Forthcoming in *Philosophy of Science* (Proceedings of PSA'18)

Towards a Taxonomy of the Model-Ladenness of Data

Alisa Bokulich
Department of Philosophy
Boston University
abokulic@bu.edu
https://bokulich.org
https://orcid.org/0000-0002-9406-3904

Model-data symbiosis is the view that there is an interdependent and mutually beneficial relationship between data and models, whereby models are not only data-laden, but data are also model-laden or model filtered. In this paper I elaborate and defend the second, more controversial, component of the symbiosis view. In particular, I construct a preliminary taxonomy of the different ways in which theoretical and simulation models are used in the production of data sets. These include data conversion, data correction, data interpolation, data scaling, data fusion, data assimilation, and synthetic data. Each is defined and briefly illustrated with an example from the geosciences. I argue that model-filtered data are typically more accurate and reliable than the so-called raw data, and hence beneficially serve the epistemic aims of science. By illuminating the methods by which raw data are turned into scientifically useful data sets, this taxonomy provides a foundation for developing a more adequate philosophy of data.

---

1. Introduction. What is the relationship between data and theoretical models? There is what we might call the *folk view of data*, which says that data and models are completely distinct sorts of things. Data just are elements of reality, offering an unmediated window onto the world, while theoretical models are human constructs. Data are always epistemically privileged, this view continues, and any tampering with data is a corruption of that data. When data and theoretical models disagree, it is always the data that win. While this folk view has long been challenged by philosophers of science (e.g., Hanson 1958), it still often shapes our thinking about data; indeed, the alternative view of data that should replace it has remained significantly undertheorized.

One of the most influential early thinkers about the relationship between data and models is Patrick Suppes (1962), who challenges the simplistic view that there are just two things, data and models, which are directly compared to each other. Rather than the raw data, Suppes argues that scientists are more typically interested in 'models of the data', which are processed and abstracted versions of the data that have been subjected to appropriate statistical analyses. Although his insights mark an important advance in philosophical thinking about data, both Suppes and most philosophers of science who follow him have largely black-boxed how data models are produced.

My aim in this paper is to shed light on the various processes by which data are turned into data models. This requires explicitly moving beyond the folk theory of data and recognizing that there is a more complicated relationship between data and theoretical models. Paul Edwards

(1999; 2010) has argued that we should view the model-data relationship as symbiotic, rather than as oppositional. By *model-data symbiosis* Edwards means that there is an interdependent and mutually beneficial relationship between data and models. There are two components to the model-data symbiosis view: On the one hand, models are data-laden, in that large amounts of data often go into the construction and calibration of theoretical models. On the other hand—and more controversially—data are also model-laden, or as Edwards puts it, "model filtered." Although insightful, Edwards' account of model-data symbiosis is limited in that it has not yet been fleshed out nor received much in the way of philosophical analysis. Discussions have also been confined to the context of climate science, potentially leading to the mistaken view that it is unique to this subfield.

In this paper I focus on the second, more controversial component of model-data symbiosis: the model-ladenness of data. More specifically, I construct a taxonomy of the different ways that theoretical and simulation models are used in the production of data. To avoid confusion, rather than Suppes's term 'data model,' I will instead speak of 'data sets'. Furthermore, I will use 'data processing' as a generic term for the methods by which so-called 'raw' data are turned into data sets that can be used by scientists as evidence for or against various claims. Data processing involves all sorts of statistical methods, and my primary focus here will be on those data processing techniques that make substantial use of theoretical or simulation models.

Each of the central sections of this paper describes a model-based data processing technique or way in which data can be model-laden.[1] The taxonomy consists of seven broad categories: data conversion, data correction, data interpolation, data scaling, data fusion, data assimilation, and synthetic data. Each of these is briefly illustrated with an example from the geosciences, specifically hydrology, volcanology, geophysics, seismology, and stratigraphy. This taxonomy, though preliminary, provides a deeper understanding of the relationship between data and models and is an essential component of developing a more adequate philosophy of data.

