

# Will Climate Change Mathematics (?)

Christopher K.R.T. Jones  
Warwick Mathematics Institute  
University of Warwick, Coventry, UK  
and  
Department of Mathematics  
University of North Carolina  
Chapel Hill, NC, USA

November 20, 2010

## 1 Introduction

It is often said that ours is the century of biology, but that may yet be reconsidered and the century end up being declared that of climate. The challenges due to climate change facing us are enormous and we will almost certainly have to harness all of our scientific knowledge, technological expertise and accumulated experience to respond. The question I wish to address here is what role we, in the mathematics community, should play in this endeavour, in particular in facilitating an understanding of our climate and effecting reliable predictions. That mathematics has a crucial role to play is undisputed. We have only one Earth and thus direct experimentation to test hypotheses is impossible. The only option open to us is the creation of mathematical replicas of the Earth, now called Earth System Models, and run experiments as computational simulations with these models.

It does not take much more than a “back-of-the-envelope” calculation, based on energy balance considerations, to see that the Earth is warming due to the accumulation in the atmosphere of greenhouse gases (GHG). Moreover, it cannot be argued that the increase of GHG is due to anything other than our burning of fossil fuels. For a comprehensive, well articulated and easily readable case I suggest Archer’s book [3] or that of Archer and Rahmstorf [4].

Given such calculations, a natural question is why we should continue to attempt much more complex calculations and predictions as are undertaken by the Intergovernmental Panel on Climate Change (IPCC) through its member climate centres, see [27]. Indeed, in terms of the global temperature average the IPCC does not improve much on these simple energy balance calculations. The main reason for going much beyond the basic energy balance is that the global temperature increase will not be uniformly distributed over the Earth. Due to

polar amplification, which is the tendency for the circulation in the ocean and atmosphere to carry heat away from the tropics, temperature increases in the polar regions will far exceed those near the Equator. Predictions of the climate models are that temperature increases in the Arctic may be closer to  $15 - 20^\circ K$  than the predicted  $2 - 4^\circ K$  in the mid-latitudes, see [27]. Since the polar regions are where most of the the land ice is located, the consequences for sea-level rise are dramatic. This is only the most prominent example of a vast array of regional variations in climate change impacts. Thus, one of the main aims of climate model simulations is to sort out these regional differences, understand their underlying physical causes and make effective and reliable predictions of them.

The use of mathematical models of the Earth puts the discipline of mathematics squarely in the frame as this very process raises a multitude of mathematical questions. It is, in many ways, surprising that mathematicians have not been more involved but I am convinced that this will change over the next decade. Since the climate is presenting us a challenge of such magnitude, societal forces will inevitably bring about a mobilization of the scientific community, including mathematics, to meet that challenge.

We do not, of course, have only models to offer us insight into the workings of the climate. There are also data: lots of data. There are data in abundance from the recent past that give a comprehensive picture of some key aspects of the current climate, for instance surface temperature. There are data from the past century or so and proxy data obtained from paleoclimate records.

There is a disciplinary split between the responsibility for the organization, analysis and inference from models, on the one hand, and data on the other. The former task falls on the shoulders of applied mathematicians, while the latter on statisticians. Models and data are, however, becoming less easy to distinguish from each other. Models themselves are resulting in massive databases which require organization and analysis. At the same time, data are often partially derived from models in subtle ways. For example, the relating of proxy data to the desired variable, such as temperature, essentially constitutes a model of the relevant process. Models are also used to enhance data, such as for capturing the dependence of precipitation fields on topography. A neat distinction, and disciplinary separation, between the analysis of data and that of models is thus becoming less viable.

My interest in this subject was firstly motivated by a commitment to direct my own work toward climate science. As a mathematician, I am also interested in how our subject will grow in response to climate issues. The emergence of mathematical biology over the past few decades has changed applied mathematics in untold ways and taken us in many new directions. It has done so because biological problems demand a different way of mathematical thinking. A critical question for us, the mathematical community, is then: how will climate research shape the work we do? Or, how will the climate change mathematics?

My contention in this essay is that the greatest challenges, as well as the greatest promise for novel and innovative mathematical thinking is at this interface between data and models. I hope to explain why I have come to this

conclusion and what mathematical issues I see arising.

## 2 Climate Models and Prediction

The first step is to see exactly what constitutes climate science at present. I figured that the only sensible way to do this was to visit a major climate centre and talk with climate scientists. So, I packed my bags and went to the National Center for Atmospheric Research (NCAR) in Boulder, Colorado. This worked out well as I had signed up to organise a Theme-of-the-Year at the Institute for Mathematics in Geosciences (IMAGE) at NCAR and I could channel my zeal into organising workshops. Before explaining what I learned from talking with NCAR scientists, it is worth giving a brief description of what constitutes a climate model.

A climate model is a collection of component models and a coupler that moderates information flow between them. The components are each models in their own right of a key part of the climate. They are usually divided up into the: atmosphere, hydrosphere (ocean etc.), cryosphere (land and sea ice), geosphere (land) and biosphere (vegetation, biota etc.). The inflow of radiative energy (short wavelength) from the Sun, the emission of radiation from the Earth (at longer wavelengths) and the interaction of the latter with GHGs is the driver in any climate model. The atmosphere then moves heat around through a variety of processes. At a global scale, the climate model redistributes heat from the equator to the poles.

The earliest models were essentially atmospheric models coupled with the energy balance equations. In these early models, the ocean is represented as a slab and only enters as a heat source, or sink, for the atmosphere. But the ocean itself plays a critical role, albeit on a much longer time scale than the atmosphere, in moving heat energy around the planet. Including a serious ocean model as part of the full climate models has been one of the great achievements of the past decade. Next comes the ice. Its importance in the energy balance is evident as its albedo (propensity for direct reflection of short wavelength solar energy) is quite large. But it also has dynamic importance through its interaction with both the ocean and the atmosphere. Then there is also the land, vegetation and role of biological populations. These pieces are only now beginning to be included in climate models.

The motion of the atmosphere and ocean therefore lie at the heart of a climate model. These are governed by the primitive equations of geophysical fluid dynamics for the velocities of a fluid particle at point  $(x, y, z)$  at time  $t$ ,

which are based on conservation laws, the first of which is that of momentum:

$$\begin{aligned}\frac{du}{dt} + fv &= -\frac{1}{\rho} \frac{\partial P}{\partial x} + F^x, \\ \frac{dv}{dt} - fu &= -\frac{1}{\rho} \frac{\partial P}{\partial y} + F^y, \\ 0 &= -g - \frac{1}{\rho} \frac{\partial P}{\partial z} + F^z,\end{aligned}\tag{1}$$

where  $\frac{d}{dt}$  is a full advective derivative (and hence nonlinear). The Coriolis term is  $f = 2|\Omega| \sin \phi$ , where  $|\Omega|$  is the rotational speed of the Earth and  $\phi$  is the latitudinal angle. The velocities are  $u = \frac{dx}{dt}$  and  $v = \frac{dy}{dt}$ ,  $P$  = pressure and  $\rho$  = density. It is assumed that horizontal motion dominates vertical motion and thus the fluid is in hydrostatic balance. As a result, the  $z$ -equation (third equation above) has reduced to a simple relationship between pressure and density. Note that these equations are given in a Cartesian frame, and thus have to be modified near the poles.

