1

A Data Assimilation Reminder

1.1 Recalling the Basic Idea of Statistical Data Assimilation

This is an expansion of the discussion of data assimilation in Abarbanel (2013). There we further developed a path integral (Abarbanel (2009); Cox (1964)) approach to the subject of data assimilation, which was illustrated by examples from nonlinear electrical circuits, chaotic fluid dynamics, and laboratory neurobiological experiments. The neurobiological experiments were performed in the laboratory of Daniel Margoliash and his students and postdoctoral fellows at the University of Chicago. The designation 'data assimilation' made its first appearance, to my knowledge, within the community of meteorologists working on numerical weather prediction (Anthes (1974); Ghil and Malanotte-Rizzoli (1991); Pires et al. (1996); Kalnay (2003); Lorenc and Payne (2007); Evensen (2009); Reich and Cotter (2015) and climate modeling. It was understood, probably from the outset of these endeavors in the 1950s, that one required knowledge of the state of the earth system, the atmosphere, and the ocean, at some time t_{final} in order to use the equations of fluid dynamics, so-called General Circulation Models, to predict forward in time for $t \geq t_{final}$. The prediction is the validation (or not) of the model proposed to represent the source of the data. Just "fitting" the observed data is a consistency check on the information transfer methods, but one must do more.

The reason one needs prediction, known in Machine Learning as 'generalization,' is that many, usually most, of the model state variables are **unobserved**. One requires them, however, to predict forward for $t \geq t_{final}$, and, as they are unobserved or even unobservable, one cannot measure them directly, so their role in prediction is the only method for probing the accuracy with which we have estimated them via data assimilation.

One problem was, and remains, that we have only an approximate idea what the state of the earth system is at **any** time with enough accuracy and spatial coverage

https://doi.org/10.1017/9781009024846.002 Published online by Cambridge University Press

to have confidence in those predictions. The situation simply got worse when, in 1963, the seminal paper of Ed Lorenz (1963) showed that the intrinsic instabilities of many nonlinear systems, certainly including fluid dynamics on a spatial grid, amplified small errors in initial conditions as well as errors in fixed physical parameters and would lead to exponential growth in those errors. It was hoped that with 'enough' measurements of the state space of the earth system one could pass 'enough' information to the physical dynamical models to rectify the discouraging prospect one faced.

In the literature that I have reviewed (Ghil and Malanotte-Rizzoli (1991); Pires et al. (1996); Kalnay (2003); Lorenc and Payne (2007); Evensen (2009); Reich and Cotter (2015)) the important question of how many measurements are actually required to make accurate predictions is not addressed. We will address this question.

This transfer of information in measured state variables to 'complete' physical models by estimating all the **unmeasured** state variables *and* all the unknown or poorly known time independent parameters of the model acquired the name 'data assimilation.'

In the discussions here and earlier (Abarbanel (2013)), we have called this process of information transfer *Statistical Data Assimilation; SDA* to emphasize its generality across many disciplines where the nonlinear Physics of the problems at hand are important and the value of viewing it as part of considerations in Statistical Physics as developed since the nineteenth century.

The statistical part of the designation SDA comes from unavoidable noise in the measurements and errors in the models. How one represents errors in a model is not at all a settled subject, but some statement must be made, and, naturally, we will do so.

In the process of SDA we require three critical ingredients to transfer information in observations to properties of models proposed to represent the source of that information.

- We should have well curated data. 'Curation' means we should understand not just the binary or ascii numbers presented as 'data,' but we also should understand the instruments used to collect the data. We should have knowledge of the calibration of these instruments and, if at all possible, we should have knowledge of the errors, whatever their source, in these data. We should also know the statistical distribution of these errors. The information here, in addition to the raw ascii numbers, is often called 'metadata', and it is not always available. A good experiment will provide the data and the metadata; be sure to ask for it.

https://doi.org/10.1017/9781009024846.002 Published online by Cambridge University Press

- We will have a **model** of the processes that produced the well curated data we receive. In a physical or biophysical setting, we may have some guidance for the construction of such a model, or we may be proposing a model whose consistency with the data we wish to examine and whose validity we wish to establish.
- We must have a method for transferring the information from the data to the model: this comprises estimating all model variables that are **unobserved** in the measurement processes as well as estimating the value of time independent parameters in the model. Some parameters are known, perhaps from other observations, but we wish to estimate all those that we do not know.

The topics in this book are primarily focused on the third item. We do not provide guidance here to propose and design experiments, in the laboratory or in the field, for every domain of science one might wish to address. We have actually worked with members of the Margoliash neurobiology laboratory at the University of Chicago to design laboratory experiments that produce excellent predictions; examples of this are discussed throughout this book.

Similarly, we do not propose to discuss specific models for all areas of scientific inquiry. We do hope to convey the overall principles and issues in SDA and to illustrate these with examples.

The third item provides a methodology to transfer the information to *a priori* ill-informed properties of the model; in particular, one wishes to estimate time independent parameters and unobserved states. The thrust of this volume is **not** to provide a model, typically in the form of a differential equation for the state variables involved in the processes generating the data. Providing a model for use in SDA rests on the experience and insight of the user, and it is not a button to push called *Give me a model, please* in some package of algorithms. We will formulate a general framework for the operations utilized to transfer the information in the data to properties of the model, once the data are collected and the models are formulated.

Models are, in a sense, the 'art' of data assimilation. It is in this that the skill of the scientist is displayed. It is a matter of insight and some experience to formulate models. The context of SDA is, first of all, to establish whether a proposed model is consistent with the data. Only consistency is possible as we have no knowledge of the unobserved state variables. So while we should be using measurements $\mathbf{y}(\tau_k)$ in the observation window we cannot check if the unobserved state variables are correct, because, well, we do not know them; similarly with the unknown time dependent parameters.

https://doi.org/10.1017/9781009024846.002 Published online by Cambridge University Press

All state variables and all parameters come into play when we want to validate (or not) the model by using it, deterministically or statistically, to predict the behavior of the observed system after $[t_0, t_{final}]$.

1.2 What Is in the Following Chapters?

We ended Abarbanel (2013) with a discussion entitled "Unfinished Business," and we will spare the diligent reader having to go over our state of ignorance those few years ago by recalling what was "unfinished."

We then will recall our notation and formulation of SDA in a section called "Remembrance of Things Path" and identify the items we'll consider in this collection of writings.

This seems a good place to apologize to Proust (Proust (1913)) and all of his dedicated readers for our choice of a pun here. How could one help but do it?

Picking up from the items of *Unfinished Business* we will address how to use the information in the waveforms of the data as a function of time. Then we will apply the ideas to a useful instructional model and to an interesting geophysics problem.

- SDA Variational Principles; Euler-Lagrange Equations for SDA Variational Calculations; Using Waveform Information; Lorenz96 Examples; Lagrangian Drifters and Shallow Water Flows
- Annealing in the Model Precision R_f
- Symplectic Integration and SDA Variational Principles; "Fokker-Planck Equation" for the SDA Standard Model
- Monte Carlo Methods; Metropolis-Hastings – Random Proposals; Hamiltonian Monte Carlo Methods – Structured Proposals or Symplectic Proposals
- SDA and Its Equivalence to Supervised Machine Learning; $\langle A(\mathbf{X})\rangle = \langle -\log P(\mathbf{X}|\mathbf{Y})\rangle$.

Many of these items were not *unfinished* in 2012; indeed, they were not even known as items to require attention at that time.

https://doi.org/10.1017/9781009024846.002 Published online by Cambridge University Press