On choosing the resolution of normative models

James H. Merrick,* John P. Weyant*

April 19, 2019

Abstract

Long time horizon normative models are frequently used for policy analysis, strategic planning, and system analysis. Choosing the granularity of the temporal or spatial resolution of such models is an important modeling decision, often having a first order impact on model results. This type of decision is frequently made by modeler judgment, particularly when the predictive power of alternative choices cannot be tested. In this paper we show how the implicit tradeoffs modelers make in these formulation decisions, in particular in the tradeoff between the accuracy of representation enabled by the available data and model parsimony, may be addressed with established information theoretic ideas. The paper provides guidance for modelers making these tradeoffs or, in certain cases, enables explicit tests for assessing appropriate levels of resolution. We will mainly focus on optimization based normative models in the discussion here, and draw our examples from the energy and climate domain.

Keywords: *Problem structuring; Validation of OR computations; Information theory; Strategic planning; OR in environment and climate change.*

Accepted Manuscript

James H. Merrick and John P. Weyant, On choosing the resolution of normative models (2019). *European Journal of Operational Research*, Article in Press.
Published version available at: https://doi.org/10.1016/j.ejor.2019.06.017

*Department of Management Science & Engineering, Stanford University. Postal Address: Huang Engineering Center, 475 Via Ortega, Stanford, CA 94305, USA. Corresponding Author: James Merrick, jmerrick@alumni.stanford.edu

1 Introduction

This paper considers how much detail to represent, and how to represent it, in normative models used in policy analysis, system analysis, and strategic planning. While the representation of detail, such as the temporal or spatial resolution of such models, is often decided by modeler judgment, particularly in those cases where predictive power of models cannot be tested, this paper provides a systematic basis for choosing an appropriate resolution. In so doing, it finds the information theoretic concept of the Information Bottleneck (Tishby et al., 1999) useful due to the natural translation of its premises to intuitive and commonly expressed model formulation goals. The resulting discussion demonstrates the tradeoff between model accuracy, given available data, and model detail, in the context of normative models. More broadly, this paper may be considered a manifestation of the trend of increased use of statistical ideas to help parameterize and formulate models.[1] Operations research has often focused on the development of methods to solve a given problem (Denton, 2017), with the process of problem formulation itself more of an art than a science.

While the current state of the art in model formulation in the space of this paper typically rests on expert judgment, from the authors' experience, these formulation decisions often appear to be made in a somewhat ad hoc manner. The ad hoc nature of the model formulation decisions can stem from data availability, computational constraints, or simply an incomplete understanding of the system being modeled. Incomplete understanding can be particularly serious in interdisciplinary modeling efforts where modelers are integrating expertise across multiple disciplines.

To address how to formulate such models systematically, particularly the choosing of model resolution, this paper begins with (a) two intuitive premises on the goals of model formulation, and (b) the premise that model formulation decisions may be assessed conditionally upon one another, and proceeds to develop principles to assess whether a model resolution is appropriate or otherwise. Importantly, these premises have grounding in disparate literatures, as opposed to solely the authors' experience of good practice.

The paper generally adopts a discursive approach, while adopting an analytic approach at points to make ideas more concrete. At some junctures, the concepts of data and model will be interchangeable, for example where the idea of compressing model detail essentially implies compressing the input data of that model. Analysts experienced with the design of models may find the resulting ideas corresponding to intuition and simply established best practice. However, an explicit logical foundation may allow the further development of best practice, along the way guiding attempts to resolve a number of outstanding applied modeling problems, and guide how, when, and if, increased data availability can help a model inform a decision. That there is value in doing so is based on the premise that such work can help harness ongoing algorithmic and computational advances to complement the skills of human analysts in addressing important questions facing society.