2. Data Conversion. The first category in the taxonomy—data conversion—is arguably the most widespread, though often not recognized. Many instruments directly measure not the quantity of interest, but rather a proxy, which then must be converted.[2] For example, the raw data provided by a thermometer is the height of a mercury column, which then gets converted into data about temperature, using information about the expansion rate of this metal. In a well-designed and well-calibrated instrument, this conversion is done automatically, essentially black-boxing the complex inferential process that went into the construction and calibration of the instrument (Chang 2004; Tal 2017).

The path from proxy to quantity of interest can be either straightforward or quite complex. In complex cases, data conversion is often mediated by a theoretical model involving idealized representations. An example from geophysics is the use of seismic reflection data to

[1] Eran Tal (2012) has argued for an even more fundamental sense in which models are involved in the production of data in his conception of measurement as a form of model-based inference. This is distinct from (though compatible with) the senses in which data are model-laden that I highlight in this paper.

[2] Data conversion is required in what Parker calls 'derived measurements' (2017, p. 281).

measure subsurface depth. The data quantity of interest is depth, namely how far below the surface some particular reservoir of gas, oil, or water is. The data quantity that is actually collected, however, is two-way travel time of an acoustic signal. In order to convert time (*t*) to depth (*km*) one needs to know the velocity of the signal. The challenge, however, is that the velocity depends on the type of rock, sand, water, etc. it is traveling through, and most subsurfaces are composed of heterogeneous layers.

In order to solve this problem, geoscientists must construct what is called a 'velocity model,' which is a 3D representation of the subsurface that allows one to distribute the attribute of velocity within the cube.[3] More specifically, one must first create a layer model of the subsurface that estimates the subsurface features, such as identifying the different types of rocks and the geometries and thicknesses of these layers. The subsurface model is itself created by combining and extrapolating from various other sorts of data, including data from boreholes, which give more precise information about the different rock types and their depths in a localized region. The information gleaned from these borehole samples then needs to be extrapolated over the larger area of interest, by estimating (e.g., through geostatistics) how the properties vary as one moves laterally away from the borehole control points. Additional data from seismic images is integrated into an idealized subsurface model, which then allows one to associate specific velocities with different points or regions of the 3D velocity model. Depending on the precision and accuracy needed for the depth data, geoscientists will construct a more or less detailed velocity model, allowing one to convert the two-way travel time data used as a proxy into the data quantity of interest, namely subsurface depth.

Depth data measured in this way not only contains uncertainties arising from the two-way travel time measurement, but also uncertainties arising from the velocity model. The important point, however, is that the depth data obtained from the velocity model conversion is less uncertain than depth data collected without such a model. Hence, model-filtered data can be more reliable (have a greater precision and accuracy) than data produced without such a model.

3. Data Correction. A second, common way in which data are model filtered is through data correction. Most data in the geosciences are collected in the field, rather than lab. Because field data are not collected in a controlled experimental environment, they are typically a complex superposition of the signal of interest and various sources of noise. Stephen Norton and Frederick Suppe (2001) introduce a helpful distinction between physical control and vicarious control. Physical control is what one finds in a well-designed controlled experiment: one isolates the variable of interest by physically removing (e.g., by purifying, isolating, or shielding from) other unwanted causes. In many cases, however, physical control is not feasible and unwanted influences must be removed vicariously, which involves measuring or estimating their influence, and removing them mathematically during data processing.

Vicarious data correction is required for gravitational data collected by gravimeters. A gravimeter is a simple but highly sensitive instrument for measuring minute differences in gravity. Gravimeters are used to detect gravitational anomalies, which are differences between locally observed and theoretically expected values of gravity, due to a variation in the density of the underlying rocks. These gravity anomalies can be used to detect geologic features in the upper crust that are relevant to natural hazards, such as earthquake faults.

[3] I owe this characterization to the geophysicist Luca Fava (personal communication).