The terms  $F^x$ ,  $F^y$  and  $F^z$  are friction forces on the fluid and will be discussed more below. These equations are then complemented by those expressing conservation of mass and energy. The conservation of mass entails:

$$\frac{\partial \rho}{\partial t} + \nabla \cdot \rho U = 0,\tag{2}$$

where  $U = (u, v, w)$ . For the ocean, which is incompressible, this can be simplified to

$$\frac{\partial \rho}{\partial t} + \nabla \rho \cdot U = 0.\tag{3}$$

The equation for conservation of energy is all-important for climate as it governs the evolution and transport of heat energy. In the following,  $T$  represents potential temperature:

$$\frac{\partial T}{\partial t} + \nabla \cdot TU = F^Q,\tag{4}$$

where  $F^Q$  includes thermal diffusivity (although this may be quite small) and all the radiative forcings, which are effective forcings due to the changes in the way the atmosphere reflects long wavelength radiation given off by the Earth. Greenhouse gas build up is manifest as a radiative forcing. The set of equations is completed by an equation of state of the form:

$$\rho = \rho(P, T, S, \dots),\tag{5}$$

which for the atmosphere is the Ideal Gas Law:  $P = \rho \exp RT$ , and for the ocean usually taken as a linear relationship between  $\rho, T$  and  $S$ . The ocean  $S$  is salinity, while in the atmosphere it is specific humidity and hence measures water vapour content. There will also be a conservation equation for  $S$

$$\frac{\partial S}{\partial t} + \nabla \cdot SU = F^S.\tag{6}$$

We now have a complete set of equations, at least in outline form; for more details, see [11], [22], and [20], . Note that all of the forcing terms  $F = (F^x, F^y, F^z, F^Q, F^S)$  have been left (deliberately) vague. Crafting them is the focus of much of the activity of the climate scientist; more on this later.

Our set of GFD equations are solved numerically on a grid. A typical grid size that is practical for climate runs, which have to be run for long periods of time (decades to centuries), is  $100km \times 100km$ . This seems like a very coarse resolution that risks missing many phenomena; that is right and they must be compensated as will be discussed in the next paragraph. Even with this resolution, a 100 year simulation might take several weeks to complete on a present day supercomputer. Of course, there is considerable work in fitting the whole computational grid together and setting the appropriate boundary conditions. The GFD equations themselves and the computational code to solve them is known as the *dynamical core* of the climate model.

The issue of dealing with effects that occur at scales below the grid is of great current interest. It is widely believed that the greatest weakness of climate models is their not resolving sub grid-scale effects. Since the grid boxes have sides of length around  $100km$ , this includes an enormous range of activity critical to weather and climate, including the behavior of clouds and convection, which is particularly important in the tropics. The strategy used now is to parametrise the sub grid-scale effects and reproduce them through introducing or modifying terms in the equations. In other words, parameters are set in the equations of motion which hopefully capture their impact on the model. An example is the parametrisation of small-scale eddies. This is most easily visualised in the ocean where we know there are all kinds of small scale turbulent eddies. This may occur on scales from meters, or even smaller, up to kilometers. All of this is well below the grid scale. A standard approach is to introduce a viscosity term, called an eddy viscosity, as the main impact on the larger scales is accepted to be diffusive.

Running a numerical experiment now involves setting the model to reflect specific conditions. These conditions will in turn reflect a specific period in history or one anticipated for the future. For instance, of interest might be the pre-industrial period. In this case, the radiative forcing is set to reflect pre-industrial  $CO_2$  emissions. Similarly, for anticipating the impact of future emission scenarios, the model is run with appropriate forcings, see [7].

There are many other specifications needed in the model. This is manifest as setting the values of parameters that appear not only in the main GCM, but also the other components of the full Earth system model, such as of land, ice, vegetation and biology. Some of these may be known a priori and thus pre-set, but others may not be so obvious and need setting through running the model and then “tuning” the parameters and nudge the model to fit the observed or known conditions.

### 3 Data Assimilation (DA)

Climate investigations are thus based on the use of models and data. Observations inform the models through parametrisations, setting of parameters and decisions about which processes to include. The subject that systematically matches information from data and models is data assimilation. Data assimilation was the starting point of my interest in climate. I have been working on assimilating Lagrangian data into ocean models for about eight years (see below for a discussion). So, I started my time of visiting NCAR with the idea that data assimilation must be central to climate science. As I discuss below, the situation is not quite as I expected. But first I think it will be helpful if I explain exactly what data assimilation is and how it works. It is a subject with a strong mathematical dimension and a commensurate need for the intervention of mathematicians in order to meet its challenges.

Assimilating data into models has the aim of achieving optimal use of both models and data. In principle, it offers the most comprehensive and accurate picture of the system state given the available information from model runs and observational data. Given the importance of models and the plenitude of data in the climate, it would seem to be an obvious candidate for an application of the techniques and ideas of data assimilation.

The formalism of DA is best described in terms of a (large) vector  $\sigma \in \mathbb{R}^n$ . The model is then a rule for advancing the state of the system in time

$$\sigma_k = M(\sigma_{k-1}, t_{k-1}). \quad (7)$$

An example would be the 2dimensional (barotropic) GFD equations on a square grid which divides a rectangular domain into  $100 \times 100$  boxes, then

$$\sigma = (u_{11}, v_{11}, \rho_{11}, P_{11}, T_{11}, S_{11}, u_{12}, v_{12}, \dots, u_{21}, v_{21}, \dots, T_{nn}, S_{nn})$$

which has length  $6 \times 10^4$ . The model would then be computational solution to the equations which advances the system from time  $t_{k-1}$  to time  $t_k$ , where  $\sigma_k = \sigma(t_k)$ .

The observations may be of a subset of the components of  $\sigma$ , or complicated combinations of them, and is expressed in terms of an observation operator

$$\eta = H(\sigma) \quad (8)$$

where  $\eta \in \mathbb{R}^m$ . In general,  $m$  will be much smaller than  $n$ . If we, say, observe temperature at certain grid points,  $H$  is then a simple projection operator onto those coordinates. Since there is no reason that measuring devices should lie at computational grid points,  $H$  would usually be a more complicated linear operator that involves an interpolation. It is not hard to conjure up a situation in which  $H$  is nonlinear if variables measured are not a subset of the state variables, for instance as occurs with proxy data.