Section 2 provides context on resolution, normative models, and assessing model formulation in general. Section 3 introduces intuitive goals of model formulation and how they can be linked to

[1] Another aspect of this trend is the advancement in methods to treat uncertainty, for example Wang et al. (2016) present an approach that exploits available data to develop a robust optimization method that excludes distributions unlikely under that data. Uncertainty, and modeling under uncertainty is at the core of this paper. If we were certain about how to model the system, we would not be having this discussion.

information theoretic ideas. Section 4 adapts these ideas to our normative modeling context, while Section 5 applies them to the question of model resolution in particular. Section 6 discusses some illustrative examples of the ideas, particularly energy and environmental applications. Section 7 attempts to distill the journey into a roadmap for the modeling community on choosing resolution, before Section 8 concludes the paper.

2 Background

2.1 Normative models

In this paper, we discuss normative models whose primary purpose is to inform a decision, through exploring or choosing *what can be* as opposed to *what is*. Furthermore, we will generally consider models that use optimization concepts to search and understand this 'what can be' space. We use the term 'normative model' broadly, as a useful distinction from other types of models that support decisions, with which there are many linkages, for example statistical / econometric / machine learning models. As an example, the models for energy/economic policy analysis discussed in Kann and Weyant (2000) fit within the focus of this paper.[2] Further examples include models for long term planning (public and private),[3] and models for the engineering of new structures and products that have not been deployed before and where prototypes/pilots are expensive. We consider these models as decision *support* tools, as opposed to models that fully automate decisions, and thus interpretation of model output is important to provide that support. In our experience, the key to interpretation is succinctly stated by Geoffrion (1976); "One must know not only *what* the optimal solution is for a given set of input data, but also *why*".

Models of this type often have the property that generation of observations along one potential decision pathway is unrepeatable due to either the cost or the simple irreversibility of certain processes, or both. Additionally, learning from past experience is the basis on which to inform our actions in the future, and when the representativeness of this past experience is in question, we find ourselves in a challenging analytical situation. An interesting feature of this broad class of models is that the factors that make prediction difficult and that strain the assumption of the past being a basis for future processes are also the factors that make the future more amenable to being shaped (Lempert, 2003). This allows a vision of modeling that fits within Simon's (1981) articulation of a 'Science of the Artificial' or 'Science of Design'. The formulation of these models is generally referred to as the art of modeling, whereas we introduce concepts for systematic assessment of one particular aspect of model formulation, the choosing of its resolution.

In some of the more abstract discussion in this paper, we will represent a model as an object, f, with the vector $\boldsymbol{x}^* \in \Re^n$ representing the output of the model f at solution. As we will discuss, a model is best evaluated in the context of the question it is designed to address. We will refer to

[2] An example in a different context is that of the model employed by Herrera et al. (1976), where the ideas discussed in this paper of would be relevant, for example, to the choosing of spatial resolution, independent of other aspects of model formulation within that different context.

[3] One such example is at the global climate policy level, where Integrated Assessment Models (IAMs) (see Weyant (2017)) inform much of the literature comprising the report of the Intergovernmental Panel on Climate Change (IPCC) Working Group III (IPCC, 2014), and by extension much of the international negotiation on climate agreements. A private planning example is a company designing a network of electric vehicle charging locations, or a similar infrastructure network.

the question asked of the model as Q, and in the more abstract discussion, it can be represented mathematically as a random variable, where model outputs inform the probabilities associated with questioner-defined states of the world.

2.2 Resolution

Choosing resolution, for example temporal, spatial, or demographic resolution, comprises an important subset of model formulation decisions. To illustrate the importance in the energy/environment domain for example, DeCarolis et al. (2017) state that the setting of spatio-temporal boundaries is a key step in formalizing best practice for energy system optimization modeling, Hamilton et al. (2015) include spatial and temporal scales amongst the ten 'salient dimensions for integrated assessment modeling', while Pfenninger et al. (2014) list 'resolving details in time and space' and 'complexity and optimization across scales' as two of four challenges in energy systems modeling for twenty-first century energy challenges. Furthermore, Merrick (2016) addresses the particular question of temporal representation in electricity capacity planning models, showing how it can influence model results, and analyzing alternate methods of representation.

