (do Practice 1.5); then, we can rewrite (1.35) in terms of $\mathbf{d}$ as (do Practice 1.5)

$$cov(\hat{\mathbf{x}}) = \left(\mathbf{H}^T\mathbf{R}^{-1}\mathbf{H}\right)^{-1} + \mathbf{d}\mathbf{R}\mathbf{d}^T. \tag{1.38}$$

The covariance of BLUE is

$$\begin{aligned} cov(\hat{\mathbf{x}}_{BLUE}) &= \mathbf{z}_{BLUE}^T\mathbf{R}\mathbf{z}_{BLUE} \\ &= \left(\mathbf{H}^T\mathbf{R}^{-1}\mathbf{H}\right)^{-1}\mathbf{H}^T\mathbf{R}^{-1}\mathbf{R}\left(\left(\mathbf{H}^T\mathbf{R}^{-1}\mathbf{H}\right)^{-1}\mathbf{H}^T\mathbf{R}^{-1}\right)^T \\ &= \left(\mathbf{H}^T\mathbf{R}^{-1}\mathbf{H}\right)^{-1}. \end{aligned} \tag{1.39}$$

By taking the difference between (1.38) and (1.39), we have

$$cov(\hat{\mathbf{x}}) - cov(\hat{\mathbf{x}}_{BLUE}) = \mathbf{d}\mathbf{R}\mathbf{d}^T. \tag{1.40}$$

Note that $\mathbf{d}\mathbf{R}\mathbf{d}^T$ is positive semidefinite (do Practice 1.5), making

$$cov(\hat{\mathbf{x}}) \geq cov(\hat{\mathbf{x}}_{BLUE}). \tag{1.41}$$

Therefore, $\hat{\mathbf{x}}_{BLUE}$ has the minimum variance (i.e., "best") among the linear unbiased estimate $\hat{\mathbf{x}}$ in (1.33).

https://doi.org/10.1017/9781108924238.003 Published online by Cambridge University Press

Colloquy 1.3

Gauss–Markov Theorem

If the observations constitute a general linear model in the form of (1.32), i.e.,

$$\mathbf{y}^o = \mathbf{H}\mathbf{x}^t + \boldsymbol{\varepsilon}^o,$$

where $\mathbf{H}$ is a linear observation operator matrix of $n \times m$, $\mathbf{x}^t$ is an $m \times 1$ vector of states to be estimated, and $\boldsymbol{\varepsilon}^o$ is an $n \times 1$ measurement error vector with zero mean and covariance $\mathbf{R}$ (i.e., $E(\boldsymbol{\varepsilon}^o) = \mathbf{0}$ and $E(\boldsymbol{\varepsilon}^o(\boldsymbol{\varepsilon}^o)^T) = \mathbf{R}$), then the WLSE in (1.31) becomes the BLUE of $\mathbf{x}^t$. That is,

$$\hat{\mathbf{x}}_{WLSE} = \hat{\mathbf{x}}_{BLUE} = \left(\mathbf{H}^T\mathbf{R}^{-1}\mathbf{H}\right)^{-1}\mathbf{H}^T\mathbf{R}^{-1}\mathbf{y}^o. \tag{1.42}$$

The covariance of $\hat{\mathbf{x}}_{BLUE}$ is

$$cov\left(\hat{\mathbf{x}}_{BLUE}\right) = \left(\mathbf{H}^T\mathbf{R}^{-1}\mathbf{H}\right)^{-1}, \tag{1.43}$$

with its diagonal elements are the minimum variance:

$$var\left(\hat{x}_i\right)_{\min} = \left[\left(\mathbf{H}^T\mathbf{R}^{-1}\mathbf{H}\right)^{-1}\right]_{ii}. \tag{1.44}$$

Practice 1.5 BLUE

Solve the following:

1. For any linear unbiased estimate $\hat{\mathbf{x}} = \mathbf{z}^T\mathbf{y}^o$ in (1.33), show that the relation in (1.35) holds, i.e.,
$$cov(\hat{\mathbf{x}}) = \mathbf{z}^T\mathbf{R}\mathbf{z}.$$
2. For $\mathbf{d} = \mathbf{z}_{BLUE}^T - \mathbf{z}^T$ in (1.37), show that
$$\mathbf{dH} = 0.$$
(*Hint*: Note that $\mathbf{z}^T\mathbf{H} = \mathbf{I}$ from (1.33).)
3. Derive the relation in (1.38) from (1.35), i.e.,
$$\begin{aligned} cov(\hat{\mathbf{x}}) &= \mathbf{z}^T\mathbf{R}\mathbf{z} \\ &= \left(\mathbf{H}^T\mathbf{R}^{-1}\mathbf{H}\right)^{-1} + \mathbf{dRd}^T, \end{aligned}$$
using (1.37) and $\mathbf{dH} = 0$.
4. Show that $\mathbf{dRd}^T$ in (1.38) is always positive semidefinite. (*Hint*: Referring to (1.18), you may define any $\mathbf{q} \in \mathcal{R}^m$ and multiply $\mathbf{q}$ or $\mathbf{q}^T$ to both sides of $\mathbf{dRd}^T$.)

https://doi.org/10.1017/9781108924238.003 Published online by Cambridge University Press