At this point, DA diverges into the so-called *sequential* and *variational* approaches. In the sequential approach, each time step is treated independently

and an update is formed using the model prediction at that time and any available observational data, again at that time. The formula for the update, often called the analysis and denoted by  $\sigma^a$  is

$$\sigma_k^a = \sigma_k^f + K \left( \eta_k - H(\sigma_k^f) \right) \quad (9)$$

where  $\sigma_k^f$  is the model prediction (forecast) at time  $t_k$ . This is then a natural linear interpolation where the  $n \times m$  matrix  $K$ , called the (Kalman) gain matrix, is to be determined. The sequential method is applied iteratively and so

$$\sigma_{k+1}^f = M(\sigma_k^a, t_k).$$

The key point is to determine  $K$  and, for that, we need to specify our respective confidence in the model and observations. This will be encoded in error covariance matrices:  $B_k$  an  $n \times n$  matrix for the model and  $R_k$  an  $m \times m$  matrix for the observations. The matrix  $K$  is then derived by requiring that  $\sigma_k^a$  minimises the cost function

$$J(\sigma) = \left\langle \sigma - \sigma_k^f, B_k^{-1}(\sigma - \sigma_k^f) \right\rangle + \left\langle \eta_k - H(\sigma), R_k^{-1}(\eta_k - H(\sigma)) \right\rangle. \quad (10)$$

A calculation then shows that

$$K = B_k H^T (R_k + H B_k H^T)^{-1} \quad (11)$$

where I have assumed that the observation operator  $H$  is linear. For an overview, see [17]

It is usually supposed that  $R$  is given as it accounts for human and instrument error, but that the error covariance matrix  $B$  for the model might be determined by the model itself. the simplest case of  $B$  being given (independently of  $k$ ) is called Optimal Interpolation. Kalman filtering, in contrast, evolves the error covariance matrix with the model. The model in this case needs to be linear. If nonlinear, then it must be linearised (this is the Extended Kalman Filter). The now-popular Ensemble Kalman Filter gives a construction of  $B_k$  from (nonlinearly) evolving an ensemble under the model, see [9].

The variational approach takes all of the observational data over a certain time period  $t = t_1, \dots, t_N$  and seeks an optimal state estimate of the initial condition  $\sigma_0$  as in inverse problem. In so-called 4DVAR, it is found as a minimiser of the cost function

$$J(\sigma) = \left\langle \sigma - \sigma_0^*, B^{-1}(\sigma - \sigma_0^*) \right\rangle + \sum_1^N \left\langle \eta_k - H(\sigma), R_k^{-1}(\eta_k - H(\sigma)) \right\rangle, \quad (12)$$

where  $\sigma_0^*$  is an initial estimate and  $B$  the error covariance associated with that estimate. In principle, this cost function is nonlinear because of the observation operator, but is often linearised in order to make the problem tractable. This minimisation is then usually done directly with some appropriate computational

method. This requires extensive software engineering and computation to approximate the minimum of the cost function. For recent reviews, see [15], and [16].

The sequential method might be preferable for short time scale problems where updates are needed regularly. On the other hand, if a reanalysis is being done to gain a better state estimate retrospectively, then an inverse variational approach is more natural. Having said this, the situation in practice is more confused as most weather centres use a variational approach but applied in a sequential manner!

## 4 Bayesian DA

The underlying philosophy of the above approach to data assimilation is that an optimal estimate of the state is being sought. But what if there are multiple minima of the cost function? Do we want just the global minimiser? A probabilist or statistician would look at this situation and say that what we have is really a probability distribution function (PDF) and we are restricting our attention to the mode. The point is that there may be critical information thrown away and the mode may be deceptive in its succinctness.

A Bayesian approach offers a very natural way to formulate DA so that the entire PDF is captured. If we apply it sequentially at each time step, the forecast is replaced by a PDF, the prior; the observation by its likelihood, which is a conditional PDF, and the analysis by the posterior PDF. This latter PDF is found by Bayes' Theorem which states that:

$$\text{Posterior} \propto \text{Likelihood} \times \text{Prior},$$

or, in the notation of Section 3:

$$P(\sigma|\eta) \propto P(\eta|\sigma)P(\sigma). \quad ()$$

An important point is that it is not just the estimates and data-sets ( $\sigma_k^f$  and  $\eta_k$  respectively) that are replaced by these PDFs, but also their error covariances. Indeed, the fact that the errors were represented by covariance matrices meant that we had effectively restricted ourselves to Gaussian PDFs. Bayes Theorem allows these PDFs to be general. If a variational approach is being used then Bayes Theorem is applied once with the prior representing the initial estimate and the likelihood PDF representing the observational error.

There is a direct relationship between this set of PDFs and the cost functions already discussed, see [21], and [1]. The posterior PDF, which corresponds to the analysis state is given by:

$$P(\sigma|\eta) \propto \exp(-J(\sigma))$$

where  $J$  is the cost function associated with the scheme under use.

The Bayesian approach gives a fundamental reframing of the DA problem. It is no longer aimed at getting the "best" estimate but at organising, probabilistically, all of the information coming from the different sources, i.e., data



and models. The result is the assignment of a probability distribution to the state space which reflects the uncertainty of the state in light of the data.

## 5 Climate and DA

It would seem to be the proverbial “no-brainer” that DA would be used extensively in climate: the subject revolves around large computational models and there are lots of data. It is, however, hardly used at all and the argument for this goes something like the following. Climate research is about prediction, the variational method may help to give us a good idea of the past climate but we are interested much more in the future! Moreover, it is not prediction over short time scales like weather where yesterday’s data can be used in a sequential approach to get a better prediction for tomorrow. For climate, it is prediction far into the future that is demanded.

Jeff Anderson is the leader of data assimilation research at NCAR, see

<http://www.image.ucar.edu/staff/jla/>.

He has developed and led an initiative in bringing the Ensemble Kalman Filter to climate and weather problems, see [21] That his efforts have met with success I learnt during the lecture of Marta Vertenstein, who won the 2010 CESM Distinguished Achievement Award at the Community Model meeting, see

<http://www.cesm.ucar.edu/news/awards/vertenstein.html>.

In describing the challenges to the software infrastructure underlying the community model, she cited the incorporation of data assimilation as one of the main four!