Assuming an optimization approach, a model can be abstractly formulated as follows

$$\text{minimize } c(\boldsymbol{x}) \qquad \text{Subject to: } \boldsymbol{x} \in \mathcal{X} \tag{1}$$

Where c is a function, and $\boldsymbol{x}$ a vector. Choosing the level of model detail, or as we term it here, choosing the model resolution, entails deciding, for the problem at hand, the nature of the set $\mathcal{X}$, that is, the dimensionality of $\mathcal{X}$ and what sort of constraints define it, whether they be integer, or linear, or conic, or convex constraints, etc. Ideally, this paper will be useful in the whole model formulation process, but should be particularly useful for those cases where one conceptually decides the model formulation, but then wants to compress it upon implementation for computational tractability, or simply for understanding of the problem. Examples of such compression steps include the aggregation of model variables and/or the relaxation of model constraints. Examples of modeling questions to which resolution relates include:

- Spatial questions: In a question of global climate policy, should every country be modeled, or is continent level aggregation, or simply one global region, appropriate?
- Temporal questions: In planning models for electricity systems, should every second be modelled, or every hour, or is a well-chosen subset of periods appropriate? How to find such a well-chosen subset?
- Demographic questions: In modeling the economic impact of policy, how to represent heterogenous demographic groups? Is an aggregate representative agent approach appropriate?

As will be briefly discussed, the choosing of resolution can also be extended to the treatment of uncertainty when formulating a model through the representation of the space modeled in stochastic optimization or in the choosing of scenarios when conducting scenario analysis.

In the literature on aggregating optimization models, Rogers et al. (1991) develop a framework for aggregation and disaggregation methodology, while Zipkin (1980a,b) develops a range of a priori and a posteriori bounds relating to aggregation. The bounds are on the objective value, which may not be the only source of useful information from model outputs. Clustering techniques play a large role in the approaches discussed. This paper attempts to harness and build on these results relating to model aggregation in the normative setting by overlaying information theoretic and decision analytic value-of-information ideas.

2.3 Model evaluation

To choose model resolution, we can consider how to evaluate different model formulations. A variety of terms are used across various scholarly fields to describe processes to assess model appropriateness. This range of terms includes model selection, model evaluation, model verification, model validation, and model confirmation. We will consider here a subset of the different concepts and approaches, highlighting the challenge of testing predictive power, and more philosophical issues with verification of numerical models.

The 'model selection problem' is a well established problem, and it can be argued that the process of finding a model that explains phenomena we observe rests at the core of much scientific research. Associated methods typically assume relevant data availability however and are more amenable to aiding the 'what is' type question as opposed to the 'what can be' type questions that are the subject of this paper. The various research communities, e.g. econometrics, machine learning, statistics itself, are not unaware of 'missing data' or out-of-sample applications of a model. For example, in the machine learning community, the inclusion of regularization parameters is important to temper risks of overfitting. The success of a model is evaluated by its performance on a previously unseen dataset that it was not trained on, i.e. the test in machine learning approaches is predictive power. The challenge for the models discussed in this paper is the lack of ability to assess predictions, in addition to prediction not necessarily being the purpose of the model in the first place. Given the large impact these prediction-tested learning models have on many different fields in recent times, it is necessary for those in the normative modeling community to consider the relevance of these developments and approaches to normative models, a topic with much further research outstanding.[4]

If we can answer the what-is type questions, i.e. understand the initial conditions and dynamics of today's system, can we then simply work out how things will evolve in the future under different actions? This is essentially the basis of the formulation of models that are used today, and improved knowledge of the current state and of dynamics undoubtedly helps our understanding of how things may evolve in the future. There are numerous limitations to this approach however, including (a) the dynamics that best represent the system today do not necessarily represent accurately how that system will change in the future due to either designed changes or changes in the environment, and (b) the possibility from chaos theory that shows how tiny variations in input data can produce outputs from a deterministic series of equations that are unpredictable in practice.[5]

Verification and validation are terms that can have definitions that vary from field to field[6], with the former sometimes relating to correctness in implementation of a model, and the latter

[4] More general than evaluation on the basis of prediction is evaluation on the basis of performance. Reinforcement learning algorithms trained to play video or strategy games, for example, are assessed upon in-game performance.

