Computer Simulation, Measurement and Data Assimilation

Wendy S. Parker

(forthcoming in the *British Journal for the Philosophy of Science*)

Abstract

This paper explores some of the roles of computer simulation in measurement. A model-based view of measurement is adopted and three types of measurement – direct, derived and complex – are distinguished. It is argued that, while computer simulations on their own are not measurement processes, in principle they can be embedded in direct, derived and complex measurement practices in such a way that simulation results constitute measurement outcomes. Atmospheric data assimilation is then considered as a case study. This practice, which involves combining information from conventional observations and simulation-based forecasts, is characterized as a complex measuring practice that is still under development. The case study reveals challenges that are likely to resurface in other measuring practices that embed computer simulation. It is also noted that some practices that embed simulation are difficult to classify; they suggest a fuzzy boundary between measurement and non-measurement.

1 Introduction

In recent years, both computer simulation and measurement have received significant attention from philosophers of science. This paper focuses on the intersection of these topics. It is inspired by what seem to be peculiar claims on the part of some scientists. In particular, scientists in various fields have begun to speak of "measuring" and "observing" the world via computer simulation.[1] Computational chemists, for instance, now talk of "observing" molecular properties via simulation (e.g. Gygi and Galli [2005], p.30). The U.S. National Institute of Standards and Technology (NIST) has introduced a "virtual measurement" program, which is investigating the prospects for "measurement *in silico*" (NIST [n.d.]). And some of today's most-used "observational" climate datasets are simulation output. These climate datasets come from a process known as data assimilation, which has been described as a new way to "measure" atmospheric properties (see Ogburn [2013]). In fact, data assimilation results were employed recently to rebut skepticism about thermometer-based estimates of global warming (Compo et al. [2013]).

[1] Often the terms are used interchangeably by practicing scientists. A meteorologist, for example, might refer to the same automated barometer reading as an "observation" of pressure in one instance and as a "measurement" of pressure in another.

Reflecting on these examples, the philosopher of science might recall Dudley Shapere's ([1982]) classic discussion of the concept of observation in science and philosophy. Motivated in part by the curious claims of astrophysicists – in particular their claim to "directly observe" the center of the sun using complex instrumentation located beneath earth's surface – Shapere argued for an extension of the philosophical concept of observation beyond its previous associations with perception, such that there can be observation by or with scientific instruments.

This paper will not argue for another extension of our philosophical concepts. It aims instead (i) to further articulate a contemporary view of measurement, (ii) to show that, on that view, computer simulations can be embedded in measurement practices in such a way that simulation results constitute measurement outcomes and (iii) to examine in some detail the practice of atmospheric data assimilation as a concrete case study. As should become clear, my interest in these matters is primarily epistemological, rather than terminological.

Section 2 brings together ideas about measurement presented recently by Bas van Fraassen ([2008]) and Eran Tal ([2012]). Following van Fraassen, measurement is characterized as an empirical information-gathering activity that locates an entity in a logical space. This locating, however, is understood to involve a form of model-based inference, along the lines suggested by Tal. Applying this view, Section 3 distinguishes three types of measurement – direct, derived and complex – that differ in the layers of inference involved in going from physical states to measurement outcomes. Section 4 turns to computer simulation. I argue that, while computer simulations on their own are not measuring processes, in principle they can be embedded in direct, derived and complex measuring practices in such a way that simulation results constitute measurement outcomes. Section 5 examines atmospheric data assimilation as a concrete case study, ultimately characterizing it as a complex measuring practice that is still under development. The case study reveals challenges related to calibration and interpretation that are likely to resurface in many other measuring practices that embed computer simulation. Section 6 notes that some practices that embed simulation are difficult to classify; they suggest a fuzzy boundary between measurement and non-measurement. Finally, Section 7 offers some concluding remarks.

2 A contemporary view of measurement

Conceptions of measurement range from the extremely permissive to the extremely conservative. Among the former, there are views like Stevens' ([1959]), on which measurement is the assigning of numbers according to a rule. At the other extreme, some physicists seem to hold that the only genuine measurement is fundamental measurement, which is supposed to involve little more than counting unit elements.[2]

A recent, moderate view is that of van Fraassen ([2008]). In general terms, van Fraassen characterizes measurement as an empirical information-gathering activity that locates an object in a logical space. A more detailed characterization is as follows:

[2] See, for example, Ellis ([1966]) for more on fundamental measurement, though he recognizes other forms of measurement too. See Batitsky ([1998]) for an argument that fundamental measurement as some authors conceive of it is a myth.

> "A measurement is a physical interaction, set up by agents, in a way that allows them to gather information. The outcome of a measurement provides a representation of the entity (object, event, process) measured, selectively, by displaying values of some physical parameters that—according to the theory governing this context—characterize that object." (van Fraassen [2008], pp.179-80)

At a physical level, in order for a procedure to be a measuring procedure, the measuring apparatus at the end of the physical interaction(s) should reliably reflect what the measured entity was like (in some respect) at the start of the interaction(s) (van Fraassen [2008], p.150).[3] This is an informal statement of van Fraassen's *criterion for the physical correlate of measurement*. Measurement is not purely physical on his view, though; it is a procedure for producing a selective *representation* of the measured entity. It should be noted here that, in characterizing measurement, van Fraassen is not aiming to offer a definition; rather, he seeks to "come to an understanding of the subject by eliciting its general features and placing them in context" ([2008], p.173, Fn.24)

In doing so, van Fraassen goes on to emphasize that measurement can incorporate theoretical and other calculation:

> "This is an important point: a measurement and its outcome can be complex, and include calculations and input from a model or theory. Such a procedure still fits the general idea of an operation performed so as to create a representation of the object; one that locates it in a certain logical space, with a location that it does not have *a priori*" ([2008], p.177).

For example, suppose a scientist produces a graph of the daily evolution of temperature in a region by making thermometer readings at a variety of stations in the region and then synthesizing the results. According to van Fraassen ([2008], p.166), both the individual station readings and the graph that is produced from those data – what we might call a *data model* – are measurement outcomes.

Tal ([2012]) delves more deeply into the epistemology of contemporary measurement. Drawing on recent work in metrology, he characterizes measurement as a form of model-based inference:[4]

> "…a necessary precondition for the possibility of measuring is the specification of an *abstract and idealized model of the measurement process*. To measure a physical quantity is to make coherent and consistent inferences from the final state(s) of a physical process to value(s) of a parameter in the model." (Tal [2012], p.17)

A model of a measurement process is a representation of how an instrument or apparatus can be used to learn about the value of a parameter; it should take account of both key physical interactions that will take place as well as any subsequent data processing (e.g. correcting, averaging) that will be performed (see Tal [2012], p.19).[5] A model of a measurement process is abstract in the sense that it is a symbolic or mental representation. It is idealized in that it inevitably does not capture the full richness of any actual measurement process. Nevertheless,

[3] This needs to be articulated carefully in the context of quantum mechanics; see van Fraassen ([2008], Ch.6).
[4] Boumans ([2005]; [2006]; [2012]) and Morrison ([2009]) also highlight the roles of models in measuring.
[5] A model of a measurement process will also assume that a particular abstract system (e.g. a numerical scale) is appropriate for representing the physical quantities or qualities of interest. Much previous work in measurement theory was concerned with the conditions under which empirical quantities can be represented with particular types of scale. Such assumptions in models of measuring processes are left implicit in the present discussion, but this is not to suggest that the choice of scale is a trivial matter.

the aim is to construct the model in such a way that its idealizations and simplifications either have an insignificant impact on the results, given the level of accuracy that is desired, or else are corrected for elsewhere in the measurement process.

To illustrate, Tal discusses the use of a torsion pendulum to make a precision measurement of the Newtonian gravitational constant, *G*, via the time-of-swing method. Theoretical analysis indicates that the oscillation frequency of a block-shaped pendulum bob should change with the angle between its equilibrium position and two flanking masses, and that this change should be systematically related to *G* (see Tal [2012], p.156; also Luo et al. [2009]). So we should be able to measure *G* as a function of the change in oscillation frequency. This theoretical analysis, however, makes several idealizing assumptions, such as that the background gravitational field is homogenous, that the pendulum bob is suspended on a perfectly rigid fiber, and so on. In practice, these ideal conditions will not be met and, since scientists are aiming for a very accurate measurement of *G*, their model of the measurement process needs to include steps to correct for many of these deviations from the ideal conditions – for inhomogeneities in the background gravitational field, for thermoelastic changes in the fiber length, etc. The magnitudes of these corrections will have some associated uncertainty, which must be factored in alongside uncertainty from other sources, such as our inability to know the positions and masses of the flanking masses perfectly precisely (see also Luo et al. [2009], Table 1). The magnitude of the total uncertainty is estimated by summing up the uncertainties from each source, if they can be considered independent, or else in a more complex way (Tal [2012], p.157). In the end, the measurement outcome for *G* is *inferred* from the pendulum behavior and takes the form of a best-estimate value plus an associated uncertainty estimate.

Of course, not every measurement involves quantifying component sources of error and uncertainty. The pendulum measurement of *G* involves what Tal, following Marcel Boumans ([2006]), calls *white-box calibration*, in which the measurement process employs a detailed uncertainty budget – a detailed accounting of the different sources of error that will impact results and of uncertainties associated with different components or steps in the measurement process. The individual station temperature measurements in van Fraassen's meteorological example, by contrast, presumably are obtained by taking thermometer indications at face value. The meteorologist still makes an inference from instrument state to measurement outcome – from thermometer state to local temperature value – but here the inference is quite simple: the level of mercury in the thermometer is understood to indicate the best-estimate temperature value, i.e. to be already calibrated, and the uncertainty associated with that estimate is obtained from the documentation supplied by the instrument's manufacturer (e.g. a 2σ accuracy of $\pm$ 0.3°C in normal operating conditions). The manufacturer may have produced this uncertainty estimate not by white-box calibration, but by comparing the instrument's indications with reference standards. This *black-box calibration* "is useful when the behaviour of the device is already well understood and when the required accuracy is not too high" (Tal [2012], p.150).