As part of the NCAR Theme-of-the-Year, Jeff Anderson and I organised a workshop on DA and climate. Our goals were to (1) Define the specific DA problems that arise in considering climate questions; (2) Assess the ways in which DA is being used effectively right now; and (3) Determine the specific DA challenges that arise in climate. See

<http://www.image.ucar.edu/Workshops/TOY2010/focus02/Agenda.shtml>.

It was evident during our organisation of the meeting and the meeting itself that there are two areas where DA is already having an impact: *paleoclimate* and *ocean*. In both of these, DA is beginning to be used in significant ways and is fast gaining acceptance. I will describe a case study of each to give an idea of some of the gains from using DA and some of the issues that arise.

### 5.1 Paleoclimate

The period between 1790 and 1820 was cold in Western Europe. Known as the Dalton Minimum, it is believed to have been part of the global cooling due to decreased solar activity and volcanic eruptions. But Western Europe was

colder than it should have been and Bjerkenes [6] hypothesised that a secondary effect was driving this regional variation. The Bjerkenes hypothesis is that an anomalous ocean-atmosphere interaction caused polar water to move south, during winter, in the eastern North Atlantic. This then lowered sea surface temperatures (SST) around Western Europe which, in turn, kept the region cool even into the summer months.

Van der Schrier and Barkmeijer [30] quantitatively corroborated Bjerkenes' hypothesis through a wonderfully innovative use of data assimilation. They assimilated sea-level pressure data (SLP) averaged over the period into the ECBilt-Clio intermediate complexity model (these are models based on GFD computations but at *much* coarser resolution than current GCMs). This model includes both active ocean and atmosphere components. The atmospheric part (ECBilt) is a quasigeostrophic (QG) model in which the ageostrophic effects are accounted for through added forcing. They then assess the southerly polar water flow in the NE Atlantic and the resulting SST around Western Europe.

The anomalous atmospheric circulation is characterised by a minimum over Labrador and a maximum over Iceland of the QG streamfunction. After the data assimilation, these are both well reproduced, see Figure 4 in [30]. Bjerkenes gives an argument that explains the origin of the cold water flow into the region around Western Europe. His argument is based on the Sverdrup Principle and the response of the ocean to the anomalous wind-stress curl.

The question posed by van der Schrier and Barkmeijer [30] was whether a free-running model that assimilated data reflecting the anomalous atmospheric circulation would reproduce this effect. The answer they give is strikingly affirmative. Moreover, they resolve another key issue: whether there is any weakening of the Meridional Overturning Circulation (MOC) involved in this scenario. The weakening or even collapse of the MOC is often cited as the usual suspect in a cooling of Western Europe. They showed that the MOC is slightly weakened, but it is not statistically significant.

There are some important points to note here: (1) there is *no* change assumed in the radiative forcing, and (2) without DA of the SLP data, the cold water flow does not occur. Their work makes clear the relationship between the SLP anomaly and the cooling of Western Europe but leaves open whether this is a secondary effect brought about a change in radiative forcing. One could imagine another experiment in which data related to changes in radiative forcing, due to solar activity and volcanic dust uploading, are assimilated into the model. Whether the anomalous atmospheric circulation is created could then be checked!

In summary, the point is that their work shows the close relationship between the anomalous atmospheric circulation and the southerly polar water flow, exactly as Bjerkenes hypothesised. It is important that the experiment is based on a free-running model and so there is no constraint on the atmospheric variability.

Van der Schrier and Barkmeijer use a dedicated assimilation scheme that does not fit neatly into what I have outlined here. This method is based on an application of forced singular vectors, a method due to Barkmeijer et al. [5]

This is a striking application of DA in that it shows how DA can uncover physical effects and relations. Overall, however, there are many issues that arise in DA for paleoclimate, see [33] for some recent developments. The data is sparse, temporally averaged and often has a low signal to noise ratio. Moreover, the timescales in paleoclimate can be long. All of these features pose a challenge to DA.

## 5.2 Ocean

The next IPCC report, entitled AR5 for Assessment Report 5 and due out in 2013-4, will adopt a new focus on decadal predictions. This is a definite change in perspective from the (close to) centennial scale predictions that were targeted in previous IPCC reports. This shift reflects an important change in perspective. Previously, the goal of the IPCC was to promote an understanding about climate change in order to motivate society to act on curbing greenhouse gas emissions. The scientific community, almost unanimously, considers that case closed and the approach is now to lend support to decision makers who must react to the inevitable changes we face over the coming decades. Indeed, based on the current levels of  $CO_2$  in the atmosphere, we are already committed to some warming regardless of future actions. The idea is thus to make predictions that will serve to support and inform those decisions, and the information needed is therefore over a decadal timescale.

It is also widely accepted that the accuracy of decadal predictions hinges on our knowledge of the state of the ocean. Recognizing that getting the ocean right is critical for decadal predictions has led to a renewed interest in ocean data assimilation. While DA has seen its development advance most rapidly in the context of the atmosphere, sustained research efforts on ocean DA have also borne fruit. For instance, the ECCO project at MIT, see

<http://www.ecco-group.org/>,

has led to a new understanding of deep water currents.

The ocean is set apart from the atmosphere for various reasons, for instance it operates on a much longer time scale. Some of the currents that form part of the so-called “ocean conveyor belt” complete their movement of water in centuries, see [34]. Most importantly, the ocean is much harder to observe. Surface properties, for instance sea surface height and temperature, can be obtained from satellite observations, but sub-surface properties and motion are much harder to ascertain. This makes the data that are available and the methods for their assimilation especially important. The measurements coming from instruments that gather subsurface data will thus be important in gaining the most accurate approximation to the state of the ocean as possible for an initialisation of the models.

### 5.3 Lagrangian DA

Dedicated techniques for the assimilation of Lagrangian data into ocean models have been developed recently. These have served to show how optimal use can be made of Lagrangian data in both updating and (re-)analysing ocean models. The challenge posed by Lagrangian data is that it is not presented in the form of state variable information. If we take a velocity field in two-dimensions, for instance a solution to the barotropic version (which just means independent of  $z$ ) of (1):

$$U = (u(x, y, t), v(x, y, t)),$$

then Lagrangian data will come from observations of a trajectory of the dynamical system describing fluid particle motion, i.e.,

$$\begin{aligned}\dot{x} &= u(x, y, t), \\ \dot{y} &= v(x, y, t).\end{aligned}$$

So, given (noisy) observations of  $(x(t), y(t))$  at, say, times  $t_1, t_2, \dots, t_k$ , we would endeavour to reconstruct the velocity field  $(u, v)$ .