[5] For more on unpredictable outputs and deterministic equations within the context of chaos theory, see Bishop (2017).

[6] Sargent (2013) discusses verification and validation in the simulation modeling context, proposing a holistic approach from conceptual model validity through to the accuracy of coding implementation.

conceptual correctness. Landry and Oral (1993) note this distinction between 'doing things right' (correctness in implementation of a model) and 'doing the right things' (conceptual correctness) when modeling, stating that the literature at that point typically focussed on the former more than the latter. The challenge of defining validity is illustrated by Déry et al. (1993), who bring in a conclusion from epistemology that there are no universal formal criteria of validity. In this vein, Oreskes et al. (1994) hold the position that a numerical model simply cannot be validated, in the sense that the truth of a statement involves starting with a closed series of premises and concluding whether the statement can be arrived at following a sequence of logical deductions. With this framing, numerical models do not meet this criterion generally, as the models are not closed systems. They are not closed systems in the sense that 'auxiliary hypotheses' are required to allow the representation of the real world run on a computer. If a model does not replicate observed data, there is no way of knowing whether it is a primary hypothesis or auxiliary hypothesis that is at fault. Similarly, when model outputs match real world data, we do not know if this model is uniquely correct without invoking some extra criterion such as Occam's Razor. This leads Oreskes et al. (1994) to conclude that numerical models can be at best confirmed, not validated.

Occam's Razor will feature in our goals for model formulation proposed later in this paper. This paper does not address all these challenges outlined above but it aims to contribute to the literature by providing a new perspective and actionable tests for a subset of model formulation decisions. Meanwhile, the conceptual challenges in model evaluation outlined here point to the necessity for quality and care in modeling practice, a topic discussed later in this section.

2.4 Value of a model

To investigate any aspect of normative model formulation, we will take a step back and consider what is the value of a model. In particular, as a model provides information, we will outline two metrics used to assess the value of information. Firstly, the Decision Analysis assessment of the value of information is whether the provided information leads to a different decision than would be made in the absence of this information, with the magnitude of the value of information the difference in expected outcome between the two decisions (Howard and Abbas, 2015). Numerous units can be used to express this value, for example dollars or utility. Secondly, a more abstract representation of the value of information is discussed by Volkenstein (2009). In this formulation, the value of information is defined to be $-\log_2(p'/p)$, with p' and p representing the probability of achieving some goal before and after the receipt of the information respectively.[7] This connection of model value to the furthering of a goal relates to the school of thought that the ultimate and simple test of a model's value is whether it is used or not. The *"Insights, Not Numbers"* objective for modeling that is frequently put forward as an alternative to a forecast accuracy criterion is related to this school of thought in the sense of a model being of value in its use through delivering new insights into a problem as opposed to the actual numbers generated.

The two measures of the value of information above are related in that probability is included in both metrics (implicitly in the decision analysis case), and cost is potentially included in both metrics, this time implicitly, and not necessarily always, in the second. Perhaps the main contribution of the second metric for our purposes is the introduction of the $\log_2 p$ term. This term is familiar

[7] In this formulation, the greater p' relative to p, the greater the value of information. If $p' = p$, the value of information is 0 ($log(1) = 0$).

in information theory and the Shannon concept of information, where the amount of information in a message is inversely proportional to the probability that you are expecting the content. This notion of surprise can be seen, for example, in the discussion on 'interestingness' by Silberschatz and Tuzhilin (1996) that can be found in the data-mining literature. These value of information ideas will be implicitly incorporated into the discussion below on appropriate choosing of model resolution.