The units of gravity measurements are cm/s^2 or Gals (Galileos), though gravimeter measurements are typically in mGals. One challenge is that most of the measured variation in the gravity field is not due to differences in the density of the underlying rock that is of geological interest. Instead, the measured gravity data is a composite of this geological signal and a number of other effects that are not geologically meaningful. Hence the "raw" gravimeter data must be corrected by modeling these other effects and subtracting them. For example, one must correct for variations in gravity due to latitude. Although we are taught that the acceleration of gravity on Earth is 9.8 m/s^2, this value actually varies from about 9.78 m/s^2 at the equator to 9.83 m/s^2 at the poles; this "latitude correction" can be up to 5,000 mGals (Keller 2018). A second data correction, the "free air correction," arises from variations in gravity due to elevation, which varies about 0.0386 mGals for every meter. A third correction is the Bouguer correction, which takes into account the mass of material between the base station and gravimeter. This requires constructing an idealized model of the topography and estimating its density. With this topographical model, the gravitational effect of the topography can be subtracted from the measured gravity value. Nearby hills and valleys can also attract the mass of the gravimeter and must be corrected for in a terrain model. In mountainous regions, these corrections can be as much as 10s of mGals (*ibid*).

As we see in this long (and not even complete) list of correction factors, the raw data of gravity measurements are not particularly helpful in detecting gravitational anomalies in the Earth's crust unless the data are first corrected. These corrections are neither negligible nor typically small compared to the effect one is trying to detect. Hence, isolating this data signal requires constructing a subtraction model that vicariously removes all these unwanted components of the data signal. Once again, the model-filtered data are much more accurate than the raw; indeed without these corrections the data are useless for geoscientists.

4. Data Interpolation. A third way in which data are model-laden is through interpolation. Data sampling is typically sparse and uneven in space and time, yet many analysis methods and projects require data values that are regularly spaced. Interpolation methods are ways to fill in the data gaps by estimating or predicting additional data points that were not directly measured. Geostatistics is the branch of statistics concerned with developing such methods. There are many different interpolation methods, ranging from simple nearest neighbor interpolation or smooth surface interpolation, to more complex model-based methods, such as Stochastic Partial Differential Equation (SPDE) model interpolation (Lindgren et al. 2011).

In their simplest form, interpolation methods assume that locations that are closer together are more similar than those far apart, and they seek to estimate the dependency of some variable (or variables) on geographical distance from the observation points. By incorporating more physical information, initial assumptions of isotropy or stationarity might need to be relaxed. SPDE models are spatial models for interpolation that can accommodate these more complicated dependency relations, as well as provide uncertainty analyses for the interpolated data.

Geoscience applications of SPDE model interpolation include constructing contour maps of magnetic field data to find magnetic anomalies indicating iron-ore deposits. The measured magnetic field data is typically sparse over the region of interest. The SPDE model-based interpolation method can be used to estimate some of the missing data points, and provide a more detailed contour map that delineates the location of the deposits. In a recent paper Guo and

colleagues test the SPDE model-based interpolation method against both other interpolation methods and real field data and conclude that

> the predicted values are quite close to the observed data. Compared to [other interpolation methods]. . . the SPDE model-based method provided the smoothest and most accurate map. The borehole data have verified that the SPDE method gives more useful detailed information. (Guo et al. 2019, p. 11)

Geophysical field data are thus model-laden in that a proper subset of the data points are not in fact directly measured or observed, but rather are estimated using sophisticated model-based statistical interpolation techniques.

5. Data Scaling. Often in geoscientific research there is a discrepancy of scales: data is collected at one scale, models operate at another, and decision makers are asking questions at yet a third scale. These discrepancies can involve differences in both spatial and temporal scales. To solve this problem, scientists must develop various methods to transfer data and information from one scale to another: When this involves going from a small (or short time) scale to a large (or long time) scale, it is called *upscaling*. When it involves going from a large (or long time) scale to a small (or short time) scale it is called *downscaling*.

Consider hydraulic conductivity, which describes the ease with which a fluid (such as water) moves through pores and cracks in a material (e.g., rock or soil); it depends on a number of factors, such as the intrinsic permeability of the material, the degree of saturation, the viscosity of fluid. The data for hydraulic conductivity is typically collected in the lab from measurements made on relatively small sediment cores. However, numerical groundwater models, often concerned with entire aquifer systems spanning hundreds of square kilometers, require values of hydraulic conductivity at the much larger scale of the model blocks (Bierkens et al. 2001, p. 4). Hence the hydraulic conductivity data must be upscaled for use in groundwater models.

A key obstacle to transferring data across scales is heterogeneity, and in such cases models are essential for upscaling or downscaling data. Hydraulic conductivity is such a heterogeneous quantity: it can vary up to several orders of magnitude within an aquifer, making upscaling nontrivial. Fernàndez-Garcia et al. (2009) discuss the use of multirate mass transfer models (MRMT) as a constitutive equation for upscaling solute transport (the movement of substances like fertilizers or pollutants in ground water) which depends primarily on hydraulic conductivity. The MRMT model is a way of representing a wide variety of mass transfer processes occurring across a wide variety of scales at the same time. They can be used to upscale the observed hydraulic conductivity data into a larger support volume, and thus make this data useful for answering questions about the transport of pollutants on the larger scale of interest.

The discrepancy of scales between data, models, and decisions can be understood as one component of what has been termed the tyranny of scales problem, which has attracted the attention of philosophers of science (e.g., Batterman 2013). The use of models to upscale and downscale data, described here, has emerged as one important strategy for taming the tyranny of scales.

6. Data Fusion. Data fusion, also known as data integration, refers to a broad family of methods for combining heterogeneous data sources into a coherent and improved data product (Wald 1999). One example of model-based data fusion is the integration of various types of stratigraphic data in the construction of global timelines. The stratigraphic record contains many different types of data: Physical events (including volcanic ash falls and tuffs, geomagnetic polarity reversals, meteorite impact deposits), chemical events (including stable isotope excursions, wide-spread anoxic events), biological events (first and last appearances of taxa, mass extinction horizons, ecological events) and others. The task is to merge these diverse types of data and construct a time ordered data set of "as many ancient evolutionary, ecological, geochemical and geophysical events as possible" (Sadler 2014, p.4). There are several challenges, however, that make this a nontrivial task.

One challenge is that these stratigraphic data are gathered locally, at particular outcrops or well cores, but need to be correlated with data gathered at many different, spatially-distant locations. Although the principle of superposition dictates that lower stratigraphic layers are older, this principle only applies to data collected at a single location. However, as Sadler notes, "global patterns of cause and effect do not emerge until geologic strata of the same age can be identified across widely dispersed locations (Sadler 2004, p. 188). Furthermore, these stratigraphic records at distant locations are incomplete, fallible, and sometimes even contradictory. Thus, despite vast quantities of data, there is a problem of underdetermination regarding how exactly the data should be coherently integrated into a single timeline. Such projects are an example of what Edwards, in the context of climate modeling, calls *making data global*, which he defines as "building complete, coherent, and consistent global data sets from incomplete, inconsistent, and heterogeneous data sources" (Edwards 2010, p. 251).

In order to integrate these vast quantities of data—in ways that respect the different levels of uncertainty and different constraints that these diverse kinds of physical, chemical, and biological data must satisfy—models are essential. There are several computer models that implement different algorithms for integrating the diverse types of data. These can differ in the ways they search for an optimal ordering of events and in the ways they measure misfit. Most of these programs contain theoretically substantive rules about permissible adjustments based on the different types of data. Sadler, who developed one such computer model, CONOP, has analogized the different types of stratigraphic data and their uncertainties to being either like nails, jacks, or clamps (Sadler 2012; Sadler et al. 2014). Volcanic ash fall or tuff, which are some of the few events that can be radiometrically dated, are analogized to 'nails' that have a fixed position in the timeline. Local taxon ranges, defined by the first appearance of a fossil taxon and its last appearance in the fossil record, are analogized to jacks that can be stretched apart to fit a composite sequence, since they are unlikely to represent the true first or last appearance of the taxon. Other data, like stable isotope excursions or paleomagnetic reversals between samples of opposite polarity, by contrast, are treated like clamps—conservative uncertainty intervals that can be squeezed to fit the same timeline of events. As Sadler explains, "sequencing algorithms seek an optimal time line to which all sections can be fit without moving any of the nails and by the least expenditure of energy to extend the jacks and squeeze the clamps" (Sadler 2012, p. 330).

These computer models are an essential tool in integrating the vast quantities of different types of stratigraphic data in physically meaningful ways. Although the time-ordering solutions for these data sets may not always be unique, "the set of equally good solutions leads directly to explicit uncertainty statements" (Sadler et al. 2014, p. 5). Model-based data fusion, thus, is

important not only for constructing data sets that can be useful as evidence for various theoretical hypotheses (in a way the nonintegrated data are not), but also provide a way to quantify the uncertainty inherent in the fused data set.

7. Data Assimilation. A particularly close form of model-data symbiosis is data assimilation, which is defined broadly as the optimal integration of data with dynamical model estimates to provide a more accurate 'assimilation estimate' of the quantity (or geophysical field) of interest. Data assimilation methods take into consideration the uncertainties associated with both data and models, and draw the estimate closer to whichever has the lower uncertainty, while incorporating relevant information from the other source. Data assimilation algorithms can be quite complex and can involve many of the previously discussed model-filtering methods, such as model-based interpolation, extrapolation, scaling, correction, and data fusion. What typically sets data assimilation methods apart is that they are aimed at producing dynamic, regularly updated, estimates and forecasts, rather than static assessments.

Data assimilation methods were first developed in meteorology, and in that context have been discussed by Wendy Parker (2016, 2017) who analyzes the implications of these methods for the epistemology of measurement. Data assimilation methods have spread to other geosciences, such as oceanography, hydrology, and volcanology. Reichle notes "in the Earth sciences, data assimilation involves nonlinear, highly complex, and exceedingly large systems with complicated error structures that defy the straightforward application of classical optimization methods" (2008, p. 1413). Data assimilation can be used to target different sources of uncertainty: in atmospheric and oceanic studies where dynamics are chaotic, the focus of data assimilation is often on the estimation of initial conditions, whereas in land surface dynamics the focus is instead on uncertain forcing or boundary conditions and model parametrizations (*ibid*).

In volcanology, data assimilations methods are being used to develop forecasting methods for volcanic eruptions (Zhang and Gregg 2017). The ascent and accumulation of magma prior to an eruption can lead to subtle ground deformation and gravity changes, which can be measured and studied by volcano geodesy. As Fernández and colleagues note,

> Recent decades have seen an explosion in the quality and quantity of volcano geodetic data. . . . [H]owever, [with this] comes a need for new approaches to analysis, modeling, and interpretation. (Fernández et al. 2017, p. 1)

In order to draw correct inferences about what is going on below the surface of the Earth, the data need to be combined with models that explore "the geometry and volume of plumbing systems, which offer critical context for interpreting the mechanisms and characteristics of unrest and eruption" (*ibid*). Both inverse and (forward) numerical models are used. The geodetic data are complicated by the fact that they are a composite of not just the 'magmatic signal', but also other signals such as tectonic changes and anthropogenic changes (e.g., due to ground water extraction); hence, subtraction models and other model-based data correction methods (section 3) are required. The power of geodetic data in forecasting eruptions is also increased by integrating it with other kinds of data (section 6), such as seismic and gas emission data (Fernández et al. 2017, p. 1). Zhang and Gregg conclude "developing data assimilation strategies to incorporate the vast array of volcano monitoring data sets into increasingly sophisticated geodynamics models is critical for future efforts to assess volcanic unrest" (2017, p. 23). As we see in this example, data assimilation is an intimate form of model-data symbiosis with the potential to lead

to more powerful and useful information than either the (geodetic) data or the (geodynamics) models alone, and moreover give near real time forecasts, relevant to natural geohazard management.

8. Synthetic Data. If we think of the mixing of models and data as a spectrum, with pure (raw) data at one end and pure simulation model products at the other, synthetic data would fall at the extreme of the model end. Synthetic data (also referred to as 'virtual observations') are data sets produced not by any measurements or observations of the world, but rather by the output of simulation models. They are included in the taxonomy because synthetic data play a crucial role in testing, refining, and sometimes implementing the data processing techniques discussed in the previous sections. Thus, they play a beneficial role in the production of real data sets, characteristic of model-data symbiosis.

Real data—though often mediated in complex ways by models and computer simulations—are distinguished from synthetic data by the fact that they are the result of a physical interaction with the world (Parker 2017). Insofar as the simulation models that produce synthetic data are based on empirically-derived theories, laws, and principles, they are not entirely divorced from the empirical world, though this point should not be overstated. As synthetic data become more common, it is critical for users to know the provenance of the data they are working with, and to what extent they are real or virtual observations (e.g., Beven, Buytaert, and Smith 2012). As Parker (2009, 2016, 2017) has cogently argued, both in the context of debates about the difference between physical experiments and simulation experiments and in the context of data assimilation, the key issue to keep sight of is the degree of epistemic reliability.

Most of the model-based data filtering methods discussed in previous sections have made use of synthetic data in testing the adequacy, and improving the reliability, of these methods. For example, in data conversion (section 2), synthetic data have been used to test the sensitivity of the time-depth converted data to errors in the velocity model. In the context of data correction (section 3), Bokulich (2018) discusses how synthetic data are used to test phylogenetic model-based data correction methods that attempt to reconstruct more accurate paleodiversity data from the incomplete and biased fossil record. By starting with synthetic, and hence exactly known, paleodiversity data, one can subject that data to various known biases and deletions, and see how well the data correction methods perform at recovering the true initial paleodiversity. Such simulation studies reveal not just the degree of reliability or robustness (under error) of these methods, but also where these data correction methods are likely to break down or systematically lead one astray.

Synthetic data have also been used to test data fusion methods (section 6), such as in the stratigraphic data-integration methods used to time correlate distant biostratigraphic events (Edwards 1984). Similarly, synthetic data were used to develop and test the reliability of the data assimilation in volcano geodesy.

Synthetic data thus play a critical role in testing and refining many of the model-based data correction methods, and help ensure the production of more accurate real data sets. As Bokulich (2018) notes, although a data-correction method's ability to work well on synthetic data does not prove its reliability in the more complicated context of real data, it does provide an informative minimum constraint: a data-correction method that fails to perform well on synthetic data is unlikely to succeed in the real world.

9. Conclusion. Here I examined seven prominent ways in which data can be model-laden: data conversion, data correction, data interpolation, data scaling, data fusion, data assimilation, and synthetic data. While these capture the most common model-based data-processing methods in the geosciences, there is further work to do in elaborating and refining this taxonomy in light of new examples. Although these categories are conceptually distinct, a model-based data-processing technique can simultaneously perform more than one of these functions, such as in the case of data assimilation, which as we saw can effect a data correction, interpolation, and fusion within one method.

With this taxonomy, we can now see how the folk view of data is inadequate. This view, recall, can be understood as consisting of four central claims: The first is that data and models are completely distinct sorts of things. As we have seen, although data and models are conceptually distinct, in practice the line between them is often blurred, not just in extreme cases such as data assimilation, but also in more mundane cases of data conversion.

The second claim of the folk view is that data just are elements of reality, or marginally better, are an unmediated window onto reality. Although data are certainly a window onto the world, they are a *mediated* window. As Sabina Leonelli has emphasized, "despite their scientific value as 'given,' data are clearly made. They are the results of complex processes of interaction between researchers and the world" (Leonelli 2016, p. 71). Hence data should not be equated with world.

The third claim of the folk view is that data are always most epistemically reliable, trumping theory and models whenever they clash. As we have seen, however, data are also subject to bias and error, and should be understood as having an associated uncertainty, just as model predictions do. A high-profile case where data were judged to be less epistemically reliable than models was the 2011 conflict between satellite and weather-balloon data and climate models, which came before a U.S. Congressional hearing (Lloyd 2012 and references therein). As Elisabeth Lloyd recounts, "in the end (and in short) it now appears that the models were mostly right and the early data were mostly wrong" (p. 391). Lloyd argues that this case shows the need for a more "complex empiricism" than has been previously recognized.

The fourth claim of the folk view of data is that any tampering with data is a corruption of the data and a lowering of its epistemic reliability. As I hope has become clear in the many examples from the geosciences discussed here, at least some "tampering" with data in fact *increases* its epistemic reliability. Most importantly, it is not a question that can be judged a priori, and instead depends on an investigation of the details of the case.

Why can't geoscientists just use the raw data? Why do they need to model-filter their data at all? As we have seen, there are many reasons. First, as we saw, geoscientific data are typically collected in the field, rather than in a controlled laboratory setting. Hence the data are a complex admixture of various signals and noise. Second, despite the vast and rapidly growing quantities of geoscientific data, they are still often insufficient to answer many of the questions of interest, and hence there is an underdetermination, unless the data are interpolated, augmented, or interpreted with the help of various models. Third, many quantities of geoscientific interest cannot be measured directly, and hence require conversion from a measured proxy quantity. Fourth, there are often multiple, diverse—and sometimes conflicting—sources of data about a quantity of interest, which need to be adjudicated and combined in complicated ways. And, finally, the geosciences must grapple with vast spatial and

temporal scales, which typically do not correspond to the scales at which data can be gathered. For all these reasons and more, the model-filtering of data is needed.

A central aim of this paper has been to make plausible the prima facie counterintuitive claim that model-filtered data can—in some instances—be more accurate and reliable than so-called raw data, and hence beneficially serve the epistemic aims of science. My claim is not that model-laden data are always more epistemically reliable; the model-based processing of data, if not done appropriately, can introduce artefacts into the data and systematically mislead researchers. All these model-based data-filtering techniques are fallible to varying degrees in different contexts, and hence their epistemic reliability should be rigorously assessed, and not assumed. It is essential for the successful development and deployment of model-data symbiosis that the uncertainties associated with both data and models be quantified and appropriately propagated. Furthermore, these data-processing techniques must be rigorously tested by all available means, including through the use of synthetic data. There is much work for methodologically reflective scientists and philosophers of science to do in sorting out cases where model-data symbiosis may be problematic or circular. The preliminary taxonomy of the model-ladenness of data given here opens up the conceptual space to undertake such further evaluative projects, and provides a valuable foundation for the continued development of a more adequate philosophy of data.

ACKNOWLEDGEMENTS

This paper was written while a visiting researcher at Institute for Advanced Study at Durham University and I gratefully acknowledge the financial support of the European Union COFUND Senior Research Fellowship under EU grant agreement number 609412. I am especially grateful to Wendy Parker for serving as my host while there and for many stimulating discussions about this topic.

REFERENCES

Batterman, R. (2013), "The Tyranny of Scales" *The Oxford Handbook of Philosophy of Physics*. Oxford: OUP, pp. 255-286.

Beven, K., W. Buytaert, and L. Smith (2012) "On Virtual Observatories and Modelled Realities (Or Why Discharge Must Be Treated as a Virtual Variable)" *Hydrological Processes* 26: 1905-1908.

Bierkens, M., P. Finke, and P. de Willigen (2000), "Upscaling and Downscaling Methods for Environmental Research" Dordrecht: Kluwer Academic. Preprint available (2001): https://www.researchgate.net/publication/40186896_Upscaling_and_Downscaling_Methods_for_Environmental_Research.

Bokulich, A. (2018), "Using Models to Correct Data: Paleodiversity and the Fossil Record" *Synthese* https://doi.org/10.1007/s11229-018-1820-x.

Chang, H. (2004), *Inventing Temperature: Measurement and Scientific Progress.* Oxford: Oxford University Press.

Edwards, L. (1984), "Insights on Why Graphic Correlation (Shaw's Method) Works" *Journal of Geology* 92: 583-597.

Edwards, P. (1999), "Global Climate Science, Uncertainty and Politics: Data-Laden Models, Model-Filtered Data" *Science as Culture* 8 (4): 437-472.

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

Fernández, J., A, Pepe, M. Poland, and F. Sigmundsson (2017), *Journal of Volcanology and Geothermal Research* 344: 1-12.