Example 1.3 BLUE with two measurements

Assume that we have measurements of humidity (q) at two places in an auditorium – say, q_1 near the air conditioner and q_2 away from the air conditioner. We define the observation model similar to (1.32), with $\mathbf{H} = \mathbf{I}$ and the observation error (ε^o) following the statistics in (1.6), as

$$q_i^o = q + \varepsilon_i^o \quad \text{for} \quad i = 1, 2. \tag{1.45}$$

The BLUE $\hat{q}$ should be 1) linear (say, $\hat{q} = c_i q_i^o$ for $i = 1, 2$), 2) unbiased (i.e., $E(\hat{q}) = q$), and have 3) minimum variance. Because $\hat{q}$ is unbiased,

$$E\left(\hat{q}\right) = q = E\left(c_1 q_1^o + c_2 q_2^o\right) = E\left(c_1(q + \varepsilon_1^o) + c_2(q + \varepsilon_2^o)\right) = (c_1 + c_2)q;$$

thus, $c_1 + c_2 = 1$. The variance of $\hat{q}$ is (do Practice 1.6)

$$E\left((\hat{q} - q)^2\right) = c_1^2\sigma_1^2 + (1 - c_1)^2\sigma_2^2, \tag{1.46}$$

making its minimum occur when

$$c_1 = \frac{\sigma_2^2}{\sigma_1^2 + \sigma_2^2}; \quad c_2 = \frac{\sigma_1^2}{\sigma_1^2 + \sigma_2^2}. \tag{1.47}$$

Then, the BLUE $\hat{q}$ becomes

$$\hat{q} = \frac{\sigma_2^2 q_1^o + \sigma_1^2 q_2^o}{\sigma_1^2 + \sigma_2^2}. \tag{1.48}$$

Practice 1.6 Variance of BLUE

Answer the following:

1. Show how (1.46) is obtained.
2. Show how to obtain c_1 in (1.47) that minimizes the variance (1.46).
3. Define a least squares cost function J as

$$J(q) = \frac{(q - q_1^o)^2}{\sigma_1^2} + \frac{(q - q_2^o)^2}{\sigma_2^2},$$

   and find the BLUE by minimizing J. Is the solution similar to the one in (1.48)?
4. Assume that q_1^o is a background (or first guess), i.e., $q_1^o = q^b$, and q_2^o is a real observation, say $q_2^o = q$. Express $\hat{q}$ in terms of the innovation (i.e., $q - q^b$) and compare it with the analysis equation (1.11).

https://doi.org/10.1017/9781108924238.003 Published online by Cambridge University Press

1.4.4 BLUE with a Background

Assume that we have a background $\mathbf{x}^b$ – the model-generated gridded observations – and a real observation $\mathbf{y}^o$, and we aim at getting the BLUE or the analysis, i.e., $\mathbf{x}^a = \hat{\mathbf{x}}_{BLUE}$. The observation model is the same as in (1.32) and the error statistics follow (1.6). We further assume that $\mathbf{x}^a$ is a linear combination of the background and the observation as:

$$\mathbf{x}^a = \mathbf{K}_x\mathbf{x}^b + \mathbf{K}_y\mathbf{y}^o, \tag{1.49}$$

with linear operators $\mathbf{K}_x$ and $\mathbf{K}_y$. We define the background error $\boldsymbol{\varepsilon}^b$ and the analysis error $\boldsymbol{\varepsilon}^a$ as

$$\boldsymbol{\varepsilon}^b = \mathbf{x}^b - \mathbf{x}^t \text{ and } \boldsymbol{\varepsilon}^a = \mathbf{x}^a - \mathbf{x}^t,$$

respectively, where $\boldsymbol{\varepsilon}^b$ assumed to have zero mean (i.e., $E(\boldsymbol{\varepsilon}^b) = 0$) and covariance $\mathbf{P}^b$. From (1.32) and (1.49), we have

$$\begin{aligned} E\left(\boldsymbol{\varepsilon}^a\right) &= E\left(\mathbf{K}_x\mathbf{x}^b + \mathbf{K}_y\mathbf{y}^o - \mathbf{x}^t\right) \\ &= E\left(\mathbf{K}_x\left(\boldsymbol{\varepsilon}^b + \mathbf{x}^t\right) + \mathbf{K}_y\left(\mathbf{H}\mathbf{x}^t + \boldsymbol{\varepsilon}^o\right) - \mathbf{x}^t\right) \\ &= (\mathbf{K}_x + \mathbf{K}_y\mathbf{H} - \mathbf{I})E\left(\mathbf{x}^t\right). \end{aligned}$$

As we are seeking the BLUE, $\mathbf{x}^a$ should be *unbiased* – i.e., $E(\mathbf{x}^a) = \mathbf{x}^t$, giving $E(\boldsymbol{\varepsilon}^a) = 0$ and hence $\mathbf{K}_x = \mathbf{I} - \mathbf{K}_y\mathbf{H}$. By inserting $\mathbf{K}_x$ into (1.49) and by simply putting $\mathbf{K}_y = \mathbf{K}$, we have

$$\mathbf{x}^a = \mathbf{x}^b + \mathbf{K}\left(\mathbf{y}^o - \mathbf{H}\mathbf{x}^b\right), \tag{1.50}$$

where $\mathbf{K}$ is called the *gain* or the *Kalman gain*, which maps from $\mathcal{R}^n$ to $\mathcal{R}^m$, and $\mathbf{y}^o - \mathbf{H}\mathbf{x}^b$ is the innovation. Note that (1.50) is equivalent to the analysis equation (1.11) except that the observation operator is linearized ($H = \mathbf{H}$) and the weight matrix $\mathbf{W}$ is replaced by $\mathbf{K}$ (i.e., $\mathbf{W} = \mathbf{K}$).

The gain $\mathbf{K}$ can be specified by finding the condition for minimum variance of $\boldsymbol{\varepsilon}^a$ – another property of BLUE. This implies that the analysis error covariance $\mathbf{P}^a$ must have the sum of diagonal elements that are the smallest among the linear estimates, that is,

$$\arg\min E\left(\boldsymbol{\varepsilon}_i^a(\boldsymbol{\varepsilon}^a)_i^T\right) = \arg\min\left(tr(\mathbf{P}^a)\right), \tag{1.51}$$

where $tr(\mathbf{P}^a)$ is the trace of a square matrix $\mathbf{P}^a$, defined for its diagonal elements p_{ii}^a as

$$tr(\mathbf{P}^a) = \sum_i p_{ii}^a.$$

For the square matrices $\mathbf{A}$ and $\mathbf{B}$, and a scalar β, the trace has the following properties:

$$\begin{aligned} &tr(\mathbf{A} + \mathbf{B}) = tr(\mathbf{A}) + tr(\mathbf{B});\ tr(\mathbf{A}) = tr(\mathbf{A}^T);\ tr(\beta\mathbf{A}) = \beta tr(\mathbf{A}), \\ &tr(\mathbf{A}\mathbf{B}) = tr(\mathbf{B}\mathbf{A});\ tr(\mathbf{B}\mathbf{A}\mathbf{B}^{-1}) = tr(\mathbf{A});\ tr(\mathbf{A}\mathbf{B}\mathbf{C}) = tr(\mathbf{B}\mathbf{C}\mathbf{A}) = tr(\mathbf{C}\mathbf{A}\mathbf{B}), \end{aligned} \tag{1.52}$$

https://doi.org/10.1017/9781108924238.003 Published online by Cambridge University Press

and

$$\nabla_{\mathbf{X}} tr(\mathbf{XB}) = \mathbf{B}^T;\ \nabla_{\mathbf{X}} tr(\mathbf{BX}^T) = \mathbf{B};\ \nabla_{\mathbf{X}} tr(\mathbf{BXC}) = \mathbf{B}^T\mathbf{C}^T,$$
$$\nabla_{\mathbf{X}} tr(\mathbf{XBX}^T) = \mathbf{X}(\mathbf{B}^T + \mathbf{B});\ \nabla_{\mathbf{X}} tr(\mathbf{X}^T\mathbf{BX}) = (\mathbf{B} + \mathbf{B}^T)\mathbf{X}. \tag{1.53}$$

Equation (1.50) is represented in terms of errors as (do Practice 1.7)

$$\boldsymbol{\varepsilon}^a = \boldsymbol{\varepsilon}^b + \mathbf{K}\left(\boldsymbol{\varepsilon}^o - \mathbf{H}\boldsymbol{\varepsilon}^b\right) = (\mathbf{I} - \mathbf{KH})\boldsymbol{\varepsilon}^b + \mathbf{K}\boldsymbol{\varepsilon}^o, \tag{1.54}$$

from which $\mathbf{P}^a$ is derived as (do Practice 1.7)

$$\begin{aligned}\mathbf{P}^a &= E\left(\left((\mathbf{I} - \mathbf{KH})\boldsymbol{\varepsilon}^b + \mathbf{K}\boldsymbol{\varepsilon}^o\right)\left((\mathbf{I} - \mathbf{KH})\boldsymbol{\varepsilon}^b + \mathbf{K}\boldsymbol{\varepsilon}^o\right)^T\right)\\ &= (\mathbf{I} - \mathbf{KH})\mathbf{P}^b(\mathbf{I} - \mathbf{KH})^T + \mathbf{KRK}^T,\end{aligned} \tag{1.55}$$

where $\mathbf{P}^b = E\left(\boldsymbol{\varepsilon}^b(\boldsymbol{\varepsilon}^b)^T\right)$ and $\mathbf{R} = E\left(\boldsymbol{\varepsilon}^o(\boldsymbol{\varepsilon}^o)^T\right)$. Following (1.51), we differentiate $tr(\mathbf{P}^a)$ with respect to $\mathbf{K}$ and set it to 0 (do Practice 1.7)

$$\begin{aligned}\nabla_{\mathbf{K}} tr(\mathbf{P}^a) &= 0\\ &= \nabla_{\mathbf{K}}\left[tr(\mathbf{P}^b) - 2tr(\mathbf{P}^b\mathbf{H}^T\mathbf{K}^T) + tr(\mathbf{KHP}^b\mathbf{H}^T\mathbf{K}^T) + tr(\mathbf{KRK}^T)\right]\\ &= -2tr(\mathbf{P}^b\mathbf{H}^T) + 2tr(\mathbf{KHP}^b\mathbf{H}^T) + 2tr(\mathbf{KR})\\ &= 2tr\left(\mathbf{K}(\mathbf{HP}^b\mathbf{H}^T + \mathbf{R}) - \mathbf{P}^b\mathbf{H}^T\right);\end{aligned} \tag{1.56}$$

thus,

$$\mathbf{K} = \mathbf{P}^b\mathbf{H}^T(\mathbf{HP}^b\mathbf{H}^T + \mathbf{R})^{-1}. \tag{1.57}$$

Alternatively, $\mathbf{K}$ is expressed as (do Practice 1.7)

$$\mathbf{K} = \left(\left(\mathbf{P}^b\right)^{-1} + \mathbf{H}^T\mathbf{R}^{-1}\mathbf{H}\right)^{-1}\mathbf{H}^T\mathbf{R}^{-1}. \tag{1.58}$$

Note that, with little or no background information, $(\mathbf{P}^b)^{-1}$ becomes very small and (1.50) reduces to (1.31) (Gelb, 1974).