2.5 Modeling practice

Kasparov (2010) tells a story of how a mediocre chess player with a mediocre computer, but with an excellent process of using the computer, can beat a chess grandmaster with a mediocre process of using a supercomputer with the most sophisticated chess algorithms. This story illustrates the importance of how models are used in harnessing their potential.

Numerous works have been written on best practice in modeling, an important subject given the lack of systematic methods for checking validity. For example, Jakeman et al. (2006) outline "ten basic steps to good, disciplined model practice", Pidd (1999) offers six principles of modeling, principles developed from modeling experience, Nestler (2011) discusses reproducible research, and Murphy (2005) in the first of a series of 3 papers, introduces 'Elements of a Theory of the Practice of Operations Research'. Morgan et al. (1998) is an extensive text on applications of modeling methods to policy analysis, offering similar recommendations to Pidd (1999) on the value of simple models based on complicated thinking.

Given the role of expert judgment in model evaluation, empirical studies on the process potentially offer a path to understanding this judgment, in turn enabling the improvement of the modeling process. An early example of such a study is Willemain (1995), where a number of OR practitioners are observed as they formulate a model, and a recent example is Hämäläinen et al. (2013), who show that behavioural effects, related to how questions are asked and how information is presented to modellers, can be exploited such that the results from a modeling exercise are completely reversed. This paper, in assessing formulation issues systematically, will ideally aid the avoidance of such undesirable effects.

If we consider modeling practice as a design process, the literature on modeling the design process itself provides a link to some further relevant literature. In particular, the 'design cycle' of Takeda et al. (1990) from this literature resonates with the Decision Analysis Cycle of Howard (1968). Both cycles comprise sequential steps through phases of the design / decision process, with feedback loops to update prior phases with learnings from the later phases. In a somewhat similar fashion, Schneeweiss (1987) systematically considers the formulation process in particular, proposing a formalization of the process of quantitative model building, with a main idea of partitioning the formulation process into an abstraction and relaxation part.

Motivated by these ideas, we take the following as a premise for the work of this paper; that the decisions that are made in formulating a model can be categorized in a hierarchical manner, particularly that some modeling decisions can be assessed conditional on upstream modeling decisions. Figure 1 illustrates this hierarchy. The subsequent ability to assess a modeling decision in (conditional) separation from other modeling decisions will allow us to apply the information theoretic concepts and associated modeling principles, next discussed, in a more systematic way for certain modeling decisions, including the choosing of model resolution.

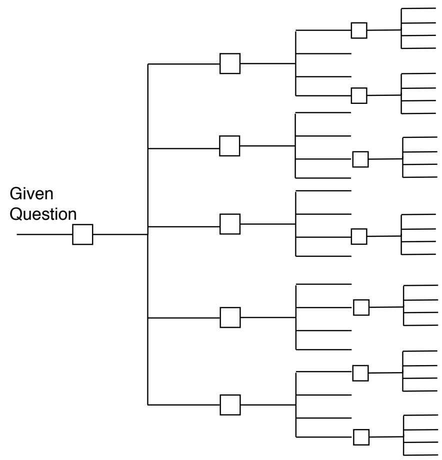

Figure 1: Schematic diagram illustrating hierarchy of modeling decisions required to answer a given question.

3 The information theoretic link

In this section, and the following, we introduce commonly expressed 'folk' goals of appropriate model formulation, typically founded on intuition built up by modeler experience, along with a set of information theoretic concepts, and show how these goals can be derived from, and grounded in, the information theoretic concepts.