Fernàndez-Garcia, D. G. Llerar-Meza, and J.J. Gómez-Hernández (2009), "Upscaling Transport with Mass Transfer Models: Mean Behavior and Propagation of Uncertainty" *Water Resources Research* 45(W10411): 1-16.

Guo, Z. X. Hu, J. Liu, C. Liu, and J. Xiao (2019), "Geophysical Field Data Interpolation Using Stochastic Partial Differential Equations for Gold Exploration in Dayaoshan, Guangxi, China" *Minerals* 9 (14): 1-12.

Hanson, N. ([1958] 2010), *Patterns of Discovery: An Inquiry into the Conceptual Foundations of Science.* Cambridge: Cambridge U. Press.

Keller, G. R. (2018), "Using and Understanding Gravity Data" (Accessed October 2018). https://research.utep.edu/Default.aspx?PageContentID=3947&tabid=38186.

Leonelli, S. (2016), *Data-Centric Biology: A Philosophical Study.* University of Chicago Press.

Lindgren, F., H. Rue, and J. Lindström (2011), "An Explicit Link between Gaussian Fields and Gaussian Markov Random Fields: The Stochastic Partial Differential Equation Approach" *Journal of the Royal Statistical Society, Series B: Statistical Methodology* 73: 423–498.

Lloyd, E. (2012), "The Role of 'Complex' Empiricism in the Debates about Satellite Data and Climate Models" *Studies in History and Philosophy of Science* 43: 390-401.

Norton, S. and F. Suppe (2001) "Why atmospheric modeling is good science" in P. Edwards, C. Miller (eds.) *Changing the Atmosphere: Expert Knowledge and Environmental Governance*. Cambridge: MIT Press, 67-106.

Parker, W. (2009), "Confirmation and Adequacy-for-Purpose in Climate Modelling" *Proceedings of the Aristotelian Society Supplementary Volume* LXXXIII: 233- 249. doi: 10.1111/j.1467-8349.2009.00180.x

Parker, W. (2016), "Reanalyses and Observations: What's the Difference?" *Bulletin of the American Meteorological Society* 97: 1565–1572. https:// doi.org/10.1175/BAMS-D-14-00226.1.

Parker, W. (2017), "Computer Simulation, Measurement, and Data Assimilation" *British Journal for the Philosophy of Science* 68: 273-304.

Reichle, R. (2008), "Data Assimilation Methods in the Earth Sciences" *Advances in Water Resources* 31: 1411-1418.

Sadler, P. (2004), "Quantitative Biostratigraphy—Achieving Finer Resolution in Global Correlation" *Annual Review of Earth and Planetary Science* 32: 187-213.

Sadler, P. (2012), "Nails, Jacks and Clamps: Mechanical Analogs as a Way to Think About Organizing Vast and Varied Stratigraphic Information Sets for Computer Algorithms that Optimize Time Lines and Calibrate Time Scales" Geological Society of America *Abstracts with Programs* 44 (7): 330.

Sadler, P., R. Cooper, J. Crampton (2014), "High-Resolution Geobiologic Time-Lines: Progress and Potential, Fifty Years after the Advent of Graphic Correlation" *The Sedimentary Record* 12: 4-9.

Suppes, P. (1962), "Models of Data" in E. Nagel, P. Suppes, and A. Tarski (eds.) Logic, Methodology and Philosophy of Science: Proceedings of the 1960 International Congress. Stanford: Stanford U. Press, pp. 252-261.

Tal, E. (2012), *The Epistemology of Measurement: A Model-Based Account*. PhD Dissertation. University of Toronto, http://hdl.handle.net/1807/34936.

Tal, E. (2017), "Calibration: Modeling the Measurement Process" *Studies in History and Philosophy of Science* 65-66: 33-45.

Wald, L. (1999), "Some Terms of Reference in Data Fusion" *IEEE Transactions on Geoscience and Remote Sensing* 37(3): 1190-1193.

Zhan, Y. and P. Gregg (2017), "Data Assimilation Studies for Volcano Geodesy" *Journal of Volcanology and Geothermal Research* 344 (2017) 13–25.