Practice 1.7 BLUE and Kalman gain

Solve the following:

1. From (1.50), derive the error equation (1.54).
2. Derive (1.55) using the properties $E(\boldsymbol{\varepsilon}^o) = E(\boldsymbol{\varepsilon}^b) = 0$.
3. Derive (1.56) using the properties of trace (1.52) and (1.53) and the fact that $\mathbf{P}^b$ and $\mathbf{R}$ are symmetric (i.e., $(\mathbf{P}^b)^T = \mathbf{P}^b$ and $\mathbf{R}^T = \mathbf{R}$).
4. Derive (1.58) from (1.57). You may start by putting (1.57) as

$$\mathbf{K} = \mathbf{IP}^b\mathbf{H}^T(\mathbf{HP}^b\mathbf{H}^T + \mathbf{R})^{-1},$$

where

$$\mathbf{I} = \left(\left(\mathbf{P}^b\right)^{-1}\mathbf{H}^T\mathbf{R}^{-1}\mathbf{H}\right)^{-1}\left(\left(\mathbf{P}^b\right)^{-1}\mathbf{H}^T\mathbf{R}^{-1}\mathbf{H}\right).$$

https://doi.org/10.1017/9781108924238.003 Published online by Cambridge University Press

5. Define a cost function as

$$J(\mathbf{x}) = \frac{1}{2}\left(\mathbf{x} - \mathbf{x}^b\right)^T \left(\mathbf{P}^b\right)^{-1}\left(\mathbf{x} - \mathbf{x}^b\right) + \frac{1}{2}\left(\mathbf{y}^o - \mathbf{H}\mathbf{x}\right)^T \mathbf{R}^{-1}\left(\mathbf{y}^o - \mathbf{H}\mathbf{x}\right).$$

Show that the optimal estimate $\mathbf{x}^a$ is given by

$$\mathbf{x}^a = \mathbf{x}^b + \mathbf{K}\left(\mathbf{y}^o - \mathbf{H}\mathbf{x}^b\right),$$

where $\mathbf{K}$ is shown as in (1.58). You may want to find $\mathbf{x}^a$ that minimizes J, i.e.,

$$\mathbf{x}^a = \arg\min J(\mathbf{x})$$

so that

$$\nabla_{\mathbf{x}} J = 0.$$

The result implies that the analysis obtained by minimizing the cost function J (i.e., *variational* analysis) corresponds to the analysis through the minimum error variance approach (i.e., BLUE).

6. Show that the analysis error $\boldsymbol{\varepsilon}^a$ is orthogonal to the analysis $\mathbf{x}^a$, that is,

$$E\left(\mathbf{x}^a\left(\boldsymbol{\varepsilon}^a\right)^T\right) = \mathbf{0},$$

and make a geometric interpretation.

1.5 Optimal Interpolation

Optimal interpolation (OI) is equivalent to the BLUE obtained intermittently in a discrete time domain when observation is available. The term “optimal” is employed in a sense that the analysis error variance is minimized – see (1.56); thus, $\mathbf{K}$ in (1.57) is actually regarded as an *optimal gain*, denoted by $\mathbf{K}^{\mathrm{O}}$. By putting $\mathbf{K} = \mathbf{K}^{\mathrm{O}}$ and substituting (1.57) into (1.55), we obtain

$$\mathbf{P}^a = (\mathbf{I} - \mathbf{K}^{\mathrm{O}}\mathbf{H})\mathbf{P}^b. \tag{1.59}$$

Therefore, the OI solution is nothing but the BLUE, represented by (1.50), (1.57), and (1.59). Solving the analysis equation requires direct inversion and is computationally expensive with all global observations. In OI, the calculation of $\mathbf{K}^{\mathrm{O}}$ is simplified by using observations only near the grid (analysis) point. That is, OI acquires the analysis $\mathbf{x}^a$ over an analysis circle (or block), consisting of a grid point and nearby observations (i.e., *localized*) within the so-called radius of influence, r (see Figure 1.6). Depending on r, some observations are used twice while some other observations are not used.

The OI scheme, for a given analysis cycle, is written in terms of the analysis circle index, i, as

$$\begin{aligned}
\mathbf{x}_i^a &= \mathbf{x}_i^b + \mathbf{K}_i^{\mathrm{O}}\left(\mathbf{y}^o - \mathbf{H}\mathbf{x}^b\right)_i \\
\mathbf{K}_i^{\mathrm{O}} &= \left(\mathbf{P}^b\mathbf{H}^T\right)_i\left((\mathbf{H}\mathbf{P}^b\mathbf{H}^T)_i + \mathbf{R}_i\right)^{-1} \\
\mathbf{P}_i^a &= \left(\mathbf{I} - \mathbf{K}^{\mathrm{O}}\mathbf{H}\right)_i\mathbf{P}_i^b,
\end{aligned} \tag{1.60}$$

https://doi.org/10.1017/9781108924238.003 Published online by Cambridge University Press

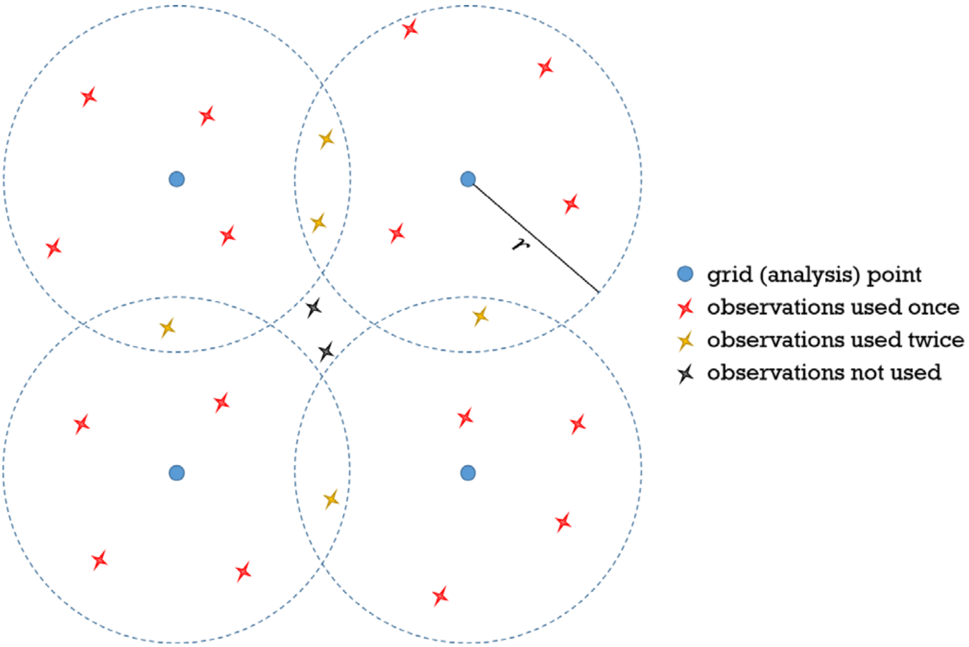


Figure 1.6 OI performed on the grid (analysis) points, centered at the analysis circles. Depending on the size of the analysis circle, which is determined by the radius of influence, r, the numbers of observations and their usage for analysis are different.

where $\mathbf{P}_i^b$ is specified through statistical measures (e.g., autocorrelation functions) and dynamic constraints (e.g., geostrophic balance). The analysis equation in OI is equivalent to (1.11) except that the weight $\mathbf{W}$ is replaced by the optimal gain $\mathbf{K}_i^{\mathrm{O}}$ and the observation operator H is linearized. That is, the analysis is given by correcting the background with the analysis increment – the product of the optimal gain and the innovation. The optimal gain is obtained as the product between the background error covariance in the observation space and the inverse of total error covariance (Kalnay, 2003).

From (1.60), by representing a fixed $\mathbf{P}^b$ as $\mathbf{B}$, the analysis increment can be expressed as

$$\mathbf{x}_i^a - \mathbf{x}_i^b = \overbrace{\left(\mathbf{BH}^T\right)_i \underbrace{\left((\mathbf{HBH}^T)_i + \mathbf{R}_i\right)^{-1} \left(\mathbf{y}^o - \mathbf{Hx}^b\right)_i}_{\mathcal{R}^n}}^{\mathcal{R}^m}, \tag{1.61}$$

where $\mathcal{R}^m$ is the model (grid) space and $\mathcal{R}^n$ is the observation space. Note that $\mathbf{H}$ performs a transformation of $\mathcal{R}^m \longrightarrow \mathcal{R}^n$ while $\mathbf{H}^T$ does that of $\mathcal{R}^n \longrightarrow \mathcal{R}^m$. The analysis increment is calculated first by computing $\left((\mathbf{HBH}^T)_i + \mathbf{R}_i\right)^{-1}\left(\mathbf{y}^o - \mathbf{Hx}^b\right)_i$ in the observation space, then by transforming it to the model space by applying $(\mathbf{BH}^T)_i$. This implies that the OI analysis is affected not only by the relevant observations $(\mathbf{y}^o)_i$ but also by the information and structure of the background error covariance $\mathbf{B}_i$. In OI, $\mathbf{B}_i$ has a stationary

https://doi.org/10.1017/9781108924238.003 Published online by Cambridge University Press

(i.e., time-invariant) structure: It is generally defined by an isotropic correlation function, depending only on the radius of influence with zero correlations for very large separations between grid points and observations, and enforces most dynamical balance properties reasonably well (see Bouttier and Courtier, 2002). With complex observation operators, calculating $(\mathbf{BH}^T)_i$ is quite difficult.

Practice 1.8 Properties of operators in analysis increment

Take the same measurements and model states as in Example 1.1; that is, the measurements are made at three levels and the model states are calculated at six levels (see the figure in Example 1.1). Then the analysis $\mathbf{x}^a$ and the background $\mathbf{x}^b$ are given by

$$\mathbf{x}^a = \left(q_1^a, \ldots, q_6^a\right) \text{ and } \mathbf{x}^b = \left(q_1^b, \ldots, q_6^b\right),$$

respectively. To avoid any confusion, set the observation levels using an alphabetical index, say, $\mathbf{y}^o = \left(q_a^o, q_b^o, q_c^o\right)^T$.

1. Construct the matrices of the background error covariance $\mathbf{B}$ and the linearized observation operator $\mathbf{H}$. [*Hint*: The elements of $\mathbf{B}$ are the covariances between grid points (b_{23}, b_{56}, etc.) and those of $\mathbf{H}$ represent interpolation between the model space and observation space (h_{a5}, h_{c2}, etc.)]
2. Using $\mathbf{B}$ and $\mathbf{H}$ from #1, evaluate the following matrices and provide interpretation on each of them:

$$\mathbf{BH}^T \text{ and } \mathbf{HBH}^T.$$

Following Hollingsworth (1986), we can extend our interpretation of OI as a filter and an interpolator by manipulating (1.61) as follows:

$$\mathbf{x}_i^a - \mathbf{x}_i^b = \underbrace{\left(\mathbf{BH}^T\right)_i \left(\mathbf{HBH}^T\right)_i^{-1}}_{\mathcal{B}} \underbrace{\left(\mathbf{HBH}^T\right)_i \left(\left(\mathbf{HBH}^T\right)_i + \mathbf{R}_i\right)^{-1}}_{\mathcal{A}} \left(\mathbf{y}^o - \mathbf{Hx}^b\right)_i, \quad (1.62)$$

where we put $\left(\mathbf{HBH}^T\right)_i^{-1} \left(\mathbf{HBH}^T\right)_i = \mathbf{I}$. By applying $\mathcal{A}$ to the innovations $\left(\mathbf{y}^o - \mathbf{Hx}^b\right)_i$, OI filters the innovations (or observations) to generate the analysis increments $\left(\mathbf{x}^a - \mathbf{x}^b\right)_i$ in the observation space. Then, by further operating $\mathcal{B}$, OI interpolates (or propagates) the analysis increments from the observation space to the model space (i.e., grid points). Note that $\mathbf{HBH}^T$ maps the background to the observation space and the observation-error covariances are given; thus, the filtering process occurs in the observation space by converting the innovations to the analysis increments. As $\left(\mathbf{BH}^T\right)_i$ represents the background error covariances operating between the observation space and the model space, the interpolating process transforms the analysis increments from the observation points to the model grid points through normalization by $\left(\mathbf{HBH}^T\right)_i^{-1}$. When the cross-variable background error correlations are provided, the process $\mathcal{B}$ can transform the analysis increment of a certain

https://doi.org/10.1017/9781108924238.003 Published online by Cambridge University Press

variable to that of another variable. Overall, the spatial propagation of the corrections by observations (i.e., innovations) is performed by the background error covariances **B**.

The OI method basically assumes that, for each state variable, the analysis increment is mainly determined by just a few relevant observations (Bouttier and Courtier, 2002). Since the dimension of $\left((\mathbf{HBH}^T)_i + \mathbf{R}_i\right)$ in (1.60) is equal to the number of observations, selecting relevant observations reduces its size (i.e., approximating a block diagonal matrix) and makes the direct inversion viable. For effective calculation of $(\mathbf{BH}^T)_i$ and $(\mathbf{HBH}^T)_i$, one should employ simple observation operators (e.g., interpolation of direct state observations) that are sparse. Due to the selection and use of local data within the analysis circles, the OI analysis fields include spurious noise and sometimes show incoherency between small and large scales (Lorenc, 1981; Bouttier and Courtier, 2002; Dance, 2004). The OI scheme had been widely used in operational centers in 1980s and 1990s (e.g., Lorenc, 1981; Lyne et al., 1982; DiMego, 1988; Kanamitsu, 1989) because of its simplicity and computational efficiency but had been replaced by a variational method due to the disadvantages mentioned above (e.g., Parrish and Derber, 1992; Courtier et al., 1998).

Algorithm 1.2 shows a general prediction process based on OI while Algorithm 1.3 depicts the computation process of $\mathbf{K}^{\mathrm{O}}$ in detail (e.g., Bouttier and Courtier, 2002; Dance, 2004).

Algorithm 1.2 Prediction process based on OI

```
/* This algorithm is based on (1.60) with $\mathbf{P}^b = \mathbf{B}$                         */
/* index $n$ denotes analysis cycle (time)                                                    */
/* index $i$ denotes analysis circle (block) centered at grid (analysis) point               */
/* $\mathbf{x}^f$ denotes the future state (forecast) and $\mathbf{x}^b$ the background       */
/* $M$ denotes the model propagator                                                           */
Initiation: $\left(\mathbf{x}^b\right)_i^0 = \left(\mathbf{x}^f\right)_i^0$ at the initial time     ! Provide a background at $n = 0$
for n=0 to nmax do                                                    ! Loop for analysis cycle (time)
    for i=1 to imax do                                                ! Loop for analysis circle (block)
        if ($\mathbf{y}^o$ exists) then                               ! When observations are available
            $\left(\mathbf{K}^{\mathrm{O}}\right)_i^n = \left(\mathbf{BH}^T\right)_i^n \left((\mathbf{HBH}^T)_i^n + \mathbf{R}_i^n\right)^{-1}$     ! Calculate the optimal gain
                                                                      ! See Algorithm 1.3
            $(\mathbf{x}^a)_i^n = \left(\mathbf{x}^b\right)_i^n + \left(\mathbf{K}^{\mathrm{O}}\right)_i^n \left(\mathbf{y}^o - \mathbf{H}\mathbf{x}^b\right)_i^n$     ! Calculate the analysis
            $\left(\mathbf{x}^f\right)_i^{n+1} = M^n \left(\mathbf{x}^a\right)_i^n$     ! Obtain the future state using the analysis
        else                                                          ! When no observations are available
            $\left(\mathbf{x}^f\right)_i^{n+1} = M^n \left(\mathbf{x}^f\right)_i^n$     ! Obtain the future state using current state
        end
        $\left(\mathbf{x}^b\right)_i^{n+1} = \left(\mathbf{x}^f\right)_i^{n+1}$     ! Assign the forecast to the background
    endfor
endfor
```

Algorithm 1.3 Process of calculating $\mathbf{K}^{\mathrm{O}}$ for each state variable in OI

```
/* Calculate $\mathbf{K}^{\mathrm{O}}$ for a given state variable, e.g., temperature, humidity, etc. */
/* This algorithm is based on (1.60) with $\mathbf{P}^b = \mathbf{B}$ */
Input: $\mathbf{x}^b, \mathbf{y}^o, \mathbf{B}, \mathbf{R}$
Output: $\mathbf{K}^{\mathrm{O}}$
begin
    Step 1: Determine the radius of influence $r$ based on empirical selection criteria to
      specify the analysis circle (block).
    Step 2: Select $l$ observations within the the analysis circle.
    Step 3: Calculate $\left(\mathbf{y}^o - \mathbf{H}\mathbf{x}^b\right)_l$ relevant to the $l$ observations.
    Step 4: Calculate $l \times l$ submatrices of $\mathbf{HBH}^T$ and $\mathbf{R}$ to form $\left(\mathbf{HBH}^T + \mathbf{R}\right)_l$.
    Step 5: Calculate a row vector $\left(\mathbf{BH}^T\right)_l$ for the given state variable, restricted to the
      $l$ observations.
    Step 6: Calculate the inverse of $\left(\mathbf{HBH}^T + \mathbf{R}\right)_l$.
    Step 7: Calculate $\mathbf{K}_l^{\mathrm{O}} = \left(\mathbf{BH}^T\right)_l \left(\mathbf{HBH}^T + \mathbf{R}\right)_l^{-1}$.
end
```

References

Barnes SL (1964) A technique for maximizing details in numerical weather map analysis. *J Appl Meteor* 3:396–409.

Bergthórsson P, Döös BR (1955) Numerical weather map analysis. *Tellus* 7:329–340.

Bonavita M, Isaksen L, Hólm E (2012) On the use of EDA background error variances in the ECMWF 4D-Var. *Quart J Roy Meteor Soc* 138:1540–1559.

Bouttier F, Courtier P (2002) *Data Assimilation Concepts and Methods*. Meteorological Training Course Lecture Series, ECMWF, Reading, 59 pp., www.ecmwf.int/node/16928

Buehner M (2005) Ensemble-derived stationary and flow-dependent background-error covariances: Evaluation in a quasi-operational NWP setting. *Quart J Roy Meteor Soc* 131:1013–1043.

Chorin A, Morzfeld M, Tu X (2010) Implicit particle filters for data assimilation. *Commun Appl Math Comput Sci* 5:221–240.

Clayton AM, Lorenc AC, Barker DM (2013) Operational implementation of a hybrid ensemble/4D-Var global data assimilation system at the Met Office. *Quart J Roy Meteor Soc* 139:1445–1461.

Courtier P, Thépaut JN, Hollingsworth A (1994) A strategy for operational implementation of 4D-Var, using an incremental approach. *Quart J Roy Meteor Soc* 120:1367–1387.

Courtier P, Andersson E, Heckley W, et al. (1998) The ECMWF implementation of three-dimensional variational assimilation (3D-Var). I: Formulation. *Quart J Roy Meteor Soc* 124:1783–1807.

Cressman GP (1959) An operational objective analysis system. *Mon Wea Rev* 87:367–374.

Daley R (1991) *Atmospheric Data Analysis*. Cambridge University Press, New York, 457 pp.

Dance SL (2004) Issues in high resolution limited area data assimilation for quantitative precipitation forecasting. *Physica D* 196:1–27.

Davies HC, Turner RE (1977) Updating prediction models by dynamical relaxation: An examination of the technique. *Quart J Roy Meteor Soc* 103:225–245.

Derber JC, Wu WS (1998) The use of TOVS cloud-cleared radiances in the NCEP SSI analysis system. *Mon Wea Rev* 126:2287–2299.

DiMego GJ (1988) The National Meteorological Center regional analysis system. *Mon Wea Rev* 116:977–1000.

ECMWF (2020) Monitoring: Data coverage. www.ecmwf.int/en/forecasts/charts/monitoring/dcover

Evensen G (1994) Sequential data assimilation with a nonlinear quasi-geostrophic model using Monte Carlo methods to forecast error statistics. *J Geophys Res Oceans* 99:10143–10162.

Gandin LS (1963) *Objective Analysis of Meteorological Fields*. Gidromet, Leningrad, 285 pp., English translation, Israel Program for Scientific Translations, 1965.

Gao J (2017) A three-dimensional variational radar data assimilation scheme developed for convective scale NWP. In *Data Assimilation for Atmospheric, Oceanic and Hydrologic Applications* (vol. III), (eds.) Park SK, Xu L, Springer International Publishing, Cham, 285–326.

Gauss CF (1809) *Theoria Motus Corporum Coelestium in Sectionibus Conicis Solem Ambientium (Theory of the Motion of Heavenly Bodies Moving about the Sun in Conic Section)*. English translation by C.H. Davis, published by Little Brown, and Co. of Boston in 1857, reprinted in 1963 by Dover Publications, Inc., New York, 374 pp.

Gelb A (ed.) (1974) *Applied Optimal Estimation*. MIT Press, Cambridge, MA, and London, 374 pp.

Gilchrist B, Cressman GP (1954) An experiment in objective analysis. *Tellus* 6:309–318.

Gilmore MS, Straka JM, Rasmussen EN (2004) Precipitation and evolution sensitivity in simulated deep convective storms: Comparisons between liquid-only and simple ice and liquid phase microphysics. *Mon Wea Rev* 132:1897–1916.

Hoke JE, Anthes RA (1976) The initialization of numerical models by a dynamic-initialization technique. *Mon Wea Rev* 104:1551–1556.

Hollingsworth A (1986) Objective analysis for numerical weather prediction. *J Meteor Soc Japan* 64A:11–59.

Houtekamer PL, Mitchell HL (1998) Data assimilation using an ensemble Kalman filter technique. *Mon Wea Rev* 126:796–811.

Ide K, Courtier P, Ghil M, Lorenc AC (1997) Unified notation for data assimilation: Operational, sequential and variational. *J Meteor Soc Japan* 75:181–189.

Kalman RE, Bucy RS (1961) New results in linear filtering and prediction theory. *J Basic Eng* 83:95–108.

Kalnay E (2003) *Atmospheric Modeling, Data Assimilation and Predictability*. Cambridge University Press, New York, 341 pp.

Kanamitsu M (1989) Description of the NMC global data assimilation and forecast system. *Wea Forecasting* 4:335–342.

Lakshmivarahan S, Lewis JM (2013) Nudging methods: A critical overview. In *Data Assimilation for Atmospheric, Oceanic and Hydrologic Applications* (vol. II), (eds.) Park SK, Xu L, Springer, 27–57.

Lakshmivarahan S, Lewis JM, Phan D (2013) Data assimilation as a problem in optimal tracking: Application of Pontryagin's minimum principle to atmospheric science. *J Atmos Sci* 70:1257–1277.

https://doi.org/10.1017/9781108924238.003 Published online by Cambridge University Press

Laplace PS (1814) *Essai philosophique sur les probabilités*. Nabu Press, French ed., 2010, 212 pp.

Le Dimet F-X, Talagrand O (1986) Variational algorithms for analysis and assimilation of meteorological observations: Theoretical aspects. *Tellus A* 38:97–110.

Legendre AM (1805) *Nouvelles méthodes pour la détermination des orbites des comètes*. Firmin Didot, Paris. Pages 72–75 of the appendix reprinted in Stigler (1986, p.56). English translation of these pages by Ruger, HA and Walker, HM in *A Source Book of Mathematics* by Smith, DE, McGraw-Hill Book Company, New York, 1929, p. 576–579.

Lewis JM, Derber JC (1985) The use of adjoint equations to solve a variational adjustment problem with advective constraints. *Tellus A* 37:309–322.

Lewis JM, Lakshmivarahan S, Dhall S (2006) *Dynamic Data Assimilation: A Least Squares Approach*. Cambridge University Press, Cambridge, 654 pp.

Lin YL, Farley RD, Orville HD (1983) Bulk parameterization of the snow field in a cloud model. *J Climate Appl Meteor* 22:1065–1092.

Lorenc AC (1981) A global three-dimensional multivariate statistical interpolation scheme. *Mon Wea Rev* 109:701–721.

Lorenc AC (2003) The potential of the ensemble Kalman filter for NWP – a comparison with 4D-Var. *Quart J Roy Meteor Soc* 129:3183–3203.

Lyne WH, Swinbank R, Birch NT (1982) A data assimilation experiment and the global circulation during the FGGE special observing periods. *Quart J Roy Meteor Soc* 108:575–594.

Navon IM, Zou X, Derber J, Sela J (1992) Variational data assimilation with an adiabatic version of the NMC spectral model. *Mon Wea Rev* 120:1433–1446.

Parrish DF, Derber JC (1992) The National Meteorological Center's spectral statistical-interpolation analysis system. *Mon Wea Rev* 120:1747–1763.

Pontryagin LS, Boltyanskii VG, Gamkrelidze RV, Mischenko EF (1961) *Matematicheskaya Teoriya Optimal'nykh Prozessov*. Fizmatgiz, Moscow, English translation *The Mathematical Theory of Optimal Control Processes* by Trirogoff, KN, published in 1962 by Interscience Publishers, John Wiley & Sons, Inc., New York, London, Sydney, 360 pp.

Sakov P, Oliver DS, Bertino L (2012) An iterative enkf for strongly nonlinear systems. *Mon Wea Rev* 140:1988–2004.

Sasaki Y (1958) An objective analysis based on the variational method. *J Meteor Soc Japan* 36:77–88.

Sasaki Y (1970a) Some basic formalisms in numerical variational analysis. *Mon Wea Rev* 98:875–883.

Sasaki Y (1970b) Numerical variational analysis formulated under the constraints as determined by longwave equations and a low-pass filter. *Mon Wea Rev* 98:884–898.

Sasaki Y (1970c) Numerical variational analysis with weak constraint and application to surface analysis of severe storm gust. *Mon Wea Rev* 98:899–910.

Shannon CE, Weaver W (1949) *The Mathematical Theory of Communication*. University of Illinois Press, Chicago, IL, 117 pp.

Smith Jr. P, Myers C, Orville H (1975) Radar reflectivity factor calculations in numerical cloud models using bulk parameterization of precipitation. *J Appl Meteor* 14:1156–1165.

Sorenson HW (1970) Least-squares estimation: From Gauss to Kalman. *IEEE Spectrum* 7:63–68.

Sugimoto S, Crook NA, Sun J, Xiao Q, Barker DM (2009) An examination of WRF 3DVAR radar data assimilation on its capability in retrieving unobserved variables and forecasting precipitation through observing system simulation experiments. *Mon Wea Rev* 137:4011–4029.

Sun J, Crook NA (1997) Dynamical and microphysical retrieval from Doppler radar observations using a cloud model and its adjoint. Part I: Model development and simulated data experiments. *J Atmos Sci* 54:1642–1661.

Thacker WC (1992) Oceanographic inverse problems. *Physica D* 60:16–37.

van Leeuwen PJ (2009) Particle filtering in geophysical systems. *Mon Wea Rev* 137:4089–4114.

Wang X, Hamill TM, Whitaker JS, Bishop CH (2007) A comparison of hybrid ensemble transform Kalman filterâŁ“OI and ensemble square-root filter analysis schemes. *Mon Wea Rev* 136:5116–5131.

Xiao Q, Kuo YH, Sun J, et al. (2007) An approach of radar reflectivity data assimilation and its assessment with the inland QPF of Typhoon Rusa (2002) at landfall. *J Appl Meteor Climatol* 46:14–22.

Zou X, Navon IM, Ledimet FX (1992) An optimal nudging data assimilation scheme using parameter estimation. *Quart J Roy Meteor Soc* 118:1163–1186.

Županski M (1993) Regional four-dimensional variational data assimilation in a quasi-operational forecasting environment. *Mon Wea Rev* 121:2396–2408.

Županski M (2005) Maximum likelihood ensemble filter: Theoretical aspects. *Mon Wea Rev* 133:1710–1726.

https://doi.org/10.1017/9781108924238.003 Published online by Cambridge University Press