3.1 Goals of model formulation

We here posit two goals of model formulation consistent with intuition and commonly expressed in best practice.[8] In addressing a question, we would like our model to simultaneously align with the following, often conflicting, goals:

- Maximize accuracy of information produced
- Minimize model detail

The first goal should be self-explanatory, while the second goal is related to the commonly applied Occam's Razor, and is similar, for example, to the 'principle of parsimony' expressed by Filar (2002), and from a practical standpoint has computational advantages. Neither goal alone necessarily makes sense in the absence of the other, and combining these goals allows us to see the tradeoff in formulating a model, in particular choosing a level of detail, to address a given question. How to combine these goals will be discussed in the next section.

Following these goals can be interpreted as the search for a model that maximizes relevant information produced. Importantly, this search may be shown to not only derive from intuitive

[8] For an example of an expression of best practice in the normative economics setting, see Gabaix and Laibson (2008).

principles, but also represent an interpretation of an established information theoretic approach to the finding of relevant information in a signal. This information theoretic approach of Tishby et al. (1999) is next outlined.

3.2 Into the information bottleneck

Tishby et al. (1999), in the context of signal processing, defines the relevant information of a signal $z \in Z$ as the information, in an information theoretic sense, that it provides about another signal $y \in Y$. Furthermore, the problem is set up to find a compressed representation of Z, $\bar{Z}$, that contains as much information as possible about Y. This 'information bottleneck' that finds the most relevant information in a compressed form, removing superfluous information, is a potential explanation for the empirical good performance of deep learning models in generalizing beyond training data (Tishby and Zaslavsky, 2015; Shamir et al., 2010). Additionally, Achille and Soatto (2018) show further desirable properties of the bottleneck approach and note how it seeks to 'get the most out of the data we have for a specific task' which is clearly desirable for our task of building normative models.

The Tishby et al. (1999) framework commences with the following optimization problem, where the mapping between Z and the compressed $\bar{Z}$, $p(\bar{z}|z)$ is chosen such that the relevant information in Z about Y is maximized in $\bar{Z}$, subject to the constraint that $\bar{Z}$ matches Z to some constant level of accuracy:

$$\text{maximize}_{p(\bar{z}|z)} I(\bar{Z}; Y)$$

such that:

$$I(\bar{Z}; Z) = \kappa$$

Where κ is some constant, $p(\bar{z}, z)$ is the mapping between $\bar{z}$ and z, and $I(.,.)$ represents the information theoretic concept of mutual information.[9] The mutual information between a pair of random variables is zero when they are completely independent, increases the more similar they are, and is bounded at a non-negative value when they match.[10] Choosing a mapping between Z and $\bar{Z}$ that maximizes the Lagrangian of the above leads to the following expression:

$$\text{maximize}_{p(\bar{z}|z)} I(\bar{Z}; Y) - \beta I(\bar{Z}; Z) \tag{2}$$

Where $\beta \geq 0$ is a scalar Lagrange multiplier, that reflects the stringency of the constraint on $I(\bar{Z}; Z)$, and as such can be a user defined parameter in the use of expression (2).[11] In the information theory context this can be interpreted as the choosing of a mapping between $\bar{Z}$ and Z such that the relevant information, expressed by the mutual information between the compressed $\bar{Z}$ and the true signal Y is maximized, less a penalty on the detail of $\bar{Z}$ relative to Z.

[9] The problem can also be expressed as the inverse, where $I(\bar{Z}, Z)$ is minimized subject to a constraint requiring a constant level of relevant information, $I(\bar{Z}, Y)$. The constraint makes sense, as if we truly only wanted to minimize detail without the constraint then we would choose the simplest possible mapping, a mapping that likely would not provide very much useful information.

[10] For an excellent and accessible overview of Information Theory concepts, see Luenberger (2006).

[11] Further, for the Lagrangian to hold, there are numerous conditions on Z, $\bar{Z}$, and Y, as outlined by Tishby et al. (1999).

Tishby et al. (1999) and Shamir et al. (2010) show links between the information bottleneck formulation and various information theory and statistics problems. In particular, Tishby et al. (1999) show that there exists a unique formal, computable solution for this problem to find the mapping between Z and $\bar{Z}$, $p(\bar{z}|z)$, under the condition of the Markov property holding between $\bar{Z}$, Z, and Y. Direct application requires knowing, as an input, the joint distribution of Y and Z, $p(y,z)$, which is frequently challenging in our setting, as will be discussed.

At this point, we can note the link between (2) and the model formulation goals introduced previously. We can map our first modeling goal of relevant information to $I(\bar{Z}, Y)$, and our second goal regarding model parsimony to $I(\bar{Z}, Z)$. This allows us to combine our goals as follows:

$$\text{maximize}\left(\texttt{Information} - \beta(\texttt{Model detail})\right) \quad (3)$$

We will develop this link in the context of normative model formulation further in the next section.

4 Application to normative modeling context

4.1 Mapping concepts to normative model formulation

To map these information theoretic ideas to our normative modeling context, we can map Y, the signal we are trying to learn about, to our question Q, Z, to our model f, and the compressed signal, $\bar{Z}$ to a compressed model f_c.

As stated in Section 2.1 above, the question Q may be represented as a random variable. For example, if the question is 'the cost of energy system decarbonization', $p(y)$ in this case could conceptually encode the probability distribution of a model producing a 'true' value. Z can represent an encoding of a model f,[12] while the joint probability distribution, $p(y,z)$, encodes the relative importance of different aspects of the model in producing the 'true' value. The search for $p(\bar{z}|z)$ then is a search for the mapping between a model and a more aggregate version, f_c or $\bar{Z}$, that maximizes the relevant information contained in the model about what we care about, our question Q. The challenge of not knowing absolutely the true signal frequently will drive much of the following discussion.

It follows from the Information Bottleneck that $I(\bar{Z}, Y) \leq I(Z, Y)$, which using our mapping would imply that a compressed model, f_c, can produce the same, or less, information about question Q as its parent model f, but not more, and this concept will be drawn upon later in the paper. From some perspectives, this relationship may seem somewhat counter to results about overfitting and generalization, where simpler models are shown to be more accurate, as more detailed models are less likely to generalize beyond the examples they are trained on.[13] There is a subtle and important difference here to note - we are not stating that a simpler model cannot be more appropriate than a more complicated model, but that a compressed model cannot say more about a

[12]In certain applications, Z could represent not the whole model necessarily, but one aspect of the input data, for example, it could encode resolution along the temporal dimension. We note that there are multiple ways of thinking about the mapping between Y and questions asked of a model also. More rigorous exploration may be useful as part of future research.

[13]For example, see Gigerenzer and Brighton (2009).

question than its parent model, noting they are built with the same data. Indeed, these overfitting concerns, along with a desire for computational tractability and the search for insights, motivate the previously introduced model formulation goal of minimizing detail, and is at the heart of the tension between the two goals espoused in the combined goal. This sets the stage for us to discuss model appropriateness in the normative model context.

4.2 Deriving model appropriateness

Given our discussions to date on model formulation goals and the underlying link to information theoretic ideas, we will define the appropriateness of a model for a particular question as the relevant information it produces (if any), which in turn is a function of the accuracy of information produced and model detail, in particular an evaluation of the objective function (3) for a particular model f and question Q:

$$\begin{aligned} g(f,Q) &= \texttt{Appropriateness[f,Q]} \\ &= \texttt{Information[f,Q]} - \beta(\texttt{Model detail[f]}) \end{aligned} \tag{4}$$

Where g is the conceptual evaluation of model appropriateness, with the more appropriate a model for a question, the higher the associated g value. For a given question and β, ideally the model chosen will maximize this function, finding the most relevant information consistent with the combined goal expressed in (3). We introduce this expression in this verbal form to emphasise the conceptual nature of this idea in our normative modeling context, and the general challenge of implementing it numerically. To recall, the challenge is in evaluating in absolute terms the true accuracy of any model in the normative setting. A more succinct notational form of (4) for convenience as we proceed is:

$$g(f,Q) = \tilde{I}(f,Q) - \beta\tilde{I}(f, f_{\texttt{datum}}) \tag{5}$$

Where $\tilde{I}(\cdot)$ is a concept inspired by, and analogous to, the concept of mutual information.[14] The first term evaluates the accuracy of a model by assessing how much accurate information the model f provides about the question Q. In the second term, the detail of a model is defined by the information it shares with a model of some datum level of greater detail.[15] Note that while the relationship between the terms is linear, the $\tilde{I}$ terms themselves are not necessarily linear. Except for boundary cases, we will never numerically evaluate this $\tilde{I}$ term in an absolute sense due to its conceptual nature, but we will be able to rank it under certain conditions.

Figure 2, inspired by Tishby and Zaslavsky (2015), illustrates the concepts further. For a given β, the relative weight of model accuracy to model detail, we have a conceptual limit of appropriateness any model representation can achieve. For β from ∞ to 0, we can trace a curve representing an 'appropriateness frontier', with β the slope at each point of the curve. We can see in this curve our goals of model formulation. That is, for a given level of model detail, a fixed point on the x-axis, we naturally want to maximize the useful information produced, along the y-axis.

[14] It is possible to set the system up such that ˜ (tilde) may be removed, and we discuss mutual information directly, however, we will use the tilde to highlight the conceptual nature at this point due to the challenge of enumerating $I(f,Q)$ in many cases. In further research to enable numerical applications in certain cases, the concept of mutual information could be directly applied.

[15] Again, the mapping between this expression and the previously introduced 'Information Bottleneck' is that the model f_{datum} represents the signal Z, f represents the compressed signal, $\bar{Z}$, and the representation of the real world implicit in Q represents the true signal we are interested in, Y.

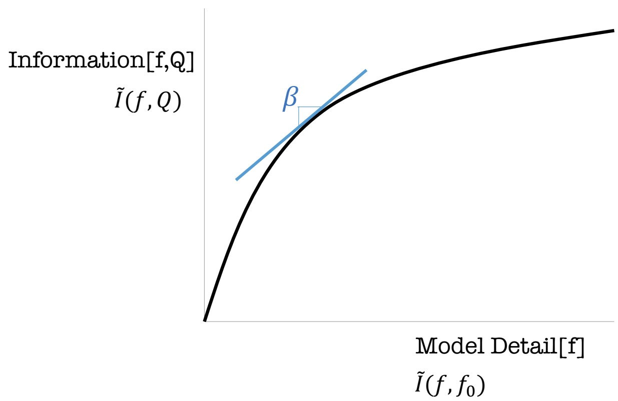

Figure 2: Appropriateness frontier. β defines the slope of curve that constitutes the frontier and represents the relative value of model detail to accuracy of model representation.

Table 1: Nomenclature

f	Model object
f_C	Compressed/aggregated model object
f_0	Datum model object
$\boldsymbol{x}^*$	Output of model at solution, $\boldsymbol{x}^* \in \Re^n$
Q	Question asked of model.
$g(f,Q)$	Appropriateness of model f for question Q
$\tilde{I}(a,b)$	Conceptual shared information between a and b
β	Relative importance of model detail to accuracy of model representation, $\beta \in \Re^1_+$

Similarly, for a given level of information on the y-axis, we want to minimize model detail along the x-axis. Generalization concerns, particularly the desire to reduce spurious noise, prevent us from always setting $\beta = 0$ and maximizing along the y-axis alone.[16]

At this point, we have introduced an abstract set of concepts where the appropriateness of the alignment between model and question is conceptually evaluated based on the relevant information the model provides, incorporating a penalty on the level of detail of the model. Our definition of appropriateness of a model, and the associated abstract g function, is tied to the model's purpose, consistent with the intