The Role of the Model in the Data Assimilation System

Richard B. Rood

University of Michigan, Ann Arbor, USA, rbrood@umich.edu

1 Introduction

The chapters in Part A, *Theory*, describe in some detail the theory and methodology of data assimilation. This chapter will focus on the role of the predictive model in an assimilation system. There are numerous books on atmospheric modelling, their history, their construction, and their applications (*e.g.* Trenberth 1992; Randall 2000; Jacobson 2005). This chapter will focus on specific aspects of the model and modelling in data assimilation.

The chapter is outlined as follows:

- Definition and Description of the Model;
- Role of the Model in Data Assimilation;
- Component Structure of an Atmospheric Model;
- Consideration of the Observation-Model Interface;
- Physical Consistency and Data Assimilation;
- Summary.

2 Definition and description of the model

Dictionary definitions of *model* include:

- "A work or construction used in testing or perfecting a final product";
- "A schematic description of a system, theory, or phenomenon that accounts for its known or inferred properties and may be used for further studies of its characteristics".

In atmospheric modelling a scientist is generally faced with a set of observations of variables, for instance, wind, temperature, water, ozone, *etc.*, as well as either the knowledge or expectation of correlated behaviour between the different variables. A number of types of models could be developed to describe the observations. These include:

- Conceptual or heuristic models which outline in the simplest terms the processes that describe the interrelation between different observed phenomena. These models are often intuitively or theoretically based. An example would be the tropical pipe model of Plumb and Ko (1992), which describes the transport of long-lived tracers in the stratosphere;
- Statistical models which describe the behaviour of the observations based on the observations themselves. That is the observations are described in terms of the mean, the variance, and the correlations of an existing set of

observations. Johnson *et al.* (2000) discuss the use of statistical models in the prediction of tropical sea surface temperatures;

- Physical models which describe the behaviour of the observations based on first principle tenets of physics (chemistry, biology, *etc.*). In general, these principles are expressed as mathematical equations, and these equations are solved using discrete numerical methods. Detailed discussions of modelling include Trenberth (1992), Randall (2000), and Jacobson (2005).

In the study of geophysical phenomena, there are numerous subtypes of models. These include comprehensive models which attempt to model all of the relevant couplings or interactions in a system and mechanistic models which have prescribed variables, and the system evolves relative to the prescribed parameters. All of these models have their place in scientific investigation, and it is often the interplay between the different types and subtypes of models that leads to scientific advance.

Models are used in two major roles. The first role is *diagnostic*, in which the model is used to determine and to test the processes that are thought to describe the observations. In this case, it is determined whether or not the processes are well known and adequately described. In general, since models are an investigative tool, such studies are aimed at determining the nature of unknown or inadequately described processes. The second role is *prognostic*; that is, the model is used to make a prediction.

In all cases the model represents a management of complexity; that is, a scientist is faced with a complex set of observations and their interactions and is trying to manage those observations in order to develop a quantitative representation. In the case of physical models, which are implicitly at focus here, a comprehensive model would represent the cumulative knowledge of the physics (chemistry, biology, *etc.*) that describe the observations. It is tacit, that an accurate, validated, comprehensive physical model is the most robust way to forecast; that is, to predict the future.

The *physical principles* represented in an atmospheric model, for example, are a series of *conservation equations* which quantify the conservation of momentum, mass, and thermodynamic energy. The equation of state describes the relation between the thermodynamic variables. Because of the key roles that phase changes of water play in atmospheric energy exchanges, an equation for the conservation of water is required. Models which include the transport and chemistry of atmosphere trace gases and aerosols require additional conservation equations for these constituents. The conservation equations for mass, trace gases, and aerosols are often called *continuity equations*.

In general, the conservation equation relates the time rate of change of a quantity to the sum of the quantity's production and loss. For momentum the production and loss follow from the forces described by Newton's Laws of Motion. Since the atmosphere is a fluid, either a *Lagrangian* or an *Eulerian* description of the flow can be used (see chapter *General Concepts in Meteorology and Dynamics*, Charlton-Perez *et al.*). The Lagrangian description follows a notional fluid parcel, and the Eulerian description relies on spatial and temporal field descriptions of the flow at a particular point in the domain. Data assimilation can be performed in either the Lagrangian or Eulerian framework. In this chapter the Eulerian framework will be the primary focus. Holton (2004) provides a thorough introduction to the

fundamental equations of motions and their scaling and application to atmospheric dynamics.

In order to provide an overarching background, it is useful to consider the elements of a modelling, or simulation, framework described in Fig. 1. In this framework are six major ingredients. The first are the boundary and initial conditions. For an atmospheric model, boundary conditions include topography, sea surface temperature, land type, vegetation, *etc.*; boundary conditions are generally prescribed from external sources of information.

Simulation Framework
(General Circulation Model, "Forecast")

Boundary Conditions	Emissions, SST, topography, ...	ε
Representative Equations	$DA/Dt = P - LA$	ε
Discrete/Parametrize	$(A_{n+\Delta t} - A_n)/\Delta t = \ldots$	$(\varepsilon_d, \varepsilon_p)$
Theory/Constraints	$\partial u_g/\partial z = -(\partial T/\partial y)R/(Hf_0)$	Scale Analysis
Primary Products (*i.e.* A)	T, u, v,) , H_2O, O_3 ...	$(\varepsilon_b, \varepsilon_v)$
Derived Products (F(A))	Pot. Vorticity, v*, w*, ...	Consistent

$(\varepsilon_b, \varepsilon_v)$ = (bias error, variability error)

Fig. 1. A schematic description of the conceptual elements of an atmospheric model formulation. The boundary conditions include, for example, emissions of trace gases, sea surface temperature (SST), and topography. There are a set of partial differential equations that are the "Representative Equations", *i.e.*, the conservation principles important in the physics (and chemistry, biology, ...) of the atmosphere. Here, there is a generic variable A, and its change with respect to time, t, is equal to its Production, P, minus Loss, which is proportional to a Loss Frequency (L) and the amount of A. These partial differential equations are, usually, specified following scale analysis and approximation for the particular application targeted by the model. The Representative Equations are represented as numerical approximations ("Discrete/Parametrize"), where the index, n, represents a step in time of increment Δt. The "Theory/Constraints" are important to robust model formulation. Here, the *geostrophic approximation* is used as an example. It is important that the numerical methods represent the theoretical constraints that are obtained, for instance, by scale analysis. The "Primary Products" are those products for which there is a prognostic equation. The "Derived Products" are either diagnosed from the primary products or as a function of the primary products. Here potential vorticity and the residual circulation are used as examples. ε represents the error that is present at all stages of the model formulation.

The next three items in the figure are intimately related. They are the representative equations, the discrete and parametrized equations, and the constraints drawn from theory. The representative equations are the continuous forms of the conservation equations. The representative equations used in atmospheric modelling are approximations derived from scaling arguments (see Holton 2004); therefore, even the equations the modeller is trying to solve have *a priori* simplification which can be characterized as errors. The continuous equations are a set of non-linear

partial differential equations. The solutions to the representative equations are a balance amongst competing forces and tendencies.

The discrete and parametrized equations arise because it is not possible to solve the representative equations in analytical form. The strategy used by scientists is to develop a numerical representation of the equations. One approach is to develop a grid of points which covers the spatial domain of the model. Then a discrete numerical representation of those variables and processes which can be resolved on the grid is written. Processes which take place on spatial scales smaller than the grid are parametrized. These approximate solutions are, at best, discrete estimates to solutions of the analytic equations. The discretization and parametrization of the representative equations introduce a large source of error. This introduces another level of balancing in the model; namely, these errors are generally managed through a subjective balancing process that keeps the numerical solution from producing obviously incorrect estimates.

While all of the terms in the analytic equation are potentially important, there are conditions or times when there is a dominant balance between, for instance, two terms. An example of this is *thermal wind balance* in the middle latitudes of the atmosphere (see Holton 2004; see chapter *General Concepts in Meteorology and Dynamics*, Charlton-Perez *et al.*). It is these balances, generally at the extremes of spatial and temporal scales, which provide the constraints drawn from theory. Such constraints are generally involved in the development of conceptual or heuristic models. If the modeller implements discrete methods which consistently represent the relationship between the analytic equations and the constraints drawn from theory, then the modeller maintains a substantive scientific basis for the interpretation of model results.

The last two items in Fig. 1 represent the products that are drawn from the model. These are divided into two types: *primary products* and *derived products*. The primary products are variables such as wind, temperature, water, ozone – parameters that are most often, explicitly modelled; that is, an equation is written for them. The primary products might also be called the resolved or prognostic variables. The derived products are of two types. The first type is those products which are diagnosed from model state variables, often in the parametrized physical processes. The second type follows from functional relationships between the primary products; for instance, potential vorticity (Holton 2004). A common derived product is the budget – the sum of the different terms of the discretized conservation equations. The budget is studied, explicitly, on how the balance is maintained and how this compares with budgets derived directly from observations or reanalysis (chapter *Reanalysis: Data Assimilation for Scientific Investigation of Climate*, Rood and Bosilovich).

In some cases the primary products can be directly evaluated with observations, and errors of bias and variability are estimated. If attention has been paid in the discretization of the analytic equations to honour the theoretical constraints, then the derived products will behave consistently with the primary products and theory. They will have errors of bias and variability, but when a budget is formed from the sum of the terms in the conservation equations, it will balance. That is, the discrete form of the conservation equation is solved. In this case the correlative relation between variables is represented and there is a “physical” consistency.

3 The role of the model in data assimilation

Data assimilation is the melding of observational information with information provided by a model (Daley 1991; Kalnay 2003; Swinbank *et al.* 2003). In assimilation for Earth system science, all types of models, conceptual, statistical, and physical, are used. Models are used in both their prognostic and diagnostic roles. First and foremost in data assimilation, the model provides an estimate of the expected value of the state variables that are observed and assimilated. The discussion, which follows, centres on this role of *state estimation*.

The focus here is on physically based models of the atmosphere formulated in an Eulerian description of the fluid dynamics. Outside of the atmospheric model (or more generally geophysical models) there are other models in the data assimilation system. Notably, because of the complexity of expressing error covariances, these are generally modelled. Also, there are forward and inverse models which transfer quantities between observed quantities, for example radiances observed by a satellite instrument, and geophysical quantities, for example corresponding temperature estimates. These types of models are discussed elsewhere in the book; see, *e.g.*, the chapters in Part B (*Observations*) and Part D (*Chemistry*), and the companion chapters in Part C (*Meteorology and Atmospheric Dynamics*).

A schematic of an assimilation system is given in Fig. 2. This is a *sequential* assimilation system where a forecast is provided to a statistical analysis algorithm that calculates the merger of model and observational information. Some assimilation methods cycle back and forth between these steps to assure maximum coherence. In this figure, errors are specified based on external considerations and methods. There is a formal interface between the statistical analysis algorithm and the model prediction which performs a quality assessment of the information prior to the merger. This interface might also include a balancing process called *initialization* (see Lynch 2003; see chapter *Initialization*, Lynch and Huang). The figure shows, explicitly, two input streams for the observations. The first of these streams represent the observations that will be assimilated with the model prediction. The other input stream represents observations that will not be assimilated. This second stream of observations could be, for example, a new type of observation whose error characteristics are being determined relative to the existing assimilation system.

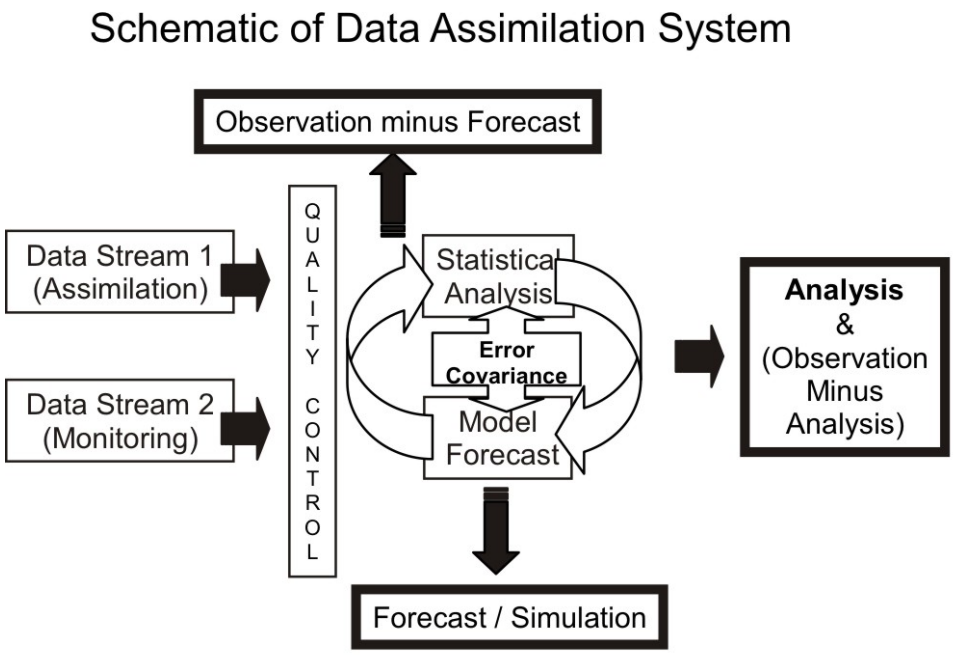

Fig. 2. A schematic of Data Assimilation System. This is a sequential assimilation system where a "Model Forecast" is provided to a "Statistical Analysis" algorithm that calculates the merger of model and observational information using "Error Covariance" information. In this figure, errors are specified based on external considerations and methods. There is a formal interface between the statistical analysis algorithm and the model prediction which performs a quality assessment ("Quality Control") of the information prior to the merger. This interface might also include a balancing process called initialization, which is not explicitly shown. There are two input streams for the observations, "Data Stream 1" and "Data Stream 2". The first of these streams represent the observations that will be assimilated with the model prediction. The other input stream represents observations that will not be assimilated. This second stream of observations could be, for example, a new type of observation whose error characteristics are being determined relative to the existing assimilation system. The products from the system are discussed more fully in the text.

From a functional point of view, the model provides a *short-term forecast* of the expected values of the state variables. This forecast is often called the first-guess, the background, or the prior. The background and the observations are mapped to the same space-time domain where they are compared. The model-provided background is used in the data quality control algorithm, as an objective assessment of the quality of the assimilation system, and as a crucial element of the statistical analysis (see, for example, Dee *et al.* 2001) – see also chapter *Error Statistics in Data Assimilation: Estimation and Modelling* (Buehner). In addition, there may be a formalized process to balance the spatial and temporal attributes of the features represented (or not represented) in both the model and the observations – initialization. In the statistical analysis, observation-based corrections to the background are determined based on

the error characteristics of both the observations and the modelled forecast. These corrections are applied to the background and replace the existing values in the model. These new, corrected values provide the initial conditions for the next model forecast.

The specification of model-error covariances and their evolution with time is a difficult problem. In order to get a handle on these problems it is generally assumed that the observational errors and model errors are unbiased over some suitable period of time, *e.g.* the length of the forecast between times of data insertion. It is also assumed that the errors are in a Gaussian distribution. The majority of assimilation theory is developed based on these assumptions, which are, in fact, not realized. In particular, when the observations are biased, there would the expectation that the actual balance of geophysical terms is different from the balance determined by the model in the assimilation process. Furthermore, since the biases will have spatial and temporal variability, the balances determined by the assimilation are quite complex. Aside from biases between the observations and the model prediction, there are biases between different observation systems of the same parameters. These biases are potentially correctible if there is a known standard of accuracy defined by a particular observing system. However, the problem of bias is a difficult one to address and perhaps the greatest challenge facing assimilation (see Dee 2005). Bias is discussed in chapter *Bias Estimation* (Ménard).

Figure 2 above shows a set of products which comes from the assimilation system. These are (see chapters *Mathematical Concepts of Data Assimilation*, Nichols; *Evaluation of Assimilation Algorithms*, Talagrand):

- ***Analysis:*** The analysis is the merged combination of *model information* and *observational information*. The analysis is the estimate of the state of the system (in this case the atmosphere) based on the optimization criteria and error estimates;
- ***Forecast/simulation:*** The forecast/simulation is a model run that starts from an initial condition defined by the analysis. For some amount of time this model run is expected to represent the state of the system with some deterministic accuracy. For this case the model run is a forecast. After a certain amount of time the model run is no longer expected to represent the particular state of the system; though, it might represent the average state and the variance (*i.e.*, the climate). In this case the model run is simply a simulation that has been initialized with a realistic state estimate at some particular time;
- ***Observation minus forecast increment:*** The observation minus forecast (O-F) increment gives a raw estimate of the agreement of the forecast information (*i.e.,* the first guess) with the observation information prior to assimilation. Usually, a small O-F increment indicates a high quality forecast, and O-F increments are used as a primary measure of the quality of the assimilation. O-F increments are exquisitely sensitive to changes in the system and are the primary quantity used for monitoring the stability and quality of the input data streams. Study of the O-F increment is useful for determining the spatial and temporal characteristics of some model errors;
- ***Observation minus analysis increment:*** The observation minus analysis (O-A) increment represents the actual changes to the model forecast that are

derived from the statistical analysis algorithm. Therefore, they represent in some bulk sense the error weighted impact of the O-F increments. If the assimilation system weighs the observations heavily relative to the forecast, then the O-A increments will have significant differences relative to the O-F increments. The opposite is also true; if the model information is weighed more heavily than the observational information then there will be little change represented by the O-F increments. If either of these extremes are realized the basic assumptions of the assimilation problem need to be reconsidered.

Assimilated data products are often said to be "value-added" (see also chapter *Data Assimilation and Information*, Lahoz *et al.*) The extra value comes from combining two sources of information under the premise that if the error sources are well represented and if the combination process is robust, then there is more information than in either individual source. The two basic sources of information are observed information and model information. Hence, if there is value added to the observed information, then that value comes from the model. Both the prognostic and the diagnostic attributes of the model contribute to the added value.

There are a number of types of information expected to come from the model. The observations are distributed in both space and time. The observations have different attributes; for instance, some observations are point values, while others represent deep layer means. The observations are not continuous; there are spatial and temporal gaps. The model represents the flow of the atmosphere. The model, therefore, takes the information from the observations and propagates that information. This fills in the gaps. Hence, at its most basic level the model is a physically based mapping routine.

From the point of view of information, the model propagates information from observed regions to unobserved regions. If the assimilation is robust, then this greatly improves knowledge of the state in unobserved regions. Further, if at one time a region is not observed and if at a future time the region is observed then, if the model has provided an improved state estimate, then the new observation can better refine the state estimate. That is, there is better use of the observational information. From a different perspective, this comparison of model prediction and observation provides a measure of quality of the assimilation system.

Another function of the model is to transfer information from one observed parameter to other observed parameters. For example, temperature and wind are related to each other. For many years, because temperature observations were by far the most prevalent observation type, temperatures were used to estimate the wind. Elson (1986) compared geostrophic estimates to a number of methods presumed to be more accurate. Such estimates of the ageostrophic wind are crucial, for instance, to weather forecasting and mass transport (see Holton 2004). One place that assimilation has had tremendous impact is in the estimate of mid latitude wind fields, where the geostrophic balance is strong and the wind is strongly influenced by the structure of the temperature field.

Perhaps best viewed as an extension of one observation type influencing another observation type, assimilation also provides estimates of unobserved quantities (see also chapter *Constituent Assimilation*, Lahoz and Errera). One quantity of specific

interest is the vertical component of the wind. Because of the strong hydrostatic stratification of the atmosphere, the vertical component of the wind is three orders of magnitude less than the horizontal components. It is difficult to measure; it remains mostly unmeasured. The vertical wind is, however, critically important to atmospheric physics, linking not only the conservation of thermodynamic energy and momentum, but it also is directly correlated with precipitation and release of latent heat through the condensation of water. Hence a goal of assimilation is to provide meaningful estimates of the vertical component of the wind through the correlated information provided by the temperature and horizontal velocity measurements. There are large errors associated with the estimates of the vertical wind.

The estimate of vertical wind follows from the divergence of the horizontal velocity field. The horizontal velocity is usually a resolved variable, by the nomenclature of Fig. 1, a primary product. Estimates of unobserved quantities also come from the parametrizations used to represent subscale processes. These quantities might include precipitation, clouds, turbulent kinetic energy, or in the case of chemistry-transport models, unobserved chemically reactive constituents or surface emissions. In general, the success of using the assimilation system to estimate unobserved quantities varies widely from one geophysical quantity to another.

Similar in spirit to estimating unobserved quantities, assimilation has the prospect of estimating incompletely observed quantities. An archetypical example is tropospheric ozone. There are many measures of the total amount of ozone in a column above a point on the surface of the Earth. There are also many measures of ozone column above the troposphere. Given the sensitivity of the ozone field to dynamical features in the atmosphere, especially synoptic-scale and planetary-scale waves, the dynamical mapping aspects of assimilation are reasonably expected to offer significant advantage in residual-based estimates of tropospheric ozone (see, for example, Štajner *et al.* 2008).

As will be discussed more fully below, the products from assimilated data sets may not be physically consistent. There are a number of ways to examine the issue of consistency. As mentioned in the discussion of Fig. 1, the equations of motion tell us that there are expected balances between variables. These balances suggest correlative behaviour between variables that reflect the physical connectivity. There is no reason that independent observations and their errors rigorously represent these balances. Similarly, the observations are not required to sample the mass field such that mass is conserved. We look to the model to develop these balances. How well the model does depends on the time-scales that connect the variables and the strength of the expected correlation and the quality of the observations and the model.

Perhaps the best way to look at the consistency problem is whether or not the conservation equation balances. In a well formulated model the conservation equation is solved; there is precise balance. The insertion of data acts like an additional forcing in the conservation equations. In general, this additional forcing will not average to zero over, say, the time-scale between data insertions. Conservation is not obtained. This is an important point to remember as many users of assimilated data sets assume that because they are essentially continuous in space and time, that the variables balance the conservation equation.

The consequences of this violation of conservation propagate through the model. There are fast modes in the model which will balance quickly and accurately. There are slow modes, for instance those balances revealed in the long-term space and time averages suitable for studying the general circulation, which will be influenced by the forcing that comes from the insertion of data. Hence, the assimilated data products might have better estimates than a free running model of primary products like temperature and wind, but the estimates of the derived products such as precipitation and the Eulerian-mean residual circulation (see Holton 2004) may be worse. That is, the analysis increments (*i.e.*, data insertion) are a major part of the forcing. Molod *et al.* (1996) was one of the first to document the representation of the moisture and energy budgets in side-by-side free-running climate simulations and assimilated data using the same predictive model as used in the climate simulation.

4 Component structure of an atmospheric model

This section lays out the component structure of an atmospheric model. The equations of motion for the atmosphere in tangential coordinates using altitude for the vertical coordinate (x, y, z) are given below (see Holton 2004). The first three equations represent the conservation of momentum components. The fourth equation is the mass continuity equation, and the fifth equation is the thermodynamic energy equation. The last equation is the equation of state (see chapter *General Concepts in Meteorology and Dynamics*, Charlton-Perez *et al.*).

$$\left.\begin{aligned}
&\frac{Du}{Dt} - \frac{u\mathrm{v}\tan(\phi)}{a} + \frac{uw}{a} = -\frac{1}{\rho}\frac{\partial p}{\partial x} + 2\Omega \mathrm{v}\sin(\phi) - 2\Omega w\cos(\phi) + \nu\nabla^2(u)\\
&\frac{D\mathrm{v}}{Dt} + \frac{u^2\tan(\phi)}{a} + \frac{\mathrm{v}w}{a} = -\frac{1}{\rho}\frac{\partial p}{\partial y} - 2\Omega u\sin(\phi) + \nu\nabla^2(\mathrm{v})\\
&\frac{Dw}{Dt} - \frac{u^2+\mathrm{v}^2}{a} = -\frac{1}{\rho}\frac{\partial p}{\partial z} - g + 2\Omega u\cos(\phi) + \nu\nabla^2(w)\\
&\frac{D\rho}{Dt} = -\rho\nabla\bullet\mathbf{u}\\
&c_v\frac{DT}{Dt} + p\frac{D\alpha}{Dt} = J \text{ or } \frac{c_p}{T}\frac{DT}{Dt} - \frac{R}{P}\frac{Dp}{Dt} = \frac{J}{T}\\
&p = \rho RT \text{ and } \alpha = \frac{1}{\rho}\\
&\frac{D}{Dt} = \frac{\partial}{\partial t} + \mathbf{u}\bullet\nabla
\end{aligned}\right\} \quad (1)$$

In Eq. 1, t is time; ϕ is latitude; a is radius of Earth, and Ω is the angular velocity of the Earth; g is gravity; ν is a coefficient of viscosity; c_v is specific heat at constant volume and c_p is specific heat at constant pressure; R is the gas constant for air; ρ is

density; T is temperature; and p is pressure. (u, v, w) = (x (zonal), y (meridional), z (vertical)) velocity; J is heating.

In addition, equations are needed which describe the conservation of trace constituents (see chapters in Part C, *Chemistry*). The generic form of these continuity equations are:

$$\frac{DQ_i}{Dt} + Q_i \nabla \bullet \mathbf{u} = P_{Q_i} - L_{Q_i} \tag{2}$$

Where Q_i is the density of a constituent identified by the subscript i; P and L represent the production and loss from phase changes and photochemistry. An equation for water in the atmosphere, $Q_i = Q_{H2O}$, is required for a comprehensive model. For water vapour, the production and loss terms are represented by evaporation and condensation. These are associated with significant consumption and release of heat, which must be accounted for in, J, the heating, *de facto* production and loss term of the thermodynamic energy equation. In general, in the atmosphere below the stratopause, heating due to the chemical reactions of trace constituents is assumed not to impact the heat budget of the atmosphere. It is possible for the spatial distribution of trace constituents, for example, ozone, to impact the absorption and emission of radiative energy; hence, there is feedback between the constituent distributions and diabatic processes in the atmosphere.

Water not only affects the atmosphere through the consumption or release of energy due to phase changes, but also affects the radiative balance of the atmosphere through both the distribution of vapour and through the distribution of clouds. Therefore, it is common in modern models to not only represent water vapour, but also to include an equation for cloud water, $Q_i = Q_{cloud}$, which is partitioned between cloud liquid and cloud ice. The episodic and local scales of the phase changes of water and clouds offer one of the most difficult challenges of atmospheric modelling. This is important for modelling weather, climate, and chemistry.

Due to their impact on both the radiative budget of the atmosphere and formation of cloud water and ice, a set of constituent conservation equations for aerosols is required in a comprehensive atmospheric model. Like water vapour, the spatial and temporal scales of aerosols are often small, below the resolved scales of the model. Again, aerosol modelling provides significant challenges, and they are important for modelling weather and, especially, climate and chemistry.

The equations of motion and a suitable set of constituent continuity equations are the representative equations of the model (see Fig. 1). The equations of motion support many types of dynamical features, for example, waves, such as, Rossby waves, synoptic- or baroclinic-scale waves, gravity waves, Kelvin waves, *etc.* and vortices, such as hurricanes, tornadoes, *etc.* There is spatial and temporal heterogeneity in the forcing terms. Hence, the atmosphere is characterized by complexity, and this complexity is confronted when trying to build a predictive model out of the above equations. Further, the complexity is increased by the fact that discrete numerical representations of the equations of motion support a whole variety of behaviour unique to numerical approximation.

Atmospheric models are usually built from components. There are several useful paradigms for organizing their construction. In the first, the model can be divided into processes, and the solution as a whole is the accumulation of these processes.

This is called "process splitting" and has been discussed in, for instance, Strang (1968), Yanenko (1971) and McCrea *et al.* (1982). Another useful way to look at models is from the perspective of systems engineering, where the whole model system is built from systems of subsystems. This systems approach is useful when formulating strategies for model evaluation and validation; the interacting subsystems determine the performance of the model as a whole. It is, therefore, often difficult to relate model performance to the design characteristics of a particular component.

Recent efforts to organize the modelling community at different laboratories and in different countries have led to the formalization of a component architecture approach to organize the structure. In this approach there are base level components and composited components which rely on the results of the base level components. The interface between the components is formalized by two-way couplers, which transfer the needed information. The model as a whole is a composite of composited, coupled components. Figure 3 shows the Goddard Earth Observing System, version 5 (GEOS-5) component architecture as expressed in the Earth System Modeling Framework (Hill *et al.* 2004; http://www.esmf.ucar.edu/).

Referring to Fig. 3, the box labelled "agcm" represents the atmospheric general circulation model. The components represented here are appropriate for climate and weather. Additional components would be required to resolve the processes above the mesosphere; for example, to support space weather (Toth *et al.* 2005; see also chapter *Assimilation of GPS Soundings in Ionospheric Models*, Khattatov). Below "agcm" are two components which represent the fluid "dynamics" and the "physics." The fluid dynamical part of the model represents both the resolved flow and the drag associated with small (subgrid) scale gravity waves. The dynamics will be discussed more fully below. The terms that form the components of the physics generally represent processes that occur on a scale smaller than resolved; again, they are subgrid. These are often called "parametrizations" (see Fig. 1). A useful, approximate concept is that those components collected under the term physics are treated as occurring only in a vertical column; hence, they can be extracted and tested in one-dimensional column models.[1] Those terms in the "dynamics" are fully three-dimensional; they connect the columns.

From left to right those components which composite as "physics" are as follows. The "surface" box represents, for an atmospheric model, the boundary conditions. Different variables characterize the transfer of momentum, mass, and energy from lakes, ice, ocean, and land (chapters *Ocean Data Assimilation*, Haines; *Land Surface Data Assimilation*, Houser *et al.*, discuss models of the ocean and land, respectively). In this particular model the "land" model is a composite of a vegetation model and a catchment basin hydrology model. The next box to the right, "chemistry," is the interface to chemical production and loss terms which take place point-by-point in both the horizontal and the vertical. This is followed by the accumulation of the processes associated with water and its phase changes, "moist process": clouds, water vapour, liquid water, ice, convection, *etc.* Next are those processes needed to represent the absorption and reflection of both solar and terrestrial (infrared)

[1] See information on column models at the National Center for Atmospheric Research - http://www.ccsm.ucar.edu/models/atm-cam/docs/scam/

"radiation." On the far right is a box labelled as "turbulence". Usually, in atmospheric models there is a separate parametrization which represents the turbulent mixing associated with the planetary boundary layer. More than the other processes in the composite of "physics," the processes in the planetary boundary layer may be connected in the horizontal; that is, they might not fit appropriately into the concept of column physics. As described here, these parametrizations connect momentum, mass, and energy in the vertical; there is transfer between model levels.

Figure 3 is a specific description of an atmospheric model, which relies on the representative equations listed above. In Fig. 1, a conceptual description for building a model was proposed. There are some obvious links. The boundary conditions appear explicitly, and Fig. 3 provides a framework for splitting up the processes of the representative equations. It remains to develop a discrete representation of equations and the identification of the primary and derived products from the model.

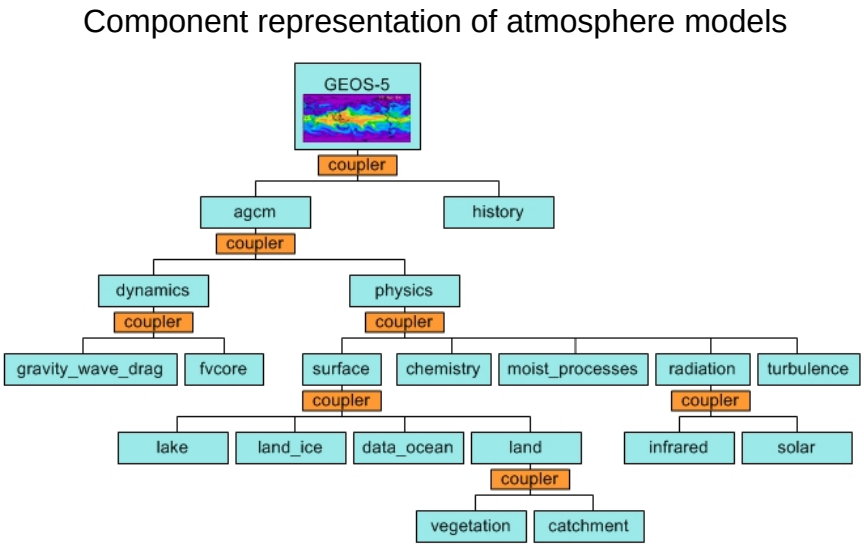

Fig. 3. Earth System Modeling Framework (ESMF) component architecture of the Goddard Earth Observing System, version 5 (GEOS-5) atmospheric model (http://www.esmf.ucar.edu/about_us/). See text for detailed discussion.

As stated at the beginning of the chapter the technical aspects of numerical modelling are left to comprehensive texts such as Jacobson (2005). Some numerical concepts will be developed here to demonstrate the art of model building. The focus will be on the "dynamics" part of the model (see Fig. 3). To demonstrate the concepts consider the thermodynamic energy equation and only the advection of temperature by the horizontal winds

$$\frac{\partial T}{\partial t} + \mathbf{u} \bullet \nabla T = \frac{\partial T}{\partial t} + u\frac{\partial T}{\partial x} + v\frac{\partial T}{\partial y} \tag{3}$$

Attention will be focused on strategies to discretize the advective transport. Figure 4 illustrates the basic concepts. On the left of the figure a mesh has been laid down to cover the spatial domain of interest. In this case it is a rectangular mesh. The mesh does not have to be rectangular, uniform, or orthogonal. In fact the mesh can be unstructured or can be built to adapt to the features that are being modelled. The choice of the mesh is determined by the modeller and depends upon the diagnostic and prognostic applications of the model (see Randall 2000). The choice of mesh can also be determined by the computational advantages that might be realized.[2]

Using the mesh, index points are prescribed to determine location. In Fig. 4a both the advective velocity and the temperature are prescribed at the centre of the cell. In Fig. 4b, the velocities are prescribed at the middle of the cell edges, and the temperature is prescribed in the centre of the cell. There are no hard and fast rules about where the parameters are prescribed, but small differences in their prescription can have large impact on the quality of the estimated solution to the equation, *i.e.*, the simulation. The prescription directly impacts the ability of the model to represent conservation properties and to provide the link between the analytic equations and the theoretical constraints (see Fig. 1; see Rood 1987; Lin and Rood 1996, 1997; Lin 2004). In addition, the prescription is strongly related to the stability of the numerical method; that is, the ability to represent any credible estimate at all.

A traditional and intuitive approach to discretization is to use differences calculated across the expanse of the grid cell to estimate partial derivatives. This is the foundation of the *finite-difference method*, and finite-differences appear in one form or another in various components of most models. Differences can be calculated from a stencil that covers a single cell or weighted values from neighbouring cells can be used. From a numerical point of view, the larger the stencil, the more cells that are used, the more accurate the approximation of the derivative. *Spectral methods*, which use orthogonal expansion functions to estimate the derivatives, essentially use information from the entire domain. While the use of a large stencil increases the accuracy of the estimate of the partial derivatives, it also increases the computational cost and means that discretization errors are correlated across large portions of the domain.

One approach to solving the model equations is to take the continuous representative equations and make term-by-term numerical approximations to variables and their derivatives. There are many approaches to discretization of the dynamical equations that govern geophysical processes (Randall 2000; Jacobson 2005). Given that these equations are, in essence, shared by many scientific disciplines, there are sophisticated and sometimes similar developments in many different fields. One approach that has been recently adopted by several modelling centres is described in Lin (2004). In this approach the cells are treated as *finite volumes* and piecewise continuous functions are fit locally to the cells. These piecewise continuous functions are then integrated around the volume to yield the

[2] Typical mesh sizes at the time of this article are 200 km for climate models down to 20 km for global weather models. Experiments are being run at resolutions as small as ~1 km. Computational resources limit resolution, but also as the resolution becomes finer the foundational assumptions of physical parametrizations must be reconsidered.

forces acting on the volume. This method, which was derived with physical consistency as a requirement for the scheme, has proven to have numerous scientific advantages. The scheme uses the philosophy that if the correlated physics are represented, then the accuracy of the scheme can be robustly built on a physical foundation. In addition, the scheme, which is built around local stencils, has numerous computational advantages.

The variables, *u, v, T,* and Q_{H2O} are often termed the resolved or prognostic variables. Models are often cast into the form that surface pressure, p_{sfc}, is the prognostic equation for conservation of mass. These variables and their gradients are explicitly represented on the grid. The hydrostatic balance is a strong balance in the atmosphere. Most current global models are hydrostatic and do not include a prognostic equation for the vertical velocity, *w*; it is a diagnostic quantity. Cloud resolving models and non-hydrostatic models do resolve the vertical velocity. Non-hydrostatic effects need to be considered if the horizontal resolution is finer than, approximately, 10 km. The importance of the consistent representation of the vertical velocity will be discussed more fully later in the chapter. Reactive constituents and aerosols can add to the list of resolved or prognostic variables. Generally, when the term prognostic is used to describe a variable, it means that a conservation equation has been written for that variable.

In contrast to the resolved or prognostic variables, there are a set of variables which are diagnosed at the grid scale. An example of this is the cloud mass flux between vertical layers of the atmosphere associated with the updrafts and downdrafts of cumulus clouds. There are some variables such as cloud liquid water and cloud ice which may be either explicitly resolved or diagnosed. This is dependent on the spatial resolution of the model. If the model has a resolution that is much larger that the scale of clouds, then cloud water and cloud ice have lifetimes too short to be advected from one grid box to another. In this case, these quantities are diagnosed in the column physics. The terminology prognostic and diagnostic are not precise; they are jargon. Many of the diagnostic variables are, in fact, predicted; therefore, they have the time-change attribute associated with the term "prognostic."

There is also a set of derived products associated with the model (see Fig. 1). For example, it is often productive to interpret atmospheric motions in terms of *vorticity* ($\nabla \times \mathbf{u}$) and *divergence* ($\nabla \bullet \mathbf{u}$). For large-scale, middle latitude dynamics, using pressure as the vertical coordinate, the relative vorticity, ζ, is related to the geopotential, Φ, by the following relationship

$$\zeta = \frac{1}{f}\nabla^2\Phi \tag{4}$$

f is the Coriolis parameter. Geopotential, Φ, is defined as $\Phi(z) = \int_0^z g dz'$, and is the variable which represents the height of a pressure surface when pressure, instead of height, is used as the vertical coordinate. (Geopotential can be related to a parameter with height dimension by dividing it by *g*; this is termed geopotential height.) The ability of the discretization method and the numerical technique to represent relationships such as the one described above is an important and underappreciated aspect of model construction. Lin and Rood (1997) show explicitly both a configuration of variables on the grid and a specification of averaging

techniques that assures that the relationship between geopotential and vorticity is maintained in the discrete equations.[3]

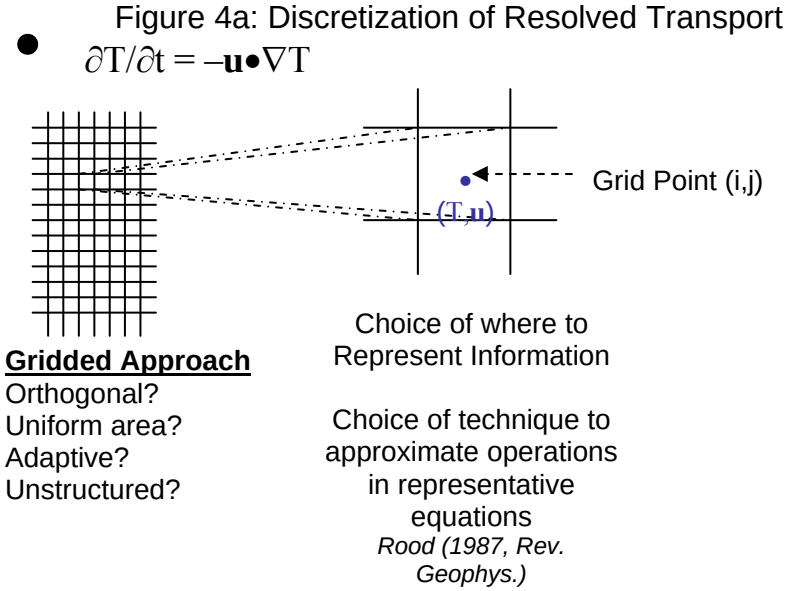

Fig. 4. The use of grids to represent the variables in the governing equations of an atmospheric model. Two examples are given to demonstrate that these choices are up to the model developer. The choices made profoundly impact the characteristics and the performance of the numerical solution. 4a) all variables at grid centre.

Returning to the grids of Fig. 4, the spatial scale of the grid is related to the smallest scales which can be resolved in the model. As guidance, it takes a minimum of 8–10 grid boxes to resolve a wave meaningfully. There is *transport* and *mixing* which occurs at smaller spatial scales. Therefore, for both physical and numerical reasons there is the need to specify a subgrid mixing algorithm. In addition, explicit filters are used to counter errors that arise because of the discrete representation of continuous fields. Subgrid mixing and filters often take on the characteristics of *diffusion*. Their role in atmospheric models is complex and not well quantified. For instance, filters have been used to remove gravity waves in weather forecasting models (see chapter *Initialization*, Lynch and Huang). Given the important role of gravity wave dissipation in climate models, such a filter minimally complicates the quantitative representation of mixing in the atmosphere. [4] There are similar

[3] This constraint, therefore, implicitly links the numerical scheme to large-scale, rotationally dominated flows. As resolution is increased, the divergent component of the flow becomes larger. Therefore, different numerical considerations are expected to be required.

[4] The initialization routine removes scales, for example, gravity waves that are detrimental to the short-term forecast. This is, in part, due to the fact that these scales are not well represented in either the model or the observations. Plus there are spurious sources of small scales related to imbalances due to scale errors and random error. It is incorrect to state that waves at these scales are not important to the atmosphere. They are important to both weather

complications associated with the boundary layer turbulence parametrization. It is important to recognize that the dynamical core of the atmospheric model includes not only an approximation to the resolved advection, but also an algorithm for subgrid mixing, and filters to remedy numerical errors. All these pieces are tightly related to each other; they are often conflated.

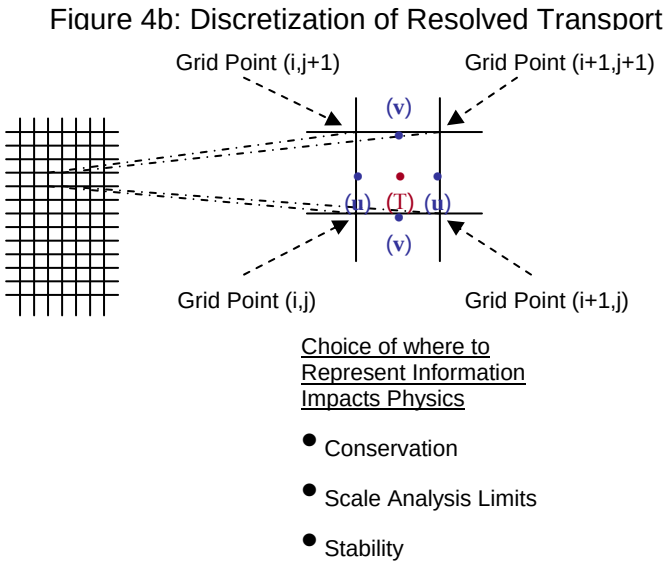

Fig. 4 (continued). 4b) Temperature, T, at grid centre, and velocity (u, v) at grid edges.

As is apparent from the discussion above, there is not a unique or defined way to build an atmospheric model. With the benefit of many years of experience, there are a set of practices which are often followed. These practices evolve as experience is gained. There are decisions in model building which balance known sources of errors. There are decisions simply to give the model viable computational attributes. In many modelling environments there are parts of the code, components, which have not been altered in many years. There remain many open problems which need to be addressed and many paths proposed to address these problems. There are decisions in design and engineering, which contain more than a small element of art.

5 Consideration of the observation-model interface

The interaction between the model and the observations takes place, ultimately, through the *statistical analysis* algorithm (see Fig. 2). There are many aspects of this interface which are described elsewhere in this book (*e.g.* chapters in Part A, *Theory*; chapter *Constituent Assimilation*, Lahoz and Errera; and chapter *Land Surface Data*

---

and climate. The behaviour of motions at these scales changes as the resolution of the model changes.

Assimilation, Houser *et al.*). The model and the observations are formally connected through the observation operator which can be as straightforward as interpolation routines or as complex as radiative transfer models which convert between the model variables and, for instance, the radiances that are observed by satellite instruments (chapter *Assimilation of Operational Data*, Andersson and Thépaut). The model provides information directly to the quality control algorithm. Information from the analysis may be returned to the model through initialization algorithms which strive to filter out dynamical scales which are not important to the short-term forecast. The model and analysis interface can be a one-time passing of information, or there are numerous strategies for cycling the information across the interface to improve the balance in the model. Four-dimensional variational techniques and the incremental analysis update (Bloom *et al.* 1996) are examples of such cycling.

This section focuses on those geophysical variables that serve as the interface variables and where these variables are updated in the component architecture of the model (Fig. 3).

In order for assimilation to be a robust approach to analysing observations, there are a number of attributes that need to be considered. For example, are there enough data to specify the state variables of the atmosphere? If there are not enough observations to specify the state, then the model predictions are not likely to be a real source of information. Alternatively, if there are so many observations that the model is essentially specified by the observations, then a model is not needed for the analysis. The model must be of sufficient quality that it can propagate information from one observing time to the next. The observed variable must have a time-scale, a lifetime, such that information lasts from one observing time to the next; that is, there is the possibility of deterministic prediction. For the assimilation to be robust both the model and observations must contribute to the analysis; the error characteristics from one source or another should not always dominate the other.

For the atmosphere the geophysical parameter with, by far, the most complete coverage is temperature (chapter *The Global Observing System*, Thépaut and Andersson). Since World War II there has been adequate coverage from surface measurements and balloons to support forecast-assimilation systems. With temperature observations it is possible to make credible estimates of winds by both the transference of information through the equations of motion and the propagation of information to chronically under-observed or unobserved regions. There is, also, a substantial body of horizontal wind observations and water vapour observations in the troposphere. Wind observations are especially important to the definition of the atmospheric state.

The temperature and wind observations are both primary products of the model; they are prognostic variables (see Fig. 1). Their spatial and temporal scales are such that their information is utilized and propagated by the model, especially in middle latitudes. From Fig. 3, these variables are directly provided by the dynamical core. The physics of the atmosphere are such that temperature and wind reflect integrated information. Temperatures and winds from assimilated data systems are often excellent estimates of the true state in the free troposphere and lower stratosphere.

Though there is a conservation equation for water vapour and water is a primary, prognostic variable, the water vapour distribution is largely determined in the "moist processes" component of the "physics" (Fig. 3). Because of phase changes, the

spatial and temporal time-scales are both small. The observations of water which come from weather balloons reflect the small spatial scales of water in the atmosphere. These scales are far smaller than those represented by the model grid. The sampling network is not always representative of the state as a whole. The error characteristics are also a strong function of environment, *i.e.*, temperature. Therefore, the representation of water from assimilated data sets is often not of quality for geophysical use (see chapter *Constituent Assimilation*, Lahoz and Errera).

One challenge that must be faced when using the observations from, for instance, balloons, is mapping the observational characteristics to the model. The model might be assumed to represent, for instance, the average temperature in a grid box. The balloon measurement might be appropriately considered a point measurement, at least relative to the model grid. Again, there is a more straightforward relation between modelled and observed temperatures and winds than modelled and observed water vapour.

Since 1979 satellite observations have come to dominate the absolute number of observations used in assimilation systems (see chapter *The Global Observing System*, Thépaut and Andersson). The first variables observed by satellites that were useful for assimilation are temperature and ozone observations. As compared with balloon measurements, satellite information is, often, smeared out over several model grid cells. It took many years to learn how to use satellite temperature observations effectively, and it was determined that mapping of the model information to the radiance space observed by the satellite often facilitated the use of observed information.

Ozone is an interesting variable to examine the model-observation-analysis interface. In some parts of the atmosphere the ozone distribution is determined by advection. Hence, the primary model information would come from the "dynamics" component (Fig. 3) in these regions. Given quality representation of the winds, ozone assimilation works well in these regions (see also chapter *Constituent Assimilation*, Lahoz and Errera). Other parts of the ozone distribution are primarily determined by processes contained in the "chemistry" component (Fig. 3). There are strong spatial gradients in the chemistry; in some places the time-scales are very short. Further, there are strong interdependencies with other gases, aerosols, and temperature (see chapter *Introduction to Atmospheric Chemistry and Constituent Transport*, Yudin and Khattatov). In these regions the assimilation is dominated by the chemical sources and sinks and the advection of ozone from other places and other times has little impact.

The examples outlined above highlight both the ability of the observing system to define the atmospheric state and the ability of the model to use the information. Experience to date shows that if the model information comes from the "dynamics" (Fig. 3) and the spatial gradients are resolved with some accuracy, then the assimilation can be robust. Alternatively, if the model information comes from the "physics" (Fig. 3) and the spatial and temporal scales are small, then the assimilation is likely to have large errors and be of little geophysical value.

Since 1979, and especially since the early 1990s, the amount, the quality, and the span of satellite observations have all grown tremendously. There are many geophysical parameters being measured in the atmosphere, and on the land, ocean, and ice surface. Some of the data, for instance, ocean surface winds have proven to

have large impact in assimilation systems. Other observations have proven more difficult to use. The same ideas as described for atmospheric assimilation hold; the observing system must be able to define the state, and the model able to use the observations. Many of these new observations would have their interface with model through the "physics" component. The spatial and temporal scales of the observations as compared to their representation within the model are often not compatible. The composites that make up the variables in the model are often not what the satellite is observing. The greatest challenges in modelling lie in the representation of "physics," and one of the primary development paths for model-data assimilation should be the development of the model variable–observed variable interface.

6 Physical consistency and data assimilation

Data assimilation has had dramatic impacts on the improvement of weather forecasts (see chapter *Assimilation of Operational Data*, Andersson and Thépaut). There has been wide-scale use of assimilated data sets in climate applications with some great success, as well as identification of a set of problems for which the assimilation analyses are not up to the application. Problems that rely on correlated behaviour of geophysical parameters that are not directly measured, *i.e.*, those estimated by the model parametrizations, are difficult to address. Two examples of such problems, *hydrometeorology* and *constituent transport*, are discussed in chapter *Reanalysis: Data Assimilation for Scientific Investigation of Climate* (Rood and Bosilovich). The rest of this chapter will explore the attributes that distinguish data assimilation for weather from data assimilation for climate.

Weather forecasting is first and foremost concerned with providing quantitative estimates of a set of key variables which define local atmospheric conditions. Successful forecasts benefit from the atmosphere being organized into dynamical structures, *e.g.* waves and vortices, whose propagation is well represented by both the equations of the motion and the numerical methods used to approximate their solution. The observation system can resolve the actual configuration of the dynamical structures at a particular time. In the midst of the forecast-assimilation cycle, the model also has a configuration of the dynamical structures. Through well-defined interfaces, the assimilation routine hands a scale-dependent, balanced initial state to the predictive model, which, in principle, corrects the observed scales in the model state. The model propagates these features forward in time. As is well established, in mid latitudes, during winter, predictions of temperature and wind are useful for several days. When precipitation is associated with large scale dynamics, the prediction of precipitation is useful. Whether that precipitation will be rain or snow is a more difficult prediction. In the tropics and in the summer, when the dynamical features are of smaller scale and the localized moist processes are important to organization of the features, the length of the useful forecast is shorter. Van den Dool *et al.* (1990) discuss measures of forecast skill for high, medium and low temporal variations in the atmosphere, including the tropics and the extra-tropics; Waliser *et al.* (1999) discuss the predictability of tropical phenomena and the relationship to their time-scale.

"Weather" is a subset of the dynamical features that make up the Earth's climate. The role of weather is to transport energy, and consequently, water and other constituents. A good weather forecast is characterized by a good forecast of wind velocity, which is essentially the flux of momentum. Since the ultimate impact of weather on climate is a transport process, climate is directly related to the divergence of fluxes.

The atmosphere, especially at mid latitudes, is dominated by rotational flows close to geostrophic and hydrostatic balance (see Holton 2004). The divergent component of the flow, that responsible for irreversible transport, is an order of magnitude smaller that the rotational part of the flow. Alternatively, the part of the flow most important to the quality of a weather forecast, the rotational part, is an order of magnitude larger than the part of the flow important for a physically consistent representation for climate, the divergent part.[5] While the fluxes are directly related to resolved variables such as the horizontal wind, the divergence of the fluxes are related to transience, non-linearity, and, ultimately, the dissipation of dynamical systems (see Andrews and McIntyre 1978).

A metric of physical consistency is whether or not the conservation equation balances. That is, when all advection, production and loss terms are accounted for, is an accurate estimate of the time rate of change of a quantity realized? If the numerical methods of a model are formulated from a foundation of physical consistency and numerical accuracy, then for a free-running model the budgets should balance. This is not achieved without attention to the details. The effects of corrective filters must be accounted for, and if the filters are a significant part of the conservation equation, then the numerical scheme must be reconsidered.

There is no reason to expect that the disparate observations of the Earth system will satisfy a conservation equation. Therefore, when the observation-based corrections are added to the model equations, imbalance is added. The observations act like a complicated source-sink term. Whether the model and observations in some time averaged sense satisfy a geophysical conservation equation depends upon many things. If there is bias between the model and the observations then the data insertion, the analysis increments, will be a forcing term of constant sign. If there is bias between the models and the observations, then that suggests that the model predicts a different mean state than is observed. If the biases in the resolved variables are "corrected" in the assimilation process, then there is an inconsistency between the values of those "corrected" resolved variables and the values that the physical parametrizations of the model generate. This inconsistency might have little or no effect on the short-term forecast; however, the data insertion is constantly forcing this imbalance, and it will impact those circulations induced by dissipating waves that are important to the climate (see Hoskins *et al.* 1985; Holton *et al.* 1995; Holton

[5] To be clear, it is the estimate of the divergent part of the wind by data assimilation that is responsible for much of the improvement of weather prediction. However, it is true that in many instances that a reasonable short-term forecast at middle latitudes can be realized by the barotropic vorticity equation; hence, the non-divergent geostrophic wind. A good weather forecast might be viewed as, "how fast is the wind blowing in my face?" This is the flux of momentum.

2004). Data insertion will impact where and how resolved scales are formed and dissipated.

In order to demonstrate the impact of data insertion more clearly and more quantitatively, two examples will be developed. The first is based on the assimilation of temperature corrections into the thermodynamic equation. The second is based on the analysis of the vertical velocity in the transformed-Eulerian mean formulation of the equations of motion (see Holton *et al.* 1995; Holton 2004).

Example 1: Observational correction to the thermodynamic equation

To demonstrate the problem of physical consistency more quantitatively, consider the thermodynamic equation written with a simple heating rate and a loss rate proportional to temperature.

$$\frac{DT_f}{Dt} = \frac{\partial T_f}{\partial t} + \mathbf{u} \bullet \nabla T_f = H - \lambda T_f \tag{5}$$

T_f is written to explicitly represent that this is the model forecast temperature. Two cases will be examined.

Example 1, Case 1: In Case 1 assume that the assimilation acts as a forcing term which relaxes the modelled temperature to an analysed temperature determined from the observations. This analysed temperature, for example, might be a gridded, least squares estimate from a variety of observations. The subscript "*a*" represents the analysis of the observational information.

$$\frac{DT_f}{Dt} = H - \lambda T_f - \lambda_a (T_f - T_a) \tag{6}$$

Note that the time-scale, $1/\lambda$ associated with the original equation follows from physical principles. The parameter $1/\lambda_a$ represents the relaxation time-scale associated with the analysis. The time-scale from the analysis follows from the design and assumptions of the data assimilation system (see Swinbank and O'Neill 1994). This appears as an additional forcing term; in the construct of Fig. 3, a "physics" term that is not in the model equations. Therefore, the estimated solution of the equation for T_f evolves in the presence of this additional forcing term.

The equation can be rearranged as

$$\frac{DT_f}{Dt} = H + \lambda_a T_a - (\lambda + \lambda_a) T_f \tag{7}$$

The analysis can be viewed as a change to the loss rate. If the observations are biased relative to the forecast, then the observations are in the time average, a heating term. If the observations are unbiased in the time average, then this heating term averages away; still however, the loss rate is changed. The conservation equation is altered.

Example 1, Case 2: Case 2 implements the data-driven correction to the model equation by correction of the prognostic variable. That is, T_f is replaced with a corrected temperature which is $T_f + \delta T_a$. On substitution into Eq. 5 and re-arranging the terms:

$$\frac{\partial T_f}{\partial t} + \mathbf{u} \bullet \nabla T_f - H + \lambda T_f = -(\frac{\partial(\delta T_a)}{\partial t} + \mathbf{u} \bullet \nabla(\delta T_a) + \lambda(\delta T_a)) \tag{8}$$

The terms on the left side are the model equations, which balance to zero in a free-running simulation. The terms on the right side represent an additional forcing to the model associated with the data insertion.

Under the condition that the model equation is satisfied, the left side of Eq. 8 is zero, then the following equation is obtained.

$$\frac{D(\delta T_a)}{Dt} = -\lambda(\delta T_a) \tag{9}$$

In this case, the increment from each data insertion relaxes to zero with time.

There are two intuitive time-scales to compare with the cooling time-scale $1/\lambda$. The first is the advective time-scale, which is a comparison of physically derived time-scales. This would divide the atmosphere into regions that are dominated by the "physics" and the "dynamics" as well as transition regions where both are important. The second time-scale is the time-scale associated with the frequency of insertion of observational information. From the point of view of "correcting" the model, the balance of these time-scales provides a mechanism to understand the system performance. In this case, following a parcel, the data insertion is a succession of corrections. These corrections relax to zero; hence, the state desired by the model. However, as the system is relaxing, the slow time-scales in the model are being constantly impacted by the observations. This impact lasts far longer than the direct impact of a single observation on the analysis increment. If there is bias between the model and the observations, then this represents a forcing to a new equilibrium state. The data insertion is, also, a source of variability at the time-scales of the observing system.

Another way to think about the role of the model in the assimilation system is to imagine the model as an instrument "observing" a suite of observations that describe the atmosphere, or more generally, the Earth. Some parts of the model directly observe this suite of observations and are well specified. Other parts of the observation suite are indirectly observed, and there are some unobserved variables which are diagnosed as part of the model "circuitry." The influence that the observations have on these indirectly determined and unobserved variables is strongly dependent on the design of the model and how information flows through the model. It is strongly dependent on the logic and construction of the numerical parametrizations.

To illustrate this, consider the component architecture of Fig. 3. Since this figure represents a solution to a set of conservation equations, then the model can be viewed as residing within a closed system. The correction of the model by observed information makes this an open system; the tenets of the conservation principle are no longer true. For the climate, which is the accumulation of the processes represented in the system, this insertion of information (forcing) from outside the system must be assumed to have consequences on the basic physical processes.

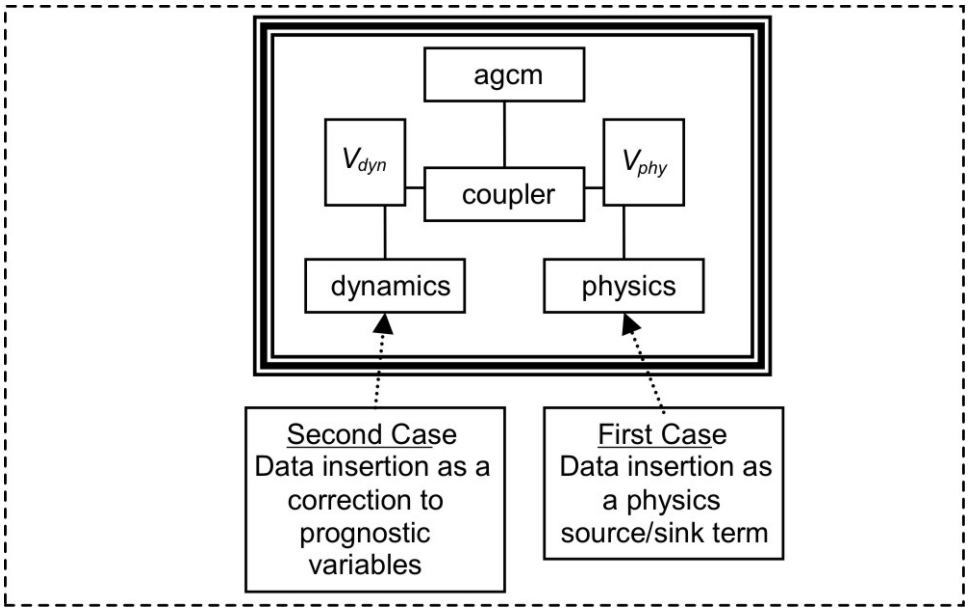

Fig. 5. Schematic of model as a closed system, which accepts forcing from outside the system. Balance here is represented symbolically as if the model was an electrical circuit with a voltage difference across the "coupler." agcm stands for "atmospheric general circulation model".

Figure 5 is a reduced version of Fig. 3; the box around the components shows that the model is a closed system. The coupler determines the transfer of information between the "physics" and the "dynamics." There are numerous time and space scales present in the coupler, as well as those intrinsic to the "dynamics" and the "physics." If the model is viewed as an electronic instrument, as posed above, then there is a network of resistors which determine flow of signal through the system. There are capacitors that represent the time-scales of processes. The balance across the coupler is represented as a voltage difference, V_{dyn}-V_{phy}. The two data insertion scenarios described above are illustrated in Fig. 5. They both explicitly bring in information from outside of the system and to different sides of the coupler. They both would change some aspect of the system, represented symbolically by a change in the voltage difference across the coupler.

Example 2: Horizontal divergence and the vertical wind

The two cases in Example 1, above, used a simple form of the thermodynamic equation. The thermodynamic variables have the property of being in local equilibrium. However their relationship to the wind fields is not always local; the winds are related to the spatially integrated thermodynamic state. There is the possibility of action at a distance as momentum that is dissipated in one region can induce circulations which impact distant regions (see Hoskins *et al*. 1985; Holton *et*

al. 1995; Holton 2004). Therefore, errors in the wind field are expected to impact the analysis in a different way than errors in the temperature field.

As pointed out above, for many years atmospheric observations were dominated by temperature measurements. Winds were estimated from the temperature observations. Assume that the horizontal winds are corrected by the temperature observations by transfer of information through the assimilation system.

$$u_c = u_f + \delta(u(\delta T_a)) \text{ and } v_c = v_f + \delta(v(\delta T_a)) \tag{10}$$

The subscript *c* is the corrected forecast; subscript *f* is the forecast. $u(\delta T_a)$ *and* $v(\delta T_a)$ are the corrections to the velocity field related to the correction in the temperature field that comes from the analysed observations. Through the mass continuity equation, the divergence of the horizontal wind is related to the vertical gradient of the vertical velocity. The divergence of the horizontal wind field, therefore, is the primary quantity that connects the thermodynamic equation and the momentum equations in the atmosphere (see Eq. 1 and Holton 2004). Schematically, the vertical velocity is a key variable connecting the "dynamics" and "physics" components of the model (see Fig. 3). Hence the vertical velocity is a key variable in the coupling of the "dynamics" and the "physics."

As an example, consider large-scale, mid latitude dynamical features. Scale analyses in the atmosphere shows that for these dynamical systems the divergence of the horizontal wind is an order of magnitude smaller than either of the individual terms that make up the divergence. That is, for a representative velocity scale *U* and length scale *L*

$$\frac{\partial u}{\partial x} + \frac{\partial v}{\partial y} \text{ scales as } 0.1\frac{U}{L} \tag{11}$$

The divergence of the assimilation-corrected horizontal wind is

$$\frac{\partial u_c}{\partial x} + \frac{\partial v_c}{\partial y} = \frac{\partial u_f}{\partial x} + \frac{\partial v_f}{\partial y} + \frac{\partial(\delta(u(\delta T)))}{\partial x} + \frac{\partial(\delta(v(\delta T)))}{\partial y} \tag{12}$$

A 10% "correction" in the wind is, potentially, a 100% error in the divergence. It follows that there are similarly large errors in the vertical velocity.

As stated in the previous section, the vertical velocity is usually diagnosed in global models. The vertical velocity can, in general, be diagnosed in two ways. Following Holton (2004), in pressure coordinates, where $\omega \equiv \frac{Dp}{Dt}$ is the vertical velocity,

$$\omega_k(p) = \omega_k(p_{sfc}) - \int_{p_{sfc}}^{p} (\frac{\partial u}{\partial x} + \frac{\partial v}{\partial y})_p \, dp \tag{13}$$

The subscript "*k*" indicates that this velocity is diagnosed from the kinematics of the flow field. p_{sfc} is the surface pressure.

The vertical velocity can also be diagnosed from the thermodynamic equation. Again, in pressure coordinates and following Holton (2004), assuming the diabatic terms, J, can be ignored,

$$\omega_T(p) = S_p^{-1}\left(\frac{\partial T}{\partial t} + u\frac{\partial T}{\partial x} + v\frac{\partial T}{\partial y}\right) \tag{14}$$

Where S_p is the static stability parameter in pressure coordinates. The subscript T indicates this estimate of the vertical velocity is from the thermodynamic equation. In this simplified case the consistency question would ask whether or not these two estimates of the vertical velocity are equal. Experience shows that this is not the case in assimilated data sets, and the errors in the divergence dominate the calculation in ω_k.

To illustrate this problem further, and to make the connection to the climate problem clearer, it is useful to consider the transformed Eulerian-mean formulation of the equations of motion (see Holton *et al.* 1995 and Holton 2004). This formulation has proven a powerful paradigm for understanding the general circulation and constituent transport and is an important example of the physical constraints discussed in Fig. 1. The transformed Eulerian-mean approximates the compensating transport of the waves (represented by a prime) and the Eulerian-mean meridional circulation (represented by an over bar). In this case the diabatic terms cannot be ignored, and one estimate of the residual mean vertical velocity, $\overline{w}^*$, is called the diabatic (subscript d) vertical velocity and should equal

$$\overline{w}_d^*(z) = \frac{R\overline{J}}{HN^2 c_p} \tag{15}$$

For convenience, the vertical coordinate, z, is log pressure-height. N^2 is the square of the buoyancy frequency, and H is a constant scale height ~7 km.

By definition the corresponding analogue to the kinematic estimate is

$$\overline{w}_k^*(z) = \overline{w} + \frac{R}{H}\frac{\partial(\overline{v'T'}/N^2)}{\partial y} \tag{16}$$

In this case the question of consistency comes to whether or not $\overline{w}_k^* = \overline{w}_d^*$ is true.

In general this equality is not realized from assimilated data sets, even in the relatively simple case of the stratosphere (see Schoeberl *et al.* 2003).

Finally, this form of the exposition of the concepts of physical consistency is illustrated in Fig. 6. The value of the vertical velocity presented to the coupler should be the same from the diagnostics via the "physics" and "dynamics." If this is not the case, then the assimilation is physically inconsistent. This particular exposition through the vertical velocity is perhaps the most relevant and important in data assimilation. It is relevant not only to climate and chemistry transport, but to weather. It poses a physical constraint for assimilation – can physically consistent thermodynamic and kinematic vertical velocities from the model be maintained in the assimilation? Or more generally - can the physical balance of the model be

maintained in the presence of the assimilation of observations? This is a formidable task.

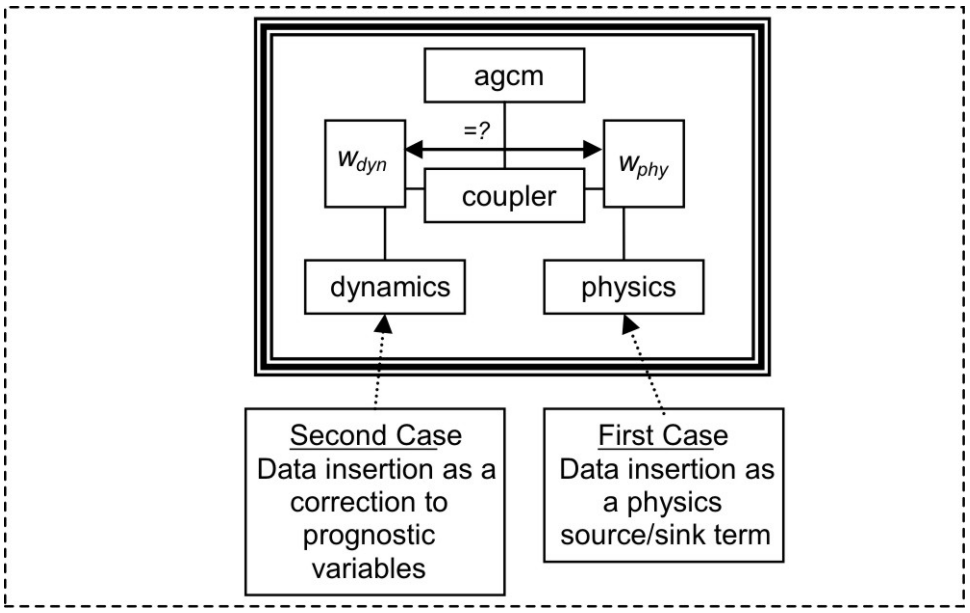

Fig. 6. Schematic of model as a closed system, which accepts forcing from outside the system. Balance here is represented as consistency between vertical velocity estimates coming from the "physics" or "dynamics" components. agcm stands for "atmospheric general circulation model".

7 Summary

This chapter introduced the fundamental ideas that a scientist needs to understand when building or using models in Earth system science research. Rather than focusing on technical aspects of modelling and data assimilation, the chapter focused on a number of underlying principles. These principles, if adhered to, will allow the models and model products to be used in quantitative, data-driven research.

With regards to stand-alone models in the absence of data assimilation, it was emphasized that the underlying physics should be well represented. Specifically, the need to represent correlated behaviour between geophysical parameters was emphasized. A strategy for meeting such a design criteria is to assure that the discrete, numerical approximation to the continuous equations honours the balance conditions that are used in the development of theoretical constructs. This emphasizes "consistency," perhaps at the expense of formal numerical accuracy, as accurate numerical techniques do not guarantee physical consistency. Data assimilation was introduced as the addition of a forcing term to the model that is a correction based on observations. This additional forcing term changes the balance of

forces. Therefore, budgets calculated from assimilated data are not expected, *a priori*, to be robust for geophysical applications.

The role of the model in data assimilation was discussed. It is the assimilation of observational information into the predictive-diagnostic model that sits at the foundation of the value and the potential value of the information produced by data assimilation. In applications ranging from mapping, to improved predictions, to generation of unobserved geophysical variables, data assimilation stands as an essential ingredient of modern Earth system science. The future development of data assimilation includes both the improvement of the models and the better use of information provided by the model. Model improvements include a more robust link in the models between resolved scales and subgrid physical parametrizations. Specifically, with regard to the link to data assimilation, the interface between the subgrid information and the observations needs more attention (see Zhang and Lin 1997). Better use of model information includes using the information developed by the model that connects the correlative physics important to the climate – how is this, first, preserved, then improved, when data is inserted into the model?

Several frames of reference were offered for thinking about models, model construction, and physical consistency. A summary version of these concepts follows. There are many time-scales represented by the representative equations of the model. Some of these time-scales represent balances that are achieved almost instantly between different variables. Other time scales are long, important to, for instance, the general circulation which will determine the distribution of long-lived trace constituents. It is possible in assimilation to produce a very accurate representation of the observed state variables and those variables which are balanced on fast time scales. On the other hand, improved estimates in the state variables are found, at least sometimes, to be associated with degraded estimates of those features determined by long time-scales. Conceptually, this can be thought of as the impact of bias propagating through the physical model (see Dee 2005). With the assumption that the observations are fundamentally accurate, this indicates errors in the specification of the physics that demand further research. The identification, the management, the correction, and the elimination of sources of bias are crucial for improving the physical robustness and self-consistency of assimilated data sets.

Acknowledgments

I thank Minghua Zhang and Ivanka Štajner for reviewing this chapter.

References

Andrews, D.G. and M.E. McIntyre, 1978. Generalized Eliassen-Palm and Charney-Drazin theorems for waves on axisymmetric mean flows in compressible atmospheres. *J. Atmos. Sci.*, **35**, 175-185.

Bloom, S.C., L.L. Takacs, A.M. da Silva and D. Ledvina, 1996. Data assimilation using incremental analysis updates. *Mon. Weather Rev.,* **124**, 1256-1271.

Daley, R., 1991. *Atmospheric Data Analysis*. Cambridge University Press, 457 pp.

Dee, D.P., 2005. Bias and data assimilation. *Q. J. R. Meteorol. Soc.,* **131**, 3323-3342.

Dee, D.P., L. Rukhovets, R. Todling, A.M. da Silva and J.W. Larson, 2001. An adaptive buddy check for observational quality control. *Q. J. R. Meteorol. Soc.,* **127**, 2451-2471.

Elson, L.S., 1986. Ageostrophic motions in the stratosphere from satellite observations. *J. Atmos. Sci.*, **43**, 409-418.

Hill, C., C. DeLuca, V. Balaji, *et al.*, 2004. Architecture of the Earth System Modeling Framework. *Computing in Science and Engineering*, **6**, 18-28.

Holton, J.R., 2004. *An Introduction to Dynamic Meteorology*. Elsevier Academic Press, 535 pp.

Holton, J.R., P.H. Haynes, M.E. McIntyre, *et al.*, 1995. Stratosphere-troposphere exchange. *Rev. Geophys.*, **33**, 403-439.

Hoskins, B.J., M.E. McIntyre and A.W. Robertson, 1985. On the use and significance of isentropic potential vorticity maps, *Q. J. R. Meteorol. Soc.,* **111**, 877-946. (Correction, *Q. J. R. Meteorol. Soc.,* **113**, 402-404, 1987.)

Jacobson, M.Z., 2005. *Fundamentals of Atmospheric Modeling.* 2nd Edition, Cambridge University Press, 813 pp.

Johnson, S.D., D.S. Battisti and E.S. Sarachik, 2000. Empirically derived Markov models and prediction of tropical Pacific sea surface temperature anomalies. *J. Climate,* **13**, 3-17.

Kalnay, E., 2003. *Atmospheric Modeling, Data Assimilation, and Predictability*. Cambridge University Press, 364 pp.

Lin, S.-J., 2004. A "vertically Lagrangian" finite-volume dynamical core for global models. *Mon. Weather Rev.,* **132**, 2293-2307.

Lin, S.-J. and R.B. Rood, 1996. Multidimensional flux-form semi-Lagrangian transport schemes. *Mon. Weather Rev.,* **124**, 2046-2070.

Lin, S.-J. and R.B. Rood, 1997. An explicit flux-form semi-Lagrangian shallow-water model on the sphere. *Q. J. R. Meteorol. Soc.,* **123**, 2477-2498.

Lynch, P., 2003. Introduction to initialization. In *Data Assimilation for the Earth System*, pp 97-11 *(*Eds. R. Swinbank, V. Shutyaev and W. Lahoz, 378 pp), Kluwer Academic Publishers.

McCrea, G.J., W.R. Gooden and J.H. Seinfeld, 1982. Numerical solution of the atmospheric diffusion equation for chemically reacting flows. *J. Comput. Phys.*, **45**, 1-42.

Molod, A., H.M. Helfand and L.L. Takacs, 1996. The climatology of parameterized physical processes in the GEOS-1 GCM and their impact on the GEOS-1 data assimilation system. *J. Climate,* **9**, 764-785.

Plumb, R.A. and M.K.W. Ko, 1992. Interrelationships between mixing ratios of long lived stratospheric constituents. *J. Geophys. Res.,* **97**, 10145-10156.

Randall, D.A. (Ed.), 2000. *General Circulation Model Development: Past, Present, and Future*. Academic Press, 807 pp.

Rood, R.B., 1987. Numerical advection algorithms and their role in atmospheric transport and chemistry models. *Rev. Geophys.,* **25**, 71-100.

Schoeberl, M.R., A.R. Douglass, Z. Zhu and S. Pawson, 2003. A comparison of the lower stratospheric age-spectra derived from a general circulation model and two data assimilation systems. *J. Geophys. Res.,* **108**, Art. No. 4113.

Štajner, I., K. Wargan, S. Pawson, *et al*. 2008. Assimilated ozone from EOS-Aura: Evaluation of the tropopause region and tropospheric columns. *J. Geophys, Res.*, **113**, D16S32, doi:10.1029/2007JD008863.

Strang, G., 1968. On the construction and comparison of difference schemes. *SIAM J. Numer. Anal.*, **5**, 506-517.

Swinbank, R. and A. O'Neill, 1994. A stratosphere troposphere data assimilation system. *Mon. Weather Rev.* **122**, 686-702.

Swinbank, R., V. Shutyaev and W.A. Lahoz (Eds.), 2003. *Data Assimilation for the Earth System*, NATO Science Series: IV: Earth and Environmental Sciences, **26**, Kluwer Academic Publishers, 378 pp.

Toth, G., I.V. Sokolov, T.I. Gombosi, *et al.*, 2005. Space Weather Modeling Framework: A new tool for the space science community. *J. Geophys. Res.*, **110**, A12226, doi:10.1029/2005JA011126.

Trenberth, K.E. (Ed.), 1992. *Climate System Modeling*. Cambridge University Press, 788 pp.

Van den Dool, H.M. and S. Saha, 1990. Frequency dependence in forecast skill. *Mon. Weather Rev.*, **118**, 128-137.

Waliser, D.E., C. Jones, J.-K.E. Schemm and N.E. Graham, 1999. A statistical extended-range tropical forecast model based on the slow evolution of the Madden-Julian Oscillation. *J. Climate*, **12**, 1918-1939.

Yanenko, N.N., 1971. *The Method of Fractional Steps*. Spinger-Verlag, 160 pp.

Zhang, M.H. and J.L. Lin, 1997. Constrained variational analysis of sounding data based on column-integrated conservations of mass, heat, moisture, and momentum: Approach and application to ARM measurements. *J. Atmos. Sci.,* **54**, 1503–1524.