This looks like a hopeless task as the velocity fields are functions and the only general information we have about them is that they satisfy the (barotropic) GFD equations. But we should not lose faith as ocean velocity fields are usually dominated by coherent structures. These include currents, the Gulf Stream being one example, and eddies, such as occur in the Gulf of Mexico which can be of the order of  $100km$  across. If the Lagrangian data we obtain nails these structures down, there is reason to believe that we might recover the flow fields from a small number of trajectories. This was the philosophy underlying the papers [19], [24], and [?], and all the simulation experiments have provided evidence that this reasoning is sound.

The most striking success came in the largest model. In [32], we showed even a very small number of trajectories could capture eddies in the Gulf of Mexico. The identical twin experiment (which means two models are run in tandem: one as “truth” or “control” and the other representing our model approximation) involved running a control with a detaching eddy that drifted to the western gulf and a model that was initialised without the eddy. Assimilating even one trajectory that stayed with the eddy led to a flow in which the eddy was evident, see [32].

The importance of this type of Lagrangian data assimilation was underscored in [26] in which it was shown that combining it with a targeted deployment strategy of Lagrangian instruments is particularly fruitful. But the technique is not without its challenges. Breakdown of standard filters, such as the (Extended) Kalman Filter, was shown in [19] to occur when a trajectory passes near a saddle stagnation point of the flow field. We have dubbed this the “saddle effect” and I will discuss some of its implications below. From the mathematical viewpoint, the interesting aspect of Lagrangian DA is the interplay between the DA and the underlying dynamics of the fluid flow.

## 5.4 A new generation of measuring devices

An article in the Guardian on 3 December, 2009 [29] described a “new weapon against climate change..” The “weapon” is new and impressive: it is a sea-glider that propels itself through buoyancy changes with wings that convert vertical motion to horizontal, and its battery, which powers it for forays that can last over several months and a few thousand kilometres, is on a rail and its movement can tilt the glider in order to change direction. But it is an instrument for taking measurements in the ocean not, as perhaps suggested by the headline, for intervening in climate processes.

Glider DA is an area that is unexplored. Of course, the data can be treated as Eulerian (in situ) data at each point it is obtained, but there is reason to believe that the (quasi-)Lagrangian nature of glider data may be significant. I call it quasi-Lagrangian because the glider, in part, uses the underlying fluid flow to move and therefore its position contains information about that flow.

Gliders can be seen then as the modern version of Lagrangian instruments that have been used for some time to gather ocean measurements and have given us considerable insight into ocean phenomena. Among these are floats, which sit at certain depths and move with the flow, and drifters which move on the surface while being drogued by a sail of some kind that sits below the turbulent Ekman layer. The former collect information about their position through sonar and triangulation using sound sources that are placed semi-permanently on the ocean floor. They upload this information by surfacing and communicating with a satellite. The latter can be followed by GPS.

The most recent, and arguably most important, version of these instruments are the ARGO floats. These floats do not use sonar but surface every ten days or so and give profiles of various physical properties (temperature, salinity etc.) from the passage between surfacings. Great international cooperation has led to significant coverage of the World’s oceans with ARGO floats.

Making full use of these rapidly accumulating data in models is an open challenge. Data assimilation schemes are needed which will be tailored to the specific nature of data obtained from Lagrangian instruments.

## 6 Calibration and Tuning

Towards the end of my time visiting NCAR , I was able to attend the annual Community Model meeting in Breckenridge, Colorado, see

<http://www.cesm.ucar.edu/events/ws.2010/Agendas/agenda.pdf>.

This meeting filled in many of the gaps for me and gave me a much deeper and broader understanding of climate models. My understanding was then deepened by the fortunate opportunity I had of interacting with a number of climate scientists at the Newton Institute during the Climate Prediction programme, see

<http://www.newton.ac.uk/programmes/CLP/>.

I have suggested that the interface between data and models is a fertile and important area. Its importance derives from an aspect of climate modeling that is key to the final climate model product but is carried on largely behind the scenes. This is the process of calibration and tuning of a model.

It is natural to think of a climate model as a living entity. Models are constantly being updated: as computers improve their processing speed, resolution can be reduced, and more processes, relating to ice, land, biology etc., can be included. The role of the IPCC, however, makes the reality slightly different. Every seven years or so, the climate centres will suspend model development and submit the current version for the next assessment report. That date is not mandated by the IPCC, but rather the necessity of published results based on the model appearing by a certain date to be counted in the next IPCC report makes such a cut-off inevitable. Thus, when the NCAR Model, or the Hadley Centre Model is referred to, it will usually be the latest version being used for IPCC runs and so will stay fixed for a period of a few years. This punctuation of the modeling process means that models under discussion often do not reflect the very latest developments, but it does help to mitigate the potential for ambiguity and misunderstanding.

The NCAR models that will be used for the IPCC AR5 runs are CCSM4 and CESM1, which were publicly released in 2010. The development of such a code is a mammoth effort involving hundreds of climate scientists, software engineers and computational scientists. A key aspect is the software engineering that facilitates the most efficient interactions between its different components. This received considerable update in CCSM4 and, as a result, its overall structure is quite different from CCSM3. As discussed above, the computational code for the (fluid) dynamical parts of the model form what is called the dynamical core. There are analogous basic model cores for each of the other processes included and a coupler that orchestrates the interactions between the model components. In addition there are parametrisations of unresolved processes. Each of these pieces has been developed, tested and set in the CCSM4. Any changes would now only be incorporated in a next-generation model.

The “model” should be viewed then as a fixed replica of our Earth upon which can be performed (computational) experiments. The different experiments require some flexibility as they will be aimed at different historical conditions. This flexibility will be through the external forcing to which the model is subjected. It will be run under different forcing scenarios that correspond to either different historical periods or conjectured future scenarios. This external forcing will include the solar insolation (although this has not varied to any great degree over the past few centuries) and the all-important radiative forcing, which encodes all the effects of changes in GHG emissions. These are the resettings of the model that will vary from one experiment to another.

There are also many settings of the model made during its formation and which make up part of the final “model” which is used for these experiments. Anyone who has worked with a complex model knows that it takes considerable time and care to get it working “right”. Initial runs will almost always lead to something unexpected. There is usually some debugging of the code, but even

after that there will be considerable necessary tuning of the model. This sounds like a potentially corrupt process. After all, what is the difference between tuning a model to get the answer you want and tuning it to get the “right” answer? The craft of the scientist is to know the difference; but it is a fine line which is one reason for multiple checks.

There are the unresolved processes related to the intrinsic physics of the models which will need “setting”. For instance: eddy diffusivity (discussed above) or gravity wave drag parameters need to be calibrated and determined. The best way to set such parameters would be to go out and directly measure them, but this may not be possible. Their determination would then be achieved through a trial and error process. This is the process of calibration and tuning of the section title.