Tal's analysis of the details of measurement practice suggests that van Fraassen's criterion for the physical correlate of measurement is too demanding. Tal shows that, even in successful measurement, the apparatus often fails to reliably reflect what the measured object was like at the start of the physical interaction; the same state of the measuring apparatus can be mapped to different measurement outcomes, depending on the background conditions,

known interfering factors, etc.[6] In the torsion pendulum case, for instance, the same differences in oscillation frequency would be mapped to different estimates of *G* if the background gravitational field were different, if the apparatus were in the presence of a strong magnetic field, etc. It is only by appeal to a model of the measuring process that we can distinguish "pertinent aspects of the measured objects" from "procedural artifacts" (Tal [2012], p.80).[7]

Otherwise, the accounts of measurement offered by Tal and van Fraassen have a number of features in common. For both, measurement is relative to a theory or model, and the outcome of a measurement is a selective representation of the entity measured. For both, coherence and consistency play important roles in measurement. An inference from a physical state (or set of states) to a measurement outcome should employ assumptions that, at least collectively, cohere with background knowledge (see van Fraassen [2008], p.145; Tal [2012], p.17). Likewise, Tal describes his *robustness criterion* for measurement, according to which measurement outcomes for the same quantity should be statistically consistent with one another once uncertainties are taken into account, as a "methodological explication" of van Fraassen's *coherence constraint* (Tal [2012], Fn.29; see van Fraassen [2008], p.153). Relatedly, neither defends the view that measurement accuracy is a matter of how close an outcome is to a true quantity value; Tal ([2012], p.20) attempts to remain agnostic about the realism of measured quantities, while van Fraassen suggests that measurement outcomes show us "what an object looks like" within the framework of a governing theory (see van Fraassen [2008], p.167).

The view of measurement that I extract from their discussions can be summarized as follows. Measurement is a kind of empirical information-gathering activity, involving physical interaction with the entity measured, which locates the entity in a logical space.[8] Especially in contemporary measurement, this locating activity often involves a form of model-based inference – an inference from the state(s) of one or more physical processes to the value(s) of one or more parameters thought to characterize the entity under study, where this inference is guided by a model of the measuring process. Assumptions of the model should cohere with background knowledge as much as possible – not just with relevant background theory but also with knowledge of interfering factors, limitations of instruments and human perception, etc.[9] The inferred parameter value(s) constitute a selective representation of the entity measured. Measurement outcomes, when complete, include not only a best-estimate value for a parameter but also a well-motivated uncertainty estimate; the latter indicates the degree to which the measuring process is expected to be informative.[10]

---

[6] Likewise, "different procedures that supposedly measure the same quantity often produce inconsistent, and in some cases even completely reversed, 'raw' orderings among objects" (Tal [2012], p.80).

[7] See also Boumans ([2012]); he argues that the homomorphisms required by the representational theory of measurement do not hold for measurement outside the laboratory.

[8] Philosophical work on observation post-Shapere (see, for example, Brown [1987], Kosso [1992]) also has tended to focus on information gathering. Kosso, for example, characterizes observation as "the acquisition of information through interaction with the world" ([1992], p.21).

[9] Though limitations of human observers will not be emphasized in this paper, they should not be overlooked. The eye itself, for instance, can be understood as an optical device whose limitations must be considered just like those of any other scientific instrument (see Brown [1985], pp.498-9). Likewise, cognitive and perceptual biases, such as a tendency to see what one expects to see, can sometimes be significant.

[10] I do not have a detailed account of informativeness. Roughly, a result is informative about *X* (for an agent) if it narrows the range of possibilities for *X* (for the agent) relative to some background set of options.

When measurement is successful, its outcomes are statistically consistent with those obtained in other measuring processes.[11]

3 Three types of measurement

With this view of measurement in mind, this section will distinguish three types of measurement, which I call *direct*, *derived* and *complex* measurement; these differ in the layers of inference involved in going from physical states to measurement outcomes.[12] The types are intended as rough methodological categories; they do not exhaust the types of measurement that might be distinguished, and it is sometimes possible for a measurement process to be classified in more than one way, depending on one's model (see Fn.23).

In introducing these types of measurement, the following terminology will be employed. An *instrument indication* is a physical state of an apparatus used in measuring, such as a pointer position or a digital display showing a numerical readout (Tal [2012], Ch.2). When that indication is 'read' in accordance with conventions of use of the apparatus, the result is a *raw instrument reading*, which assigns a preliminary value to some parameter about which the apparatus, in favorable circumstances, is supposed to be informative.[13] A *measurement outcome* is a selective representation of the system under measurement, inferred from one or more instrument indications; it consists of values for a single parameter (a best estimate plus uncertainty range) or, in some cases, values for multiple parameters (each with a best estimate plus uncertainty range) that together can represent system states, trajectories, fields, flows, and so on.

In *direct measurement*, an instrument indication is produced via a process that involves no explicit symbolic calculation, and the raw instrument reading assigns a preliminary value to the parameter that is ultimately of interest.[14] Calculations might still be performed in inferring a measurement outcome from the raw instrument reading, but they will be calculations that correct for interfering factors or that estimate uncertainties; they will not transform the raw instrument reading into a value for a different parameter. A detailed model of a direct measurement process will represent physical interactions culminating in the instrument indication, including interfering interactions, as well as any subsequent processing

---

Measurement processes are more informative for an agent to the extent that they allow that agent to infer (in a manner consistent with her background knowledge) a narrower range of values for a parameter or, in van Fraassen's terminology, to locate the entity in a smaller region of a theory's logical space. As Tal ([2013]) notes, there is a need for more work on the notion of information employed in discussions of measurement.

[11] Note however that, as scientists seek increasingly accurate measurements, statistical consistency may be achieved and then later break down. Precision measurements of the gravitational constant, *G*, are currently inconsistent (see Quinn et al. [2013]).

[12] I have employed what seem to me apt labels. Some of these labels have been used by previous writers on measurement, but with different meanings. For instance, as will be clear below, by "direct" measurement I do not mean "requiring no other measurement" (Ellis [1960]; Kyburg [1984]). Likewise, my characterization of "derived" measurement is broader than that of Ellis ([1966]) but perhaps narrower than that of Batitsky ([1998]).

[13] This applies to instruments that have conventions of use that associate their readings with physical parameters, e.g. temperatures are 'read' from thermometer indications. If an apparatus used in measurement has no such associated conventions, we can speak instead of what Tal calls "processed indications" ([2012], p.144), which are numbers that represent instrument features but not any physical parameter under measurement.

[14] Instruments used in direct measurement are examples of what Humphreys ([2014]) calls "non-computational instruments".

to be performed to correct for that interference. Parameter values obtained by direct measurement are *direct measurement outcomes*.

For example, an ordinary rain gauge is used to directly measure rainfall depth. A model of the measurement process might assume that the gauge is a perfect collector, except for the interference of wind, which acts to reduce the catch of the gauge. The raw gauge reading assigns a preliminary value to the 'rainfall depth' parameter, but a correction factor – itself a function of average wind speed during the rain event – is then applied in arriving at the final measurement outcome. (Note that measurements of wind speed are thus also required as part of this direct measurement of 'rainfall depth'.) Likewise, an ordinary mercury thermometer is used to directly measure 'ambient temperature', with calibration already performed by the manufacturer.

A second type of measurement is *derived measurement*, which involves at least one additional layer of inference: measurement outcomes are calculated or derived from (in the simplest case) directly-measured values for other parameters, using reliable scientific principles or definitions (see Figure 1).[15] Ordinary derived measurement is *synchronic*: the directly-measured and derived parameter(s) represent conditions for the same time, or for times within the same period. For example, a value for 'relative humidity' at time *t* might be derived, using physical equations or a psychrometric chart, from direct measurement outcomes for 'pressure', 'wet bulb temperature' and 'dry bulb temperature' at time *t*.[16] A detailed model of the measuring process in derived measurement will include assumptions not just about the physical interactions and data processing involved in the direct measurement part of the process, but also about how the directly-measured parameters relate to the derived one(s) that are ultimately of interest. Insofar as that relation (e.g. the formula for calculating 'relative humidity' from the pressure and temperatures) is idealized in ways that will significantly impact the results, the model should also include steps for correcting the preliminary derived results – analogous to a raw instrument reading – as part of a process of calibration.[17]

We can also identify a *diachronic* species of derived measurement, in which derived parameter value(s) represent conditions for times either before or after the time(s) for which the direct measurements are made. For example, if we directly measure the temperature of a small, heated iron ball at time *t* as well as the ambient temperature at that time, we can calculate, using Newton's law of cooling, what the temperature of the iron ball was twenty minutes prior (see Gockenbach and Schmidtke [2009]).[18] As in the synchronic case, a detailed model of the measurement process will include assumptions not just about the direct measurement part of the process, but also about how the directly-measured parameters relate to the derived ones that are ultimately of interest. If that relation (e.g. Newton's law of

[15] In more complicated cases, the parameter values that serve as inputs to the derivation can include derived or complex measurement outcomes.

[16] Tal's ([2012]) torsion pendulum measurement of *G*, discussed above, is another example. In that case the measured parameter is a constant.

[17] The derivation/calculation step is sometimes performed inside an instrument or apparatus, such that the raw instrument reading assigns a value to one or more derived parameters of interest; this assignment may still require correction/calibration if the derivation relies on significantly idealized assumptions.

[18] An empirically-estimated cooling coefficient for iron will be needed for the calculation, but this can be obtained (see, for example, Gockenbach and Schmidtke [2009]). I also assume that the ball has been placed in a room where the ambient temperature is nearly constant over the period of interest.

cooling) is idealized in respects that can be expected to introduce significant errors in results, then the model should also include steps for correcting for some of those idealizations.

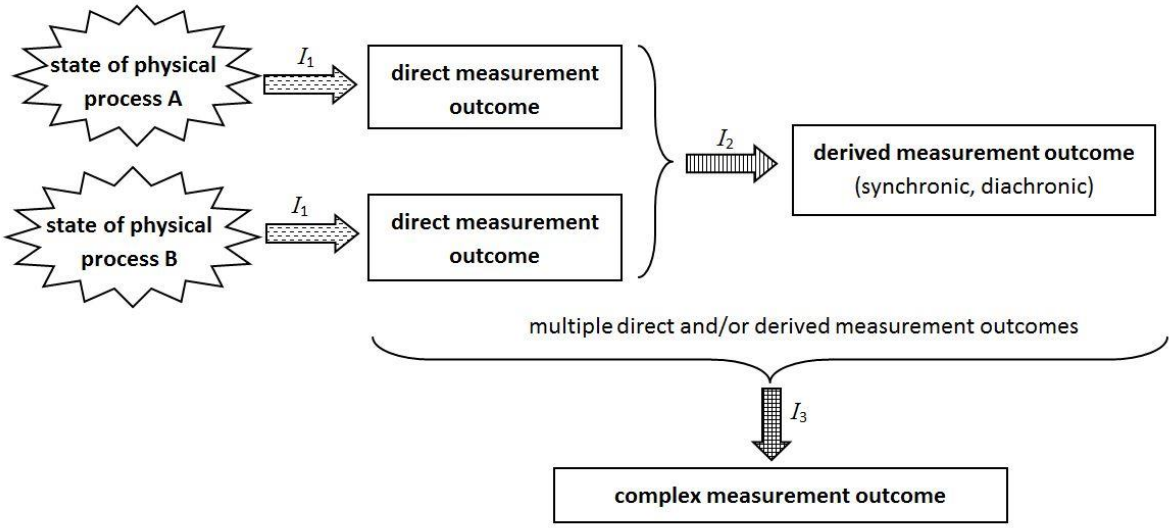

Figure 1. Relations among three types of measurement outcome in simple cases. Arrows represent layers of inference. In direct measurement, outcomes are inferred (I_1) from instrument indications in conjunction with assumptions about interfering factors. In derived measurement, there is a second layer of inference, I_2, involving assumptions about how directly-measured parameters relate to the derived parameters that are ultimately of interest. In complex measurement there is an additional layer of inference, I_3, which appeals to statistical and/or physical assumptions in order to combine information about parameter(s) of interest obtained from multiple direct or derived measurements.

Including diachronic derived measurement – and thus some predictive and retrodictive practices – as a species of measurement is a departure from earlier discussions. However, diachronic derived measurement seems to differ from ordinary (synchronic) derived measurement only in relying on principles or definitions that relate parameters at different times, rather than at the same time. As long as such a process can deliver estimates of parameter values whose accuracy can be specified reliably, it seems reasonable to consider it a species of derived measurement. Yet the following objections might be leveled for cases in which future values of parameters are calculated from past or present values: it does not make sense to "measure" conditions or states of affairs before they occur; given that the future is open, estimates of parameter values for future times differ fundamentally from estimates for the past or present; and estimating parameter values for future times is more aptly characterized as *prediction*. Some of these objections, particularly the second but perhaps also the first, do have some force. So perhaps we should limit measurement to parameters that represent properties at times that are already past. In that case, we cannot measure parameter values for future times, but a result that is now a prediction may later be considered a derived measurement outcome (e.g. if the conditions assumed when making the prediction closely enough match the conditions that occur in the interim).

A third type of measurement is *complex measurement*, which involves making multiple direct and/or derived measurements and then using their results together, with the aim of obtaining a measurement outcome that is more informative than could be achieved using just a subset of those results.[19,20] The measurement process thus involves an additional layer of inference beyond those involved in making any direct or derived measurements that figure in the process; this additional layer of inference involves assumptions about how to combine information about the parameter(s) of interest obtained via multiple measurements. Complex measurement practices come in a variety of forms, not all of which will be discussed here. Some involve combining multiple measurements of the same parameter; others involve multiple measurements that serve as samples from which to estimate an aggregate or population-level parameter; still others involve a set of measurements to which structure is added to arrive at a data model. Combinations of these are also possible, as we see below.

A simple example of the first form is a measurement process that assigns values to a 'weight' parameter by weighing a person on three different but calibrated scales, averaging the raw instrument readings and calculating the uncertainty in the result using the accuracy information provided by the scale manufacturers, under the assumption that the individual measurements are independent. A detailed model of this sort of complex measurement process thus includes assumptions not only about the individual direct measurements, but also about relations among them (e.g. that they are independent). If assumptions about those relations are significantly idealized, the model may also need to include steps to correct for those idealizations.

A more complicated example, involving both the second and third forms of complex measurement mentioned above, is van Fraassen's temperature graph example (see Section 2), in which a trajectory of temperature for a region over the course of a day is produced from direct measurements of temperature made periodically at a several different stations in the region. A detailed model of this complex measurement process will include assumptions not only about the direct measurement of the station temperatures and about how different station measurements made at a single time should be combined to arrive at a regional temperature estimate for that time, but also about how a continuous trajectory for temperature over the course of the day – a data model – should be produced from the regional temperature estimates for specific times. It also may include steps to correct for idealized assumptions made elsewhere in the measurement process.

4 Can computer simulations measure real-world target systems?

In a computer simulation study, an agent uses a digital computer to execute a special kind of algorithm: one designed to repeatedly solve dynamical equations, at least some of whose variables are understood to represent properties of a real or imagined system, the target system. Such an algorithm, when implemented on a particular digital computer, is a computer simulation model. From a specification of values for the model's variables at an initial time,

[19] In more complicated cases, the measurement outcomes that are used together in complex measurement might themselves be complex measurement outcomes.

[20] While derived measurement also often has multiple "input" measurements, typically these are only jointly sufficient for obtaining an informative measurement outcome, whereas in complex measurement a subset would be sufficient for an informative outcome.

t_0, the computer solves the dynamical equations to produce values for a later time, t_1; from the values for t_1, it calculates values for t_2, and so on for some number of time steps. Strictly speaking, the time steps need not proceed from earlier to later; in some cases an algorithm might calculate backward in time instead. Either way, the computer produces a time-ordered sequence of selective representations of a target system. Usually, this sequence of representations – assignments of values to model variables – is saved as the output or "results" of the computer simulation.

Can running a computer simulation model and observing the results be a process for measuring the properties of the system being simulated? Measuring is an activity that involves, among other things, physical interaction with the system being measured. But running a computer simulation model and observing the results involves no physical interaction with the target system being simulated (see also Giere [2009] and Parker [2010] responding to Morrison [2009]). Such a study involves physical interaction with the simulat*ing* system – the programmed digital computer – but not with the simulat*ed* system.[21] A computer simulation, on its own, is not a process for measuring properties of the system being simulated.

Nevertheless, it is possible for computer simulations to be embedded in measurement practices and, indeed, for them to be embedded in measurement practices in such a way that simulation results constitute raw instrument readings or even measurement outcomes.[22] The remainder of this section illustrates how this might happen in each of the three types of measurement identified in Section 3.

In direct measurement, computer simulation results can be measurement outcomes when computer simulation is employed to correct for the effects of interfering factors. An example of van Fraassen's – unrelated to computer simulation – illustrates why we sometimes need corrections of a type that simulation could help provide. Suppose we are interested in measuring the temperature of a very small cup of hot tea at time t_0, and we insert a mercury thermometer at that time; we wait a short while for the mercury to stop rising in the tube and take a reading. But thermodynamic theory tells us that the thermometer itself will affect the temperature of the tea and hence the reading obtained. To arrive at a more accurate temperature estimate for t_0, our measurement process will need to include a step that corrects the thermometer reading for this interference. This might involve calculating the earlier temperature of the tea using the thermometer reading, thermodynamic theory and our knowledge of the initial temperature of the thermometer (see van Fraassen [2008], p.146). In this example, the equation that needs to be solved to obtain the corrected value might be solved directly, but in other cases corrected values might be obtained with the help of computer simulation. In those cases, simulation results can be direct measurement outcomes.[23]

---

[21] Perhaps in some cases running a computer simulation model and observing the results can be a process for measuring properties of the programmed digital computer (but see Barberousse et al. [2009], Fn.3, for some complexities).

[22] Morrison ([2009]) suggests that computer simulation models can function as measuring instruments. I agree, but only in the sense that a computational instrument/apparatus (such as a calculator or a programmed computer) can be part of a measurement process.

[23] Activities that involve simulation in this way could be characterized as either direct or derived measurement processes, depending on the model of the measurement process that is assumed. If we think of the thermometer reading as a biased indicator of the original temperature of the tea that needs correcting, we describe the process

It is easier to see how derived measurement practices could embed computer simulation in such a way that simulation results constitute raw instrument readings or derived measurement outcomes: computer simulation could be used to perform the derivation step in diachronic derived measurement. Suppose, for example, that we make direct measurements of the relative positions of the sun, moon and earth today and use these, along with previously-made measurements of their masses, as initial conditions for a Newtonian simulation of the earth-moon-sun system, in order to estimate the relative positions of these bodies five days or five months ago.[24] The earlier positions indicated by the simulation could be either raw instrument readings or the best-estimate portion of the measurement outcome, depending on our model of the measurement process; it depends in particular on whether that model calls for corrections to be applied to the simulation results or whether the results are expected to be accurate enough that they require no subsequent calibration.[25]

What sorts of correction might be required? Corrections for at least two kinds of error related to the simulation process: dynamical model error and numerical error.[26] *Dynamical model error* is error in the simulation results due to the ways in which the equations of the simulation model are idealized, simplified or otherwise distorted representations of the actual relations among target system properties. In the model used for the Newtonian simulation, for example, the gravitational influence of other bodies in the universe will be handled in a simplified way (e.g. as small perturbation term) for the sake of computational tractability. *Numerical error* is error in the simulation results due to the use of numerical methods that calculate only approximate solutions to the dynamical modeling equations. These two types of error cannot always be quantified separately, but if there is reason to think that they are either singly or jointly significant, given the accuracy of the result we are seeking, then corrections to the simulation results will need to be applied to arrive at the measurement outcome. If this is not feasible, because the errors are not well understood (more on this below), we may be left with simulation results as raw instrument readings from which we do not know how to infer measurement outcomes.

---

as a direct measurement process. If we think of the reading as a measurement outcome for a different parameter than the one that interests us – a parameter representing the later temperature of the tea, then we characterize the same activity as a diachronic derived measurement process. Tal ([2012], pp.19-20) notes that the same physical measurement process can be modeled in different ways; as this example illustrates, it is not just the level of detail of the model that can vary but, in some cases, even whether the process is modeled as a direct or derived measurement process. This does not mean that a model of a measurement process can take any form one likes, at least not if one's goal is successful measurement; the requirement that the model cohere with background knowledge constrains the choice.

[24] The use of backward simulation here is analogous to the use of carbon dating to estimate, on the basis of isotope measurements today, what the age of an ancient artifact was in the year 1000.

[25] Giere ([2009]) briefly discusses a very similar example but argues that the results are not measurement outcomes because there is no "causal interaction with the target system" (p.60). Since there is such interaction in the course of measuring the initial condition and mass parameters, perhaps he means causal interaction with the target system at the time for which the measurement outcomes are meant to represent conditions. If so, this would be to deny the diachronic species of derived measurement. Alternatively, perhaps he envisions a slightly different example in which the position and mass measurements were all previously-made for some other purpose, so that the activity at hand encompasses just the computational step (see also Section 6 below).

[26] There are other possible sources of error as well, including round-off error and hardware error. These too should be considered but, as they are less often the source of trouble today, they are omitted from the discussion here for the sake of brevity.

Lastly, computer simulation can be embedded in complex measuring procedures in such a way that simulation results constitute raw instrument readings or measurement outcomes. Some studies employing data assimilation methods, such as those involving four-dimensional variational (4D-Var) assimilation, illustrate how this might work. These complex measuring procedures are designed to make use of two sets of measurement outcomes for times within the same period $[t_0, t_n]$, known as an *assimilation window*. One set consists of (a) direct and/or synchronic derived measurements of parameters at various times within the window, often called the *observations proper* (Talagrand [1997]). The other set consists of (b) diachronic derived measurements (predictions) of parameters for a sequence of times within the window; these predictions are generated via computer simulation, using initial conditions that are direct, derived or complex measurement outcomes for a time prior to t_0. These diachronic derived measurements constitute a *first-guess forecast* for the period $[t_0, t_n]$. Finally, the measuring procedure makes use of (c) a 4D-Var algorithm that revises (b) in light of (a). The algorithm aims to find, from among those sequences of parameter values for $[t_0,t_1,\ldots\ t_n]$ that are consistent with the relevant physical and dynamical 'laws' of the system – as represented in a simulation model – the sequence that best fits the measurement outcomes in (a) and (b), given their respective uncertainties. As described below, this search procedure involves running a number of computer simulations, one of which is chosen to provide the updated estimate of conditions for $[t_0,t_1,\ldots\ t_n]$.

An example will help to illustrate. Suppose we want to estimate how the state of a simple damped pendulum – its angular position and velocity – evolves over a period of time, $[t_0, t_n]$. The length of the pendulum is obtained by direct measurement. A complex measuring procedure involving 4D-Var might proceed as follows. (a) Observations proper of position and velocity are made at a series of times $[t_0,t_1,\ldots\ t_n]$; let us suppose that they are known to be unbiased but to have significant associated uncertainty (noise), which is specified. (b) A computer simulation model, initialized with observations proper of the position and velocity of the pendulum at a time $t < t_0$ is run to produce predictions of the pendulum position and velocity for the same times $[t_0,t_1,\ldots\ t_n]$ for which observations proper were made; suppose that these predictions also are subjected to a process of calibration, to produce estimates that are unbiased but that have some associated uncertainty (noise), which is specified. (c) A 4D-Var algorithm then looks for a sequence of position and velocity values for $[t_0,t_1,\ldots\ t_n]$ that is both consistent with the dynamical 'laws' according to which the pendulum behaves and a better fit to the measurement outcomes in (a) and (b), given their specified uncertainties, than any other sequence of position and velocity values that is consistent with the dynamical laws.[27]

4D-Var typically attempts to find such a best-fitting sequence of values in roughly the following way (see Figure 2).[28] A cost function is defined whose value is determined by (i) the differences between a candidate sequence of values and the observations proper and (ii) the difference between that sequence's values for t_0 and the first-guess forecast for t_0.[29]

[27] In this way, 4D-Var seeks to extract information contained not just in individual observations, but in their spatiotemporal distribution: "Observations distributed irregularly in time contain information in their spatio-temporal distribution. With 4D data assimilation (4DDA), it is possible to extract this information and one of the 4DDA algorithms is 4D-var …. The assimilation…imposes a dynamical constraint on the solution that pulls out the information contained in the time consistency of the observations" (Gauthier [1998], p.2).

[28] See also Bouttier and Courtier ([1999]) or Kalnay ([2003]) for a more detailed technical discussion.

[29] Note that computing (i) requires making the quantities comparable or "subtractable" (Lorenz [1985]; Smith [2006]). Grid point values in a simulation often are assumed to represent temporal and/or spatial averages –

Starting with the first-guess forecast, the 4D-Var algorithm identifies directions in the state space of the simulation model in which incremental shifts in the forecasted values would bring the largest reductions in the value of the cost function. It then uses an adjoint model to work backward to a set of initial conditions that (hopefully) will produce a simulation for $[t_0,t_n]$ that differs from the current one in those cost-reducing directions. A new candidate model trajectory (i.e. sequence of values) is then produced by running the simulation model with those initial conditions. The entire procedure is then repeated with that candidate trajectory, and so on, for some number of iterations. If all goes well, the 4D-Var algorithm converges to the model trajectory / sequence of values that minimizes the cost function.[30]

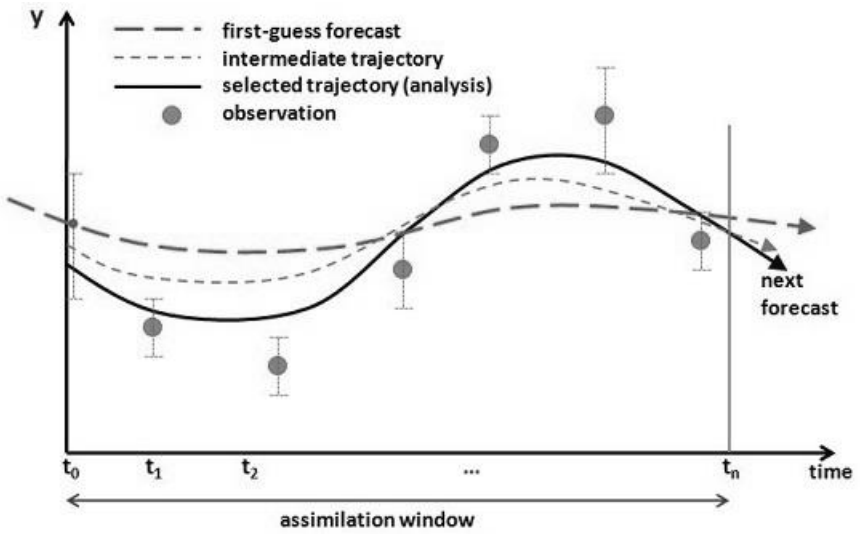

Figure 2. Elements of four-dimensional variational data assimilation (4D-Var). The result (thick solid line) is a model trajectory selected in light of the first-guess forecast (dashed line), observations (large dots), information about the errors and uncertainties associated with each (error bars / confidence intervals) and the physical laws of the system under measurement as represented in a simulation model (not shown). An intermediate trajectory considered by the iterative 4D-Var algorithm is also depicted (thin dashed line).

The model trajectory selected by the 4D-Var algorithm – in this case a discrete-time state trajectory for the pendulum – can be either a raw instrument reading or a complex measurement outcome; as in derived measurement, it depends on whether there is reason to think that sources of error will impact the results enough to require a subsequent calibration step. In the case of 4D-Var, the errors to be considered include not only dynamical model error and numerical error, but also error introduced because the methodology used to identify

---

more on this below – while conventional observations often take the form of point measurements. Thus in order to compute (i), an "observation operator" may be needed to map the observations to the model space (grid point values) or to map the grid point values to the observation space.

[30] A sequence of values for one parameter/variable is shown in Figure 2, but in practice the state trajectory selected by the assimilation algorithm will include a sequence of values for each state parameter/variable in the model. In the pendulum case, there are just two state variables at each time step; in atmospheric data assimilation, there can be 10^8 or more.

the best-fitting trajectory relies on idealized or simplistic assumptions (e.g. that the topology of the cost function is such that the 4D-Var algorithm will not get stuck in local minima). Once again, if there is reason to think that these errors are significant, given the accuracy of the result we are seeking, then corrections to the simulation results will need to be applied to arrive at the measurement outcome. And once again, if this is not feasible, because the errors are not well understood, we may be left with simulation results as raw instrument readings from which we do not know how to infer measurement outcomes.

To review, according to the conception of measurement adopted in Section 2, computer simulations on their own cannot be processes by which we measure the target systems being simulated, because they do not involve interaction (or even attempted interaction) with those target systems. Nevertheless, in principle computer simulations can be embedded in studies that do involve this interaction and, indeed, can be embedded in them in ways such that results from simulations constitute raw instrument readings or even measurement outcomes. In the next section we examine a real case.

5 Case study: atmospheric data assimilation

As mentioned in Section 1, atmospheric science is one context in which computer simulation results sometimes are referred to as "observations" and are used in the roles of traditional measurement outcomes. We are now in a position to understand which computer simulation results are characterized and used in this way: results of atmospheric data assimilation studies involving methods like 4D-Var. Results from similar studies that employ other data assimilation methods, which give results in the form of a blending of simulation results and conventional observations, also frequently are characterized in this way – as "observations" or "measurements" of atmospheric properties. I suggested above that results like these in principle can be complex measurement outcomes. In this section, after providing some background, I explain why currently atmospheric data assimilation should be understood as a complex measurement practice that is still under development.

5.1 Why data assimilation?

The original impetus for atmospheric data assimilation came from numerical weather prediction (NWP), in which a computer is used to calculate an estimate of later atmospheric conditions from an estimate of current ones, by solving fluid dynamical and other equations representing physical processes in the atmosphere.[31] Even before NWP, the depiction of the atmosphere from which a weather forecast was generated had been given a special name: *the analysis*. It was produced by hand, by plotting irregularly-spaced observations on maps and drawing lines of equal pressure, temperature, etc., using past experience and physical understanding to do this in a sensible way. With the analysis as the starting point, graphical techniques and rules of thumb were used to generate forecast maps of conditions at later times (see also Friedman [1989]; Monmonier [1999]).

With the advent of operational NWP in the 1950s, the analysis needed to be tailored to the computer model used to make the forecast: in each forecast cycle, initial values were needed for each of the model's variables ('temperature', 'pressure', 'easterly wind speed',

[31] For an insightful and detailed historical discussion of the development of both computer modeling and data assimilation in the study of weather and climate, see Edwards ([2010]). For a more technical introduction to data assimilation methods, see Kalnay ([2003], Ch.5).

etc.), for each of the model's grid points, at both the surface and a number of levels above. This amounted to thousands (and later millions) of initial values. The number and types of conventional observations being made were also increasing. A new approach to generating the analysis was developed, termed *objective analysis* (see Cressman [1959]). It automated the ingestion and quality control of observations as well as the generation of a gridded analysis from those observations.

Some early objective analysis algorithms produced gridded analyses by spatial interpolation: values for grid point variables, such as temperature, were estimated as a function of observations near the grid point location, though in some cases these observations were quite distant. Such algorithms relied on basic assumptions about the smoothness and general structure of atmospheric fields. It was soon recognized, however, that better analyses could be produced using data assimilation methods that took into account the information available in NWP forecasts for the analysis time. This additional information is especially useful for filling in large gaps in the observations proper. Consider the extreme illustration in Figure 3.[32] Since the 1960s, a number of different data assimilation methods have been developed and deployed in support of improved NWP forecasting, including optimal interpolation, 3D- and 4D-Variational assimilation, Kalman filtering and others (see Kalnay et al. [2003]).

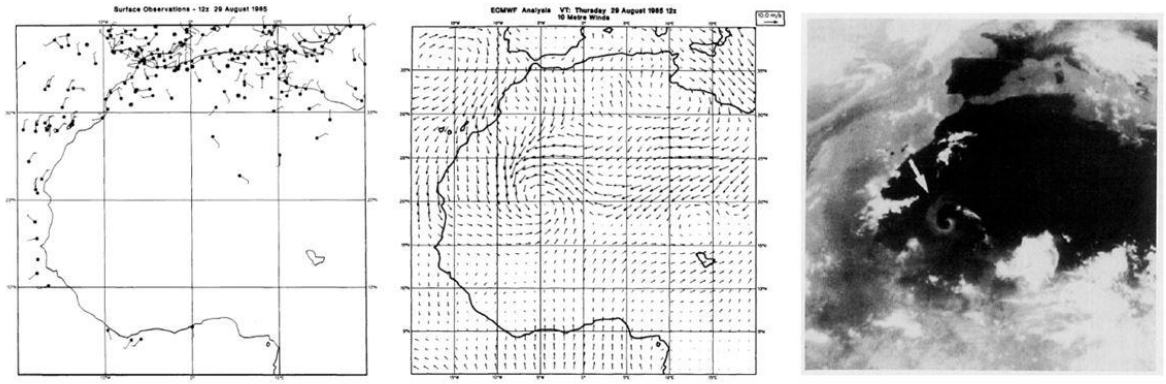

Figure 3. On this date in 1985, a technical problem at a regional telecommunications center prevented the transmission of many of the observations proper made in North Africa. The resulting large gaps in plots of these observations, including surface wind observations (left panel), were filled in during the data assimilation process using information from the NWP forecast. The resulting analysis for the near-surface wind field is shown in the middle panel. Independent confirmation of the prominent vortex shown in the analysis was obtained from satellite observations not used in the assimilation (right panel). (Adapted from Figures 3, 4 and 6 of Bengtsson, L. and Shukla, J. [1988]: 'Integration of Space and In Situ Observations to Study Global Climate Change', *Bulletin of the American Meteorological Society*, **69**(10), pp. 1130-43. ©American Meteorological Society. Used with permission.)

[32] Examples like these led some commentators to remark that a realistic forecast model "can be viewed as a unique and independent observing system that can generate information at a scale finer than that of the conventional observing system" (Bengtsson and Shukla [1988], p.1134). I would argue that the forecast model on its own is not an observing (or measuring) system independent of the conventional observing system, since the latter informs the initial conditions from which the model generates the forecast; the forecast model is part of a complex observing/measuring system that includes both the conventional observing system and the data assimilation procedure. See also Fn.22 above and Humphreys ([2013]; [2014]) on "causal-computational instruments."

Beginning in the 1990s, data assimilation also was used to construct long-term global datasets for past periods, in a process known as retrospective analysis or *reanalysis*. Constructing such datasets from conventional observations alone faces many of the same difficulties encountered when seeking initial conditions for NWP, such as large spatial gaps. In reanalysis, the basic idea is to break the past into a series of assimilation windows (e.g. a series of 12-hour-long windows) and perform data assimilation for each window, using some or all of the conventional observations that are now available for that past window. After the assimilation for the first window is complete, the algorithm moves on to the second window, and so on, up to the present. The NWP model and assimilation method are held fixed throughout the study, which helps to avoid some of the spurious jumps and trends that are seen when sequences of operational NWP analyses are examined; those spurious jumps and trends reflect occasional changes that forecasting centers make to their NWP models and assimilation methods, with the aim of improving forecast skill.

Numerous atmospheric reanalyses now have been performed, with several ongoing. Some are global in coverage and span several decades or more, while others target a specific region during a shorter time period. These reanalysis datasets are in heavy use. For example, the NCEP-NCAR global reanalysis dataset (Kalnay et al. [1996]), which covers the period 1948-present, now has more than 12,000 citations.[33] Practitioners routinely use these datasets in the roles of conventional observations and measurements. They are used to investigate atmospheric dynamics (e.g. Kidston et al. [2010]), to evaluate climate models (e.g. Gleckler et al. [2008]), as data in which to look for the presence of greenhouse gas fingerprints (e.g. Santer et al. [2004]), and for various other purposes. Some practitioners refer to reanalysis data cautiously as "reference data" while other routinely call them "observational data" or simply "observations".

One recent use of reanalysis results deserves special mention. Compo et al. ([2011]; [2013]) employ reanalysis results to rebut concerns about the reliability of global warming estimates made previously from thermometer readings. Some skeptics contend that the apparent warming over land seen in those thermometer-based studies is an artifact of the siting of thermometers, heat island effects, etc. Compo et al. ([2013]) investigate this possibility using results from the 20th Century Reanalysis (20CR), a reanalysis that is unusual in at least two respects. First, it covers the period 1871 to the present, the longest time period of any reanalysis available today. Second, in 20CR, over land regions only conventional observations of surface pressure are used in the assimilation; thermometer data, wind data, satellite data, etc. are deliberately omitted. Compo et al. show that 20th century global warming over land estimated from 20CR results – without the use of thermometer data – is quite similar to the thermometer-based estimates. They describe this as "independent *observational* confirmation" (Compo et al. [2013], p.3172) of such warming.[34] In a *Scientific American* piece discussing these findings, 20CR is described as an alternative approach to "measuring" changes in surface temperature (Ogburn [2013]).

[33] This is according to the Web of Science database, accessed on March 19, 2014.

[34] They consider it "observational" because conventional observations of some sort are assimilated at each time step. They do not consider ordinary computer simulations of 20th century climate change to provide observational confirmation of that warming, because the simulations are not designed to simulate the actual trajectory of the climate system but rather to produce a representative trajectory (or set of trajectories) for specified boundary conditions (see Compo et al. [2013]).

5.2 A complex measuring practice under development

Atmospheric data assimilation involves combining information about atmospheric conditions from two sources: the first-guess NWP forecast and conventional observations. The former is intended to serve as a diachronic derived measurement outcome, while the latter are direct or synchronic derived measurement outcomes, just as in the pendulum example of Section 4. Exactly how information from those sources is combined depends on which data assimilation method is used, but the aim is to deliver a more informative estimate of atmospheric properties (such as temperature, pressure, wind speed, etc.) than could be produced from either alone. Atmospheric data assimilation thus looks like an attempt at complex measurement.

Nevertheless, it seems best to characterize atmospheric data assimilation as a complex measurement practice that is *still under development*. Ultimately, this is because today's data assimilation systems are not subjected to a rigorous process of calibration that provides well-motivated uncertainty estimates. Closely related is an issue of interpretation: some fundamental questions about how to interpret the results of today's atmospheric data assimilation studies remain unanswered. These matters are discussed in turn below.

Difficulties related to calibration begin with the first-guess forecast. Such forecasts, while informative, are known to err in nontrivial ways at least some of the time, indicating that some calibration is required; the output of the forecast model is more like a raw instrument reading than a measurement outcome. But calibrating the forecast is far from trivial. White-box calibration would require identifying and correcting for the impacts of component sources of error, including imperfections in the nonlinear equations of the simulation model (dynamical model error).[35] But such sources of error are not well-enough understood to apply corrections and estimate uncertainties in a confident way (see also Whitaker and Hamill [2012]).[36]

Black-box calibration is the alternative, but here too there are obstacles. Most obviously, there is no independent reference standard available, i.e. no accurate three-dimensional atmospheric state estimate with which to compare; else there might be no need for data assimilation. In some regions, forecasted grid point values can be compared with observations proper, but even this is more complicated than one might think, since the observations are often point measurements, while the forecasted grid point values are not (see Fn.29 and the issue of interpretation below). In practice, it is sometimes assumed that forecast bias and uncertainty can be estimated by taking statistics on the differences between the first-guess forecasts and the analyses produced by a data assimilation system, but this approach has limitations. Not only does it assume that the analyses are accurate, but it may fail to account for the fact that forecast errors and uncertainties are to some extent state (or regime) dependent: forecasts launched from some states will have different errors and uncertainties than forecasts launched from other states. Biases and uncertainties estimated without regard for such regime-dependence will gloss over these differences. How fine-grained the reference classes should be is an empirical question whose answer is not yet clear. (But consider the

[35] Not to mention the fact that the initial conditions of the forecast – i.e. the previous analysis! – are also not yet well-calibrated, as we will see.

[36] Bogen ([2002]) identifies a similar problem related to calibration in the case of fMRI images, which are also produced with the help of complex computational methods.

number of states that in principle might be individuated for an atmospheric model with ~10^8 state variables, even if each variable is imagined to have only ten possible values.)

Calibration of the results of the full assimilation process also remains a challenge. This is not only because the first-guess forecast is one of the ingredients of the assimilation, but also because many of the same challenges encountered when considering how to calibrate the first-guess forecast (e.g. state dependence, lack of comprehensive reference standards) are also present when the data assimilation results themselves are under consideration. In fact, when it comes to the assimilation results, we add to the mix the possibility of error due to idealized assumptions of the assimilation algorithm, such as assumptions about the topology of the cost-function in 4D-Var. In practice, nontrivial errors are expected to be introduced because of those idealized assumptions (see e.g. the discussion of Talagrand [2010]), but it is not easy to identify and correct them.

Given these difficulties related to calibration, it appears that data assimilation at present fails to meet the coherence and consistency requirements of the model-based account of (successful) measurement. The assumptions employed in inferring data assimilation results from instrument indications fail to cohere with background knowledge in some ways that can be expected to introduce substantial errors in results, but exactly when and to what extent they do so – and thus how to reliably correct for these errors in a process of calibration – remains a topic of ongoing research.[37] This in turn is one reason why results obtained via data assimilation and reanalysis often are not accompanied by uncertainty estimates: often only best-estimate values are given. As a consequence, these results cannot yet be shown to display the statistical consistency that is a hallmark of successful measurement.[38]

One reanalysis study that does provide uncertainty estimates is 20CR, introduced in Section 5.1 above. Upon "eyeballing" plots of global mean temperature change estimates from both 20CR – obtained without assimilating station-based thermometer data – and a number of thermometer-based studies, there looks to be impressive agreement. Compo et al. ([2013]) note, however:

> Although the agreement between 20CR and the station-based data sets is strong, the mean square differences between them are somewhat larger than expected from their respective confidence intervals…. This suggests that the data sets underestimate their uncertainty, particularly 20CR during the periods of disagreement…. (ibid, pp.3171-2)

This is a nice illustration of the sort of consistency check that Tal's ([2012]) analysis recommends; in this case it confirms that there is still some work to be done, since rigorous consistency has not yet been achieved. Note, however, that the main message of the Compo

[37] In practice, such corrections, if they are made at all, tend to be made in a way that is recognized to be simplistic. For example, discussing 20CR, Compo et al. ([2011], p.21) remark: "Several components of the current algorithm concerned with accounting for uncertainties arising from the use of a finite ensemble, an imperfect model, and imperfect observations are rather simplistic."

[38] It is not that there is no insight into which results are likely to be more and less informative. Conventional observations are denser in some regions and time periods than in others, NWP models are known to be more reliable when it comes to some fields and weather situations than others, and results for some variables have been compared with independent measurements collected in special field campaigns. But it remains unclear just how accurate we can expect results for different reanalysis fields to be in different regions and at different times. Practitioners tend to treat as less reliable those reanalysis fields that are derived from the model's state variables and for which no conventional observations are assimilated (see the distinction among A, B and C fields in Kalnay et al. [1996]), but this is just a reasonable rule of thumb.

et al. ([2013]) article is that the 20CR results confirm the thermometer-based estimates; even if rigorous statistical consistency has not been achieved, there is clear agreement among independent estimation techniques on the basic picture of a warming world and even on many of the regional-scale temperature changes.

A second remaining issue, in practice related to calibration but at the same time more fundamental, relates to the interpretation of results from data assimilation studies. In particular, the mathematical techniques used in the simulation portions of atmospheric data assimilation make it more difficult to interpret some of the results.

In the pendulum example in Section 4, the measurement outcome came in the form of a discrete-time state trajectory. This makes sense from the point of view of the mathematics involved in the measurement process: the pendulum's dynamics are thought to be described by ordinary differential equations, for which numerical methods are used to estimate solutions in discrete time steps by the computer; those numerical methods are designed to ensure (at least) that, when several time steps in a row are considered, the estimated state trajectory will resemble the continuous trajectory of the differential equations.

In atmospheric data assimilation, the dynamical laws include partial differential equations, closely related to the Navier-Stokes equations of fluid dynamics. Numerical methods are used to estimate solutions to a spatially and temporally discretized version of these equations.[39] Assuming such solutions exist, the numerical methods used are designed to give results that resemble those solutions when several spatial grid points are considered together over several time steps; they do not guarantee that any particular grid point value will be "very close" to the average value for some region of the continuous field (e.g. a volume) that we associate with that grid point.[40] From a mathematical point of view, the measurement outcome in an atmospheric data assimilation study should be a function of (at least) several grid points considered over several time steps.

There is a tendency in practice, however, to use reanalysis datasets as if results for individual grid point variables at individual time steps are measurement outcomes, representing the average conditions in a volume of atmosphere that we choose to associate with the grid point. In principle, this might be justified on empirical grounds – by showing that reanalysis results for particular grid point variables in fact do closely track reliable, independent measurements of conditions in such atmospheric volumes. This is not ruled out on mathematical grounds; it is just not guaranteed. But as noted above in connection with black-box calibration, such reliable independent measurements are not available in a comprehensive way (else we wouldn't need reanalyses). The upshot is that a fundamental question about today's atmospheric data assimilation systems remains unanswered, namely, the question of what exactly we can expect them to give us measurements of – state estimates, state trajectories, volume-average parameters, point parameters or something else – and with what associated uncertainties.

[39] The existence and smoothness of solutions to the full Navier-Stokes equations has not been proven yet; doing so is a Millennium Prize problem carrying an award of $1 million. See http://www.claymath.org/millennium-problems/millennium-prize-problems.

[40] Thanks to Leonard Smith for helping me begin to understand the issues involved here.

5.3 Epistemic iteration

Recent historical and philosophical work on scientific measurement has called attention to the long journey that precedes the achievement of a stable measuring practice. It is common to analyze measurement from the point of view of the end of the journey, after what is thought to be a successful measuring procedure has been achieved, but it is also important to consider what goes on during the journey.[41] Hasok Chang ([2004]) argues that the journey often involves *epistemic iteration*, "a process in which successive stages of knowledge, each building on the preceding one, are created in order to enhance the achievement of certain epistemic goals" (p.45).