The procedure for calibration and tuning in climate models relies heavily on the expert judgment of climate scientists. Runs will be carried out and output collected and visualised. For instance, a model will be run under “twentieth century” conditions. The results will then be evaluated against standard observational data sets and key qualitative features of the climate system. Each domain scientist will have their favorite signature effect they are looking for. An oceanographer might look for whether the model gets key features of the major currents correctly: perhaps checking that the separation point of the Gulf Stream from the North American East Coast is somewhere around Cape Hatteras. An ice scientist might look for the thickness of the winter Arctic ice sheet. The judgement made by the climate scientist is based on an understanding of what value or configuration their key signature should have. This understanding comes from accumulated experience with the phenomenon and is ultimately based on observations of that quantity or feature. For instance, as more data on ice sheet thickness comes in, we can make better judgements about whether the GCMs are capturing it well.

After this process of calibration, the model will then be tuned to improve its reproduction of the quantity or effect that has been the subject of calibration. There are many parameters that are subject to tuning; some of these will have emerged during the process of parametrisation of unresolved processes and therefore are not directly measurable physical quantities. Of course, tweaking parameters to achieve agreement on one signature may have a knock-on effect on other quantities and cause them to become “out of tune”. The whole process is then a massive balancing act and it constitutes a lot of the scientific skill and understanding that goes into a model. Much insight into this process in the context of CCSM3 can be gleaned from reading the papers in the dedicated issue of the Journal of Climate, see [10]. See also [14] for a different perspective on model tuning. Statisticians have laid out a perspective for calibration, see [18], and [12].

The end-product that is *the model* includes the result of all this calibration and tuning. This process should be carefully distinguished from the modification of forcings, such as radiative forcing due to GHG build-up that will be modulated to capture different scenarios of future emissions or to capture different time periods. It is helpful to think of the model, such as the CCSM4 of NCAR or

HadCM3 of the Hadley Centre as an immutable entity which encodes all of our best current understanding of the Earth system. Each is a mathematical replica of the Earth that can be used for experimentation: to test hypotheses and develop understanding of the workings of the Earth under different forcing scenarios. Their importance derives from the fact that we cannot run such experiments on the actual Earth and so need a substitute.

There are innumerable science questions that drive climate science and these computational models offer us a “laboratory” for addressing them. To mention two: it is known that the poles will warm at a much faster rate than the tropics, as mentioned earlier, but we do not know the extent of this polar amplification nor what all the mechanisms are that drive it. The models include all the known drivers: Hadley cells in the atmosphere, meridional overturning circulation in the ocean, and yet they appear to underestimate it. So, what else is at work? The Mauna-Loa data set for  $CO_2$  (see [3]) in the atmosphere is the benchmark for amounts of atmospheric  $CO_2$ , but we know that if all the  $CO_2$  coming from fossil fuel emissions accumulated in the atmosphere, the Mauna-Loa record should be even larger. So what happens to the rest of the carbon? Fears are that it is accumulating in the ocean, but this is not conclusively known.

## 7 Prediction and Uncertainty

Although the climate science community sees models as experimental tools for answering science questions, those outside see their main purpose as being for prediction. The public face of climate science is, in other words, made up of the predictions and projections (a term used by the IPCC for predictions under different emission scenarios) made using these models. As we know, there are forces lined up to discredit these predictions in part on the basis of the uncertainty that climate modelers openly attach to their predictions. This politically fraught atmosphere has made it difficult for a properly informed discussion on uncertainty to take place.

If stakeholders are to use model predictions as a basis for decision-making it is very important that we develop the clearest and most comprehensive picture of uncertainty possible. The standard approach to quantifying model uncertainty in climate is to form an ensemble of models which are grounded in common features but with enough differences to create a spread of predictions. That spread is then used to represent the uncertainty. This is the approach of the IPCC AR4 where the set of ensembles is provided by the different participating climate centres. These are often known as “ensembles of opportunities” as no specific effort has been made to create a representative sample with a pre-envisioned variety. Indeed, the spread relies on the fact that the various climate centres have different histories and priorities and work within their own unique political context which demand an emphasis on particular aspects of the climate. An alternative is to create ensembles within a single model by introducing perturbations to the physics. This approach has been pioneered by the climateprediction.net group, see [28], with some very interesting results.



Both of these approaches, however, rely on the model(s) to generate assessments of their own uncertainty and this approach is open to debate. There is an important recognition underlying all of these ideas which is that the uncertainty should be represented by a probability distribution. To give error bounds is just too crude for the same reasons as mentioned earlier. In terms of the PDF, they would become confidence intervals, but if the PDF has a complex multi-modal structure, this is overly simplistic and might be deceptive.

In the last section, I described briefly how the model is put together from the dynamic core, parametrisations and tuning. A reasonable approach to addressing uncertainty would be to assess it for each of these parts of the modeling process as conditional probabilities and then use Bayes' Law ( ) to calculate the cumulative uncertainty. This would require unpacking the model and making separate judgements about the uncertainty of the model and of the data. While this has many disadvantages, in terms of efficiency, by comparison with the ensemble approach used in the context of a fixed and complete model, it gives an ultimately unbiased assessment of uncertainty: information is put into the final model that comes from physics, to which we assign a representational uncertainty, and data, with its human and instrument error to which we assign a certain confidence. The model uncertainty given all available information is then calculated using Bayes' Theorem.

To the best of my knowledge, this comprehensive approach is not taken to uncertainty assessment. Indeed, it would mix together the process of model formation and quantification of its uncertainty. If it were done, however, data assimilation would be the appropriate vehicle for its implementation.

Data assimilation has been used to understand the failings of perturbed models within the climateprediction.net experiment and with striking results. In [23], Rodwell and Palmer show that outliers in a climateprediction.net experiment can be exposed as failing on short time scales through using the data to assess the spread-which is another way of describing data assimilation.

Data assimilation is used by statisticians for model calibration but largely in much simpler systems. It would work in the case of climate models by achieving the tuning mentioned above through a process of assimilating the data that is at the basis of the expert judgements being made in the model tuning. If done in a Bayesian way, the net uncertainty for the "tuned" state would emerge as part of the process. The procedure could also ensure a level of dynamic consistency as well as a balance between the demands of data from different sources.

So, why is this not done? The answer is that data assimilation, in its full Bayesian form, is not even close to being equipped to deal with problems at the scale of climate models.

## 8 Nonlinearity and Dimension

The difficulty in creating a practical data assimilation scheme is dealing with nonlinearity in the model at the same time as its intrinsic high dimension. When the dynamical core of a climate model is discretised on a current generation

grid, the dimension of the resulting equation has an order of at least  $10^6$ . Data assimilation was developed largely for use in Numerical Weather Prediction (NWP) and high dimensional systems were being considered from the start. What I described at the beginning as the applied mathematics perspective in which a “best estimate” is sought deal, at least in principle, relatively easily with this dimension issue. Both the Kalman filter and the variational (least squares) method adapt easily to high-dimensional problems. But they do so by linearising the system at some level. Even the Ensemble Kalman Filter which uses the underlying (nonlinear) flow to generate ensembles makes a linearity assumption at the update step by effectively enforcing a Gaussian straitjacket onto the PDF of the prior. Of course, actually carrying this out raises many challenges as it involves some very high-dimensional numerical linear algebra, and the computations become extremely challenging.

In the dynamic core, nonlinearity is present in the advective term, which is hidden in the total derivative of (1). Nonlinearity will further manifest itself in the many processes included in a full Earth system model. In the Bayesian formulation, there is no linearisation necessary. Although observation and model errors are usually represented by Gaussians, the nonlinear evolution of the model can generate non-Gaussian PDFs (a linear system would preserve the Gaussian structure) and combining PDFs in Bayes’ Theorem has no requirement of Gaussianity.

Nonlinear models can lead to arbitrarily complex PDFs. In [13], a (log-) posterior PDF for the Lorenz equations is shown. The distribution is a jagged mess and is clearly far from Gaussian and it is not evident that a “best estimate” (which would be the mode of the distribution) would be at all useful. Understanding what is happening in such chaotic systems as the Lorenz equations clearly requires a more complete characterisation of the posterior PDF.

The Bayesian formulation can account for nonlinear effects, but that is not to say these complex PDFs can be calculated. There are many statistical methods that have been developed to deal with this issue: Markov-Chain Monte-Carlo methods including Metropolis-Hastings algorithms, Hybrid Monte-Carlo, particle filters, and others are arising all the time. But they have largely proven effective only in low dimensions. The challenges to implementation of these algorithms in the kind of dimensions required by climate models are great and genuine.

As I have already confessed, I came to this subject through looking at what seems to be the very special problem of assimilating Lagrangian data. But I came to realise that it exposes many of the issues arising due to nonlinearity, and does so in a way that may offer lessons for the larger issue of dealing with nonlinearity and DA in global models.. The saddle effect discussed in Section 5.3 causes a breakdown in Gaussianity. This can be easily seen through a thought experiment: imagine passing a unimodal (say Gaussian) PDF of initial conditions toward a saddle point in a low-dimensional system. If the unstable manifold is one-dimensional, it will break into a bimodal (so non-Gaussian) distribution very quickly. Since chaotic systems can be viewed as ones in which repeated passages near saddles occur, it can readily be seen how a PDF like that in [13] for the Lorenz system will come about. Even the nonlinear effect of a

shear can cause filter breakdown, see [2]. For a recent survey on this nonlinearity vs. dimension issue, see [31]

Lagrangian DA also suggests a possible approach to resolving this problem. It creates a natural separation into a low-dimensional piece (the Lagrangian trajectory part) and a high dimensional piece (the Eulerian flow field). A suggestion made by Salman [25] is to treat these separately, using a statistical method on the Lagrangian part and a linearisation based method on the high-dimensional part. A natural question is whether such a strategy could work more generally. In other words: isolate a low-dimensional part that captures the nonlinear aspects of the problem. This is, of course, not a new thought as the desire for reduction to low dimensions has underpinned much mathematical research, particularly in the context of fluids. But, the proposal here is not for a reduction but an operational division, and this may open up new avenues.

## 9 So, What's The Vision?

I have taken it as given that the issues surrounding climate will be a major driving force for scientific research over the coming decades. The question posed here is whether and how the mathematical sciences will be impacted. I have argued that the need to reconcile the data and models available will demand the development of new mathematical ideas and techniques.

Climate raises this issue of optimising the use of data and models in a particularly poignant way as both play an equal and critical role. I had mentioned biology at the beginning of this paper and it is worth thinking why data and model analysis have not become more interwoven under its influence. I suspect that the answer lies in many areas of biology being relatively data-rich and model-poor; I am thinking of genetics as an example. In comparison, climate is model-rich and data-poor. But while models are indispensable, data are equally important as they are what is used to force the models to accord with the actual climate we live in. This is perhaps a manifestation of a phenomenon common to all complex systems, such as the atmosphere-ocean, namely that model replicas cannot be constructed without guidance and validation from measurements.

The mathematical community develops its own mindsets. If we think of three of the main areas that carry the applications of mathematics: partial differential equations, dynamical systems and stochastic processes, each has its own way of looking at a problem and this leads to a particular formulation. The perspective of that area is manifest in a language used to describe the problem that is common to the researchers in the area. This is obviously critical to our functioning as it means we can work together with some mutual understanding. If we did not have such common perspectives and languages, we would have to completely reboot at the start of any communication and would presumably never get beyond the rebooting phase!

Each of the three areas mentioned above are built around mathematical models for relating to the real world. A physical (biological, chemical, sociological etc.) situation is encoded in an ordinary, partial or stochastic differential

equation, and some notion of “solving” the system is sought. Any observations of the physical state being modeled will come in after the fact.

My contention is that this will change over the coming decades and data will start to come into the mathematical formulation of an applied problem at the most fundamental level. If we read a book now on evolution equations, for example, we will see the basic formulation as consisting of the specification of an equation, an initial condition and possibly boundary conditions and external forcing (some of which may be incorporated into the equation). This is an abstract representation of a “model” and an entity that takes on a life of its own, independently of the real world. Applied mathematicians will come back after performing their analysis and relate it to the actual physical situation, but this is usually, at best, in the background during the analysis phase. This abstract formulation is important as it is the starting point of the set of ideas that is built upon it and of our communication with each other.

I am suggesting that this will change and new formulations will include observational data as part of the basic construction. So, the hypothetical book on evolution equations might start by laying out an equation, initial condition etc., *and* a sequence of observations at later times. Luckily, this will not be built in a vacuum as this is already the paradigm of such areas as stochastic control. But if it infiltrates applied mathematics as a whole, it will change the landscape significantly and generate all kinds of new mathematical problems.

Climate science, with its balance of dependence on data and models, will demand of us to make this shift.

## 10 Acknowledgements

The work of the author is supported by the U.S. National Science Foundation under grant DMS-0940363 and also by a Wolfson Research Merit Award from the Royal Society. The work on Lagrangian data assimilation of the author and collaborators has been supported by the Office of Naval Research under N00014-05-1-0791. Much of this work was written while the author was a Visiting Fellow at the Isaac Newton Institute for Mathematical Sciences and he expresses heartfelt thanks for their hospitality and the stimulating environment for development of these thoughts and ideas. Thanks also to the Institute for Mathematics in Geosciences at the National Center for Atmospheric Research for supporting the Theme-of-the-Year on Mathematics and Climate.

I am particularly grateful to Doug Nychka (IMAGE-NCAR) and Andrew Stuart (Mathematics-Warwick) who read earlier versions of this essay and made invaluable comments.

## References

- [1] Apte, A., Stuart, A.M., Voss, J., and Jones, C. “Data assimilation: mathematical and statistical perspectives,” *Int. J. Num. Meth. Fluids* **56**, 1033-

1046 (2008)

- [2] Apte, A., Stuart, A.M., and Jones, C. “A Bayesian approach to Lagrangian data assimilation,” *Tellus A* **60** 336-347 (2008)
- [3] Archer, D. *Global Warming: Understanding the Forecast*, Blackwell, Oxford (2007)
- [4] Archer, D., Rahmstorf, S. *The Climate Crisis: An Introductory Guide to Climate Change*, Cambridge University Press (2009)
- [5] Barkmeijer, J., Iversen, T., and Palmer, T.N., “Forcing singular vectors and other sensitive model structures,” *Q. J. R. Meteorol. Soc.* **129** 2401–2423 (2005)
- [6] Bjercknes, J., “Atmosphere-ocean interaction during the ‘Little Ice Age’,” In: WMO-IUGG Symposium on Research and Development Aspects of Long-Range Forecasting WMO-162. TP. 79, 77–88 Technical Note No. 66 (1965)
- [7] Davidson, O. and Metz, B., “IPCC Special Report on Emissions Scenarios,” see <http://www.ipcc.ch/pdf/special-reports/spm/sres-en.pdf>
- [8] Dijkstra, H., *Dynamical Oceanography*, Springer, Berlin (2008)
- [9] Evensen, G., *Data Assimilation: The Ensemble Kalman Filter*, Springer-Verlag, Heidelberg (2006)
- [10] Gent, P., “Preface to Special Issue on Community Climate System Model (CCSM),” *J. Climate* **19** 2121–2121 (2006)
- [11] Gill, A.E., *Atmosphere-Ocean Dynamics*, Academic Press, San Diego, CA (1982)
- [12] Goldstein, M., and Rougier, J.C., “Reified Bayesian Modelling and Inference for Physical Systems,” *Journal of Statistical Planning and Inference* **139** 1221-1239 (2009)
- [13] Hairer, M., Stuart, A.M., and Voss, J., “Signal Processing Problems on Function Space: Bayesian Formulation, Stochastic PDEs and Effective MCMC Methods,” To appear in *The Oxford Handbook of Nonlinear Filtering*, Crisan, D., and Rozovsky, B. (2009)
- [14] Huybers, P., “Compensation between Model Feedbacks and Curtailment of Climate Sensitivity,” *J. Climate* **23** 3009-3018 (2010)
- [15] Ide, K., and Jones C. (eds.) “Data Assimilation: Special Issue,” *Physica D: Nonlinear Phenomena* **230** (2007)
- [16] Ide, K., “Mathematical Advances in Data Assimilation: Special Issue,” *Monthly Weather Review* **137** (2009)

- [17] Kalnay, E., *Atmospheric Modeling, Data Assimilation and Predictability*, Cambridge University Press (2003)
- [18] Kennedy, M. and O’Hagan, A., “Bayesian calibration of computer models (with discussion),” *Journal of the Royal Statistical Society, Series B.* **63** 425-464 (2001)
- [19] Kuznetsov, L., Ide, K. and Jones, C. “A method for assimilation of Lagrangian data,” *Monthly Weather Review* **131**, 2247-2260 (2003)
- [20] Marshall, J., and Plumb, R.A., “*Atmosphere, Ocean, And Climate Dynamics*,” Elsevier Academic Press, London, UK (2008)
- [21] Nychka, D. and Anderson, J.L., “Data Assimilation” in *Handbook of Spatial Statistics*, Gelfand, A.E., et al. (eds.), 477-494 CRC Press, Boca Raton, USA (2010)
- [22] Pedlosky, J. *Geophysical Fluid Dynamics*, 2nd. edition, Springer-Verlag, New York (1987)
- [23] Rodwell, M.J., and Palmer, T.N., “Using numerical weather prediction to assess climate models,” *Q. J. R. Meteorol. Soc.* **133**, 129–146 (2007)
- [24] Salman, H., Kuznetsov, L., Ide, K., and Jones, C. “A method for assimilating Lagrangian data into a shallow-water equation ocean model,” *Monthly Weather Review* **134**, 1081-1101 (2006)
- [25] Salman, H., “A Hybrid Grid/Particle Filter for Lagrangian Data Assimilation. Part I: Formulating the passive scalar approximation and Part II: Application to a model vortex flow,” *Q. J. R. Meteorol. Soc.*, **134**, 1539-1565 (2008)
- [26] Salman, H., Ide, K. and Jones, C., “Using flow geometry for drifter deployment in Lagrangian data assimilation,” *Tellus A* **60** 321-335 (2008)
- [27] Solomon, S., Qin, D., Manning, M., Chen, Z., Marquis, M., Averyt, K.B., Tignor, M., and Miller, H.L. (eds.) *Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change*, Cambridge University Press (2007)
- [28] Stainforth, D.A., Aina, T., Christensen, C., Collins, M., Faull, N., Frame, D.J., Kettleborough, J.A., Knight, S., Martin, A., Murphy, J.M., Piani, C., Sexton, D., Smith, L.A., Spicer, R.A., Thorpe, A.J., and Allen, M.R., “Uncertainty in predictions of the climate response to rising levels of greenhouse gases,” *Nature* **433**, 403–406 (2005)
- [29] Tremlett, G., “New weapon against climate change on verge of completing first transatlantic voyage,” *The Guardian*, Thursday December 3 (2009)
- [30] van der Schrier, G., and Barkmeijer, J., “Bjerknes’ hypothesis on the coldness during AD 1790–1820 revisited,” *Climate Dynamics* **24**, 355-371 (2005)

- [31] van Leeuwen, P.J, “Nonlinear Data Assimilation in geosciences: an extremely efficient particle filter ,” Q. J. R. Metereol. Soc, In Press, (2010)
- [32] Vernieres, G., Ide, K., and Jones, C. “Capturing eddy shedding in the Gulf of Mexico from Lagrangian observations,” to appear in Physica D (2010)
- [33] Widmann, M., Goosse, H., van der Schrier, G., Schmur, R., and Barkmeijer, J., “Using data assimilation to study extratropical Northern Hemisphere climate over the past millennium,” Clim. Past **6** 627-644 (2010)
- [34] Wunsch, C., *The Ocean Circulation Inverse Problem*, Cambridge University Press (1996)