Current atmospheric data assimilation practices can be viewed as the latest stage in the development of methods for reliably estimating the three-dimensional state of the atmosphere. Earlier stages include that in which hand-drawn atmospheric analyses were made from sparser collections of conventional observations (in the 19th and early 20th centuries) as well as that in which atmospheric analyses were produced using automated spatial interpolation methods (the original "objective analysis" schemes of the mid-20th century). This process of development, and the accompanying expansion of the conventional observing system, exemplifies Chang's epistemic iteration: successive stages of methodologies for estimating the atmospheric state, each building on the preceding one, have been created in order to enhance the achievement of epistemic goals – such as skillful weather forecasting, understanding atmospheric phenomena, quantifying climate change, and so on. With respect to some of these goals, such as skillful weather forecasting, there has been significant progress. This progress is important to recognize in the face of the remaining limitations discussed in Section 5.2.

Nevertheless, the warnings that some scientists have issued regarding the use of reanalysis datasets do have merit. For example, Gavin Schmidt ([2011]) and Dick Dee et al. ([2014]) have cautioned that reanalysis results should not be equated with "real observations". It is tempting to assume that what motivates and justifies this warning is the fact that reanalysis results are simulation-dependent and, in some cases, come in the form of simulation output. But the discussion of Sections 3 and 4 suggested that the involvement of simulation is not necessarily a problem. The real issue seems to be that the extent to which data assimilation systems give accurate results, and what exactly those results represent, has not been established. Insofar as more familiar measuring instruments, such as thermometers and barometers, have undergone a careful process of calibration and give results that we know how to interpret, there is indeed an important difference between reanalysis results and the observations and measurements obtained from these familiar instruments, one that should be kept in mind when using reanalysis datasets.

6. The boundaries of measurement

Section 4 explained how computer simulations could be embedded in direct, derived and complex measurement practices in such a way that simulation results constituted raw instrument readings or measurement outcomes. Reflecting on that analysis, it is easy to imagine simulation-embedding investigations that are more difficult to classify. As discussed below, these include investigations in which the initial conditions for the simulation are a

[41] van Fraassen calls the former a view of measurement "from above" and the latter a view "from within" (see [2008], p.91).

mixture of measurement outcomes and guesses, as well as investigations in which all of the initial conditions for the simulation come from "previously-made" measurements. These cases, all of which have straightforward analogues involving unassisted calculation rather than simulation, suggest that the boundary between measurement and non-measurement is not clear cut.

In Section 4, a derived measurement practice embedding computer simulation was envisioned as follows: direct measurements of target system parameters are made by investigators, who then use the results as initial conditions for a simulation model, which performs the derivation step of the measurement. We can contrast this with a case in which all of the initial conditions for a computer simulation are obtained by guessing;[42] in that case, the investigation would not constitute a process for measuring the target system according to the view adopted in Section 2, since it would lack the relevant physical interaction with the target system.

Questions then arise about intermediate cases, in which only some of the initial conditions for the simulation are obtained by measurement, while the others are guesses. If the simulation results of interest are determined almost entirely by the measured initial conditions, then it seems reasonable to say that they can be derived measurement outcomes. At the other extreme, if the results of interest are determined almost entirely by the initial conditions that are guesses, then presumably this is not a case of measurement, even if it turns out that the results agree closely with the results that other measurement processes give for the same parameters.[43] Between these two extremes, we can envision a continuum of intermediate cases; it is difficult to imagine drawing a non-arbitrary line between those that do constitute measurements and those that do not. This suggests that the boundary between measurement and non-measurement is fuzzy.[44]

Cases in which all of the initial conditions for a simulation are outcomes from "previously-made" measurements – ones made in the past for other purposes – also seem difficult to classify. On one hand, it is natural to say that in this situation the investigators running the simulation are not engaged in a process of measurement (of the target system); they are performing calculations using available background information. On the other hand, one might argue that what matters is that the investigation produces accurate results using information obtained via the right sort of physical interactions with the target system – interactions that did occur when those past measurements were made; what the people arranging those interactions had in mind is irrelevant.[45] It is unclear whether we have anything more than a terminological dispute here.

---

[42] Here I mean guessing that is not informed by relevant interaction with the target system.

[43] Here, a model of the (candidate) measurement process will have no real account of how the process manages to provide desired information about the simulated system. In fact, if the uncertainty associated with the guesses is handled properly, presumably the simulation results should be accompanied by huge uncertainty bounds, indicating that the process is not very informative.

[44] This is hardly a shocking suggestion, since many concepts are fuzzy in this way.

[45] Moreover, insisting that physical interactions must have been arranged for the purpose of measuring the parameter(s) of interest would have some odd implications. Consider reanalysis studies that provide information about what conditions were like in time periods before the 'invention' of reanalysis. Reanalysis results produced for those time periods cannot be measurement outcomes, since the conventional observations on which they depend obviously were not made for the purpose of supporting reanalysis efforts. Yet results produced in the same study using the same methods but representing conditions for more recent time periods, when many

It should be obvious that the harder-to-classify cases identified above have straightforward analogues involving unassisted calculation rather than simulation; it is not simulation per se that is the source of trouble. The point of discussing such cases is not to encourage further debate about how to distinguish "genuine" measurement from other practices but simply to acknowledge that there will be cases whose classification is not clear cut, given the conception of measurement adopted here, and to indicate what some of those cases might look like.

7. Epistemology, not terminology

As emphasized from the start, my concern in this paper was not primarily with terminology or labeling. I did suggest that some practices that normally are not called "measuring" practices, including some predictive and retrodictive practices, might reasonably be classified as such. But the grounds for this were epistemological: these practices do not seem fundamentally different, epistemically, from more familiar forms of derived measurement.

My main interest was in exploring how computer simulation might be embedded in existing forms of measurement and in understanding what such practices might look like in their detailed epistemology. I pursued this using Tal's model-based framework for the epistemology of measurement, which I attempted to flesh out further for three different types of measurement: direct, derived and complex. The analysis showed that in principle computer simulation can be embedded in these measuring practices in such a way that simulation results constitute measurement outcomes.

Atmospheric data assimilation, in which model-based forecasts and conventional measurements are combined to estimate the atmospheric state, was examined as a case study. It is of particular interest because practitioners today often call its results "observations" and use them as if they were measurement outcomes. At the same time, some atmospheric scientists have cautioned against thinking of data assimilation results as "real" observations or measurements. I argued that this caution is justified, but not because data assimilation's embedding of computer simulation is inherently problematic. The problem, rather, is largely that of calibration: results obtained from today's atmospheric data assimilation systems are not subjected to a rigorous calibration process that also provides well-motivated uncertainty estimates. This is in contrast to many conventional measuring instruments, which have been carefully calibrated. The prospects for achieving such calibration for atmospheric data assimilation systems remain unclear. At present, atmospheric data assimilation seems best characterized as a complex measuring practice that is still under development.

The case study revealed some complications and challenges that may well resurface in many other simulation-embedding measuring practices. Some of these, such as those stemming from discretization and the use of numerical methods, are absent from typical measuring practices. For instance, we saw that properties of the numerical methods used to estimate solutions to partial differential equations can complicate the interpretation of values assigned to grid point parameters. Others challenges look like standard ones, just exacerbated. Calibration, for example, is in general a challenging business, but it can be especially difficult when the measuring process embeds the complex, motley and nonlinear models that are typical of computer simulation. Insofar as the trend in contemporary science

---

conventional observations have been made in part to support data assimilation / reanalysis efforts, could be measurement outcomes.

is toward embedding computer simulation in an ever-wider range of practices, including observing and measuring practices, it seems worthwhile to examine these complications and challenges, and the prospects for overcoming them in real cases, in greater detail than could be done here.

Lastly, for those readers who are concerned with terminology and who object to the moderately permissible conception of measurement adopted in this paper, perhaps there is a simple solution: to treat the foregoing as a discussion of how computer simulation might be embedded fruitfully in practices that aim to find out the values of target system parameters, whatever further labels we decide to apply to those practices. In the end, agreeing on which simulation-embedding practices we call "measurement" is less important than getting clear on what we can expect to learn about the world via those practices and why.

Acknowledgements

For helpful discussion and feedback, the author wishes to thank Leonard Smith, Harold Brown, Philip Ehrlich, Paul Humphreys, Gavin Schmidt, Reto Knutti, two BJPS referees as well as audiences at the 2012 Eastern APA meeting, the 2013 Bergen Philosophy of Science Workshop, the Bielefeld Dimensions of Measurement conference, the University of Colorado - Boulder, Aarhus University and Northern Illinois University. This material is based upon work supported by the U.S. National Science Foundation under Grant No. SES-1127710.

Wendy S. Parker
Department of Philosophy
Durham University
50 Old Elvet
Durham DH1 3HN, UK
wendy.parker@durham.ac.uk

References

Barberousse, A., Franceschelli, S. and Imbert, C. [2009]: 'Computer simulations as experiments', *Synthese*, **169**, pp. 557-74.

Batitsky, V. [1998]: 'Empiricism and the Myth of Fundamental Measurement', *Synthese*, **116**, pp. 51-73.

Bengtsson, L. and Shukla, J. [1988]: 'Integration of Space and In Situ Observations to Study Global Climate Change', *Bulletin of the American Meteorological Society*, **69**(10), pp. 1130-43.

Bogen, J. [2002]: 'Epistemological Custard Pies from Functional Brain Imaging', *Philosophy of Science*, **69**: pp. S59-S71.

Boumans, M. J. [2005]: 'Measurement outside the laboratory', *Philosophy of Science*, **72**(5), pp. 850-63.

Boumans, M. J. [2006]: 'The difference between answering a 'why' question and answering a 'how much' question', in J. Lenhard, G. Küppers and T. Shinn (eds) *Simulation: Pragmatic Construction of Reality,* Dordrecht: Springer, pp. 107-24.

Boumans, M. J. [2012]: 'Modeling Strategies for Measuring Phenomena In- and Outside the Laboratory', in H. W. de Regt, S. Hartmann and S. Okasha (eds) *EPSA Philosophy of Science: Amsterdam 2009: The European Philosophy of Science Association Proceedings*, Dordrecht: Springer, pp. 1-11.

Bouttier, F. and Courtier, P. [1999]: 'Data assimilation concepts and methods', <http://old.ecmwf.int/newsevents/training/lecture_notes/pdf_files/ASSIM/Ass_cons.pdf>.

Brown, H. I. [1985]: 'Galileo on the Telescope and the Eye', *Journal of the History of Ideas*, **46**(4), pp. 487-501.

Brown, H. I. [1987]: *Observation and Objectivity,* New York: Oxford University Press.

Chang, H. [2004]: *Inventing Temperature*: *Measurement and Scientific Progress,* New York: Oxford University Press.

Compo, G. P., Whitaker, J. S., Sardeshmukh, P. D., Matsui, N., Allan, R. J., Yin, X., Gleason, B. E., Vose, R. S., Rutledge, G., Bessemoulin, P., Brönnimann, S., Brunet, M., Crouthamel, R. I., Grant, A. N., Groisman, P. Y., Jones, P. D., Kruk, M. C., Kruger, A. C., Marshall, G. J., Maugeri, M., Mok, H. Y., Nordli, Ø., Ross, T. F., Trigo, R. M., Wang, X. L., Woodruff, S. D. and Worley, S. J. [2011]: 'The Twentieth Century Reanalysis Project', *Quarterly Journal of the Royal Meteorological Society*, **137**(654), pp. 1-28.

Compo, G. P., Sardeshmukh, P. D., Whitaker, J. S., Brohan, P., Jones, P. D. and McColl, C. [2013]: 'Independent confirmation of global land warming without the use of station temperatures', *Geophysical Research Letters*, **40**, pp. 3170-4.

Cressman, G. [1959]: 'An Operational Objective Analysis System', *Monthly Weather Review*, **87**(10), pp. 367-74.

Dee, D., Fasullo, J., Shea, D., Walsh, J. and National Center for Atmospheric Research Staff [2014]: 'The Climate Data Guide: Atmospheric Reanalysis: Overview & Comparison Tables', <https://climatedataguide.ucar.edu/climate-data/atmospheric-reanalysis-overview-comparison-tables>.

Edwards, P. N. [2010]: *A Vast Machine: Computer Models, Climate Data and the Politics of Global Warming,* Cambridge, MA: MIT Press.

Ellis, B. [1960]: 'Some fundamental problems of basic measurement', *Australasian Journal of Philosophy*, **38**(1), pp. 37-47.

Ellis, B. [1966]: *Basic Concepts of Measurement,* New York: Cambridge University Press.

Friedman, R.M. [1989]: *Appropriating the Weather: Vilhelm Bjerknes and the Construction of a Modern Meteorology*, New York: Cornell University Press.

Gauthier, P. [1998]: 'Development of 4D Data Assimilation at the Atmospheric Environment Service', <http://collaboration.cmc.ec.gc.ca/cmc/cmoi/product_guide/docs/lib/4dvar_en.pdf>.

Giere, R. [2009]: 'Is computer simulation changing the face of experimentation?', *Philosophical Studies*, **143**(1), pp. 59-62.

Gleckler, P. J., Taylor, K. A. and Doutriaux, C. [2008]: 'Performance Metrics for Climate Models', *Journal of Geophysical Research*, **113**, pp. D06104.

Gockenbach, M. and Schmidtke, K. [2009]: 'Newton's law of heating and the heat equation. *Involve: A Journal of Mathematics*', **4**(2), pp. 417-35.

Gygi, F. and Galli, G. [2005]: '*Ab initio* simulation in extreme conditions', *Materials Today*, **8**(11): pp. 26-32.

Humphreys, P. [2013]: 'What are data about?, in E. Arnold and J. Duran (eds) *Computer Simulations and the Changing Face of Experimentation*, Cambridge: Cambridge Scholars Publishing, pp. 12-28.

Humphreys, P. [2014] 'X-ray Data and Empirical Content', in P. E. Bour, G. Heinzmann, W. Hodges and P. Schroeder-Heister (eds) *Logic, Methodology and Philosophy of Science. Proceedings of the Fourteenth International Congress (Nancy)*, London: College Publications, pp. 1-15.

Kalnay, E. [2003]: *Atmospheric Modeling, Data Assimilation and Predictability*, Cambridge: Cambridge University Press.

Kalnay, E., Kanamitsu, M., Kistler, R., Collins, W., Deaven, D., Gandin, L., Iredell, M., Saha, S., White, G., Woollen, J., Zhu, Y., Leetmaa, A., Reynolds, R., Chelliah, M., Ebisuzaki, W., Higgins, W., Janowiak, J, Mo, K. C., Ropelewski, C., Wang, J., Jenne, R. and Joseph, D. [1996]: 'The NCEP–NCAR 40-Year Reanalysis', *Bulletin of the American Meteorological Society*, **77**(3), pp. 437-71.

Kidston, J., Frierson, D. M. W., Renwick, J. A., Vallis, G. A. [2010]: 'Observations, Simulations, and Dynamics of Jet Stream Variability and Annular Modes', *Journal of Climate*, **23**, pp. 6186-99.

Kosso, P. [1992]: 'Observation of the Past', *History and Theory*, **31**(1), pp. 21-36.

Kyburg, H. E. [1984]: *Theory and Measurement*, Cambridge: Cambridge University Press.

Lorenz, E. N. [1985]: 'The growth of errors in prediction', in M. Ghil (ed.) *Turbulence and Predictability in Geophysical Fluid Dynamics*, North Holland: Amsterdam, pp. 243-265.

Luo, J., Liu, Q., Tu, L.-C., Shao, C.-G., Liu, L.-X., Yang, S.-Q., Li, Q. and Zhang, Y.-T. [2009]: 'Determination of the Newtonian Gravitational Constant *G* with Time-of-Swing Method' *Physical Review Letters*, **102**, pp. 240801.

Monmonier, M. [1999]: *Air Apparent: How Meteorologists Learned to Map, Predict and Dramatize Weather*, Chicago: University of Chicago Press.

Morrison, M. [2009]: 'Models, measurement and computer simulation: The changing face of experimentation', *Philosophical Studies*, **143**(1), pp. 33-57.

NIST [n.d.] 'Virtual Measurement', <http://www.nist.gov/itl/vm/>.

Ogburn, S.P. [2013], 'New Method Proves – Again – Climate Change is Real', *Scientific American*. <http://www.scientificamerican.com/article.cfm?id=new-method-proves-climate-change-is-real>.

Parker, W.S. [2010] 'An Instrument for What? Digital Computers, Simulation and Scientific Practice', *Spontaneous Generations: A Journal for the History and Philosophy of Science*, **4**: pp. 39-44.

Quinn, T., Parks, H., Speake, C. and Davis, R. [2013]: 'Improved determination of *G* using two methods', *Physical Review Letters*, **111**: pp. 101102.

Santer, B.D., Wigler, T. M. L., Simmons, A. J., Kållberg, P. W., Kelly, G. A., Uppala, S. M., Ammann, C., Boyle, J. S., Brüggemann, W., Doutriaux, C., Fiorino, M., Mears, C., Meehl, G. A., Sausen, R., Taylor, K. E.. Washington, W. M., Wehner, M. F. and Wentz, F. J. [2004]: 'Identification of anthropogenic climate change using a second-generation reanalysis', *Journal of Geophysical Research*, **109**: pp. D21104.

Schmidt, G. [2011]: 'Reanalyses 'R' Us', <http://www.realclimate.org/index.php/archives/2011/07/reanalyses-r-us/>.

Shapere, D. [1982]: 'The concept of observation in science and philosophy', *Philosophy of Science*, **49**(4), pp. 485-525.

Smith, L.A. [2006]: 'Predictability Past, Predictability Present', in T. Palmer and R. Hagedorn (eds) *Predictability of Weather and Climate*. Cambridge: Cambridge University Press, pp. 217-50.

Stevens, S.S. [1959] 'Measurement, psychophysics and utility', in C. W. Churchman and P. Ratoosh (eds) *Measurement: Definitions and Theories*, New York: Wiley, pp. 18-63.

Tal, E. [2012]: *The Epistemology of Measurement: A Model-Based Approach*, Ph.D. Dissertation, University of Toronto.

Tal, E. [2013]: 'Old and New Problems in Philosophy of Measurement', *Philosophy Compass*, **8**(12), pp. 1159–73.

Talagrand, O. [1997]: 'Assimilation of observations: An introduction', *Journal of the Meteorological Society of Japan*, **75**, pp. 191–205.

Talagrand, O. [2010]: 'Variational Assimilation', in W. A. Lahoz, B. Khattatov and R. Ménard (eds) *Data Assimilation: Making Sense of Observations*, Berlin Heidelberg: Springer-Verlag GmbH, pp. 41-67.

Whitaker, J.S. and Hamill, T.M. [2012]: 'Evaluating Methods to Account for System Errors in Ensemble Data Assimilation', *Monthly Weather Review*, **140**, pp. 3078–89.

van Fraassen, B. C. [2008]: *Scientific Representation*, Oxford: Oxford University Press.

van Fraassen, B. C. [2012]: 'Modeling and Measurement: The Criterion of Empirical Grounding', *Philosophy of Science*, **79**(5), pp. 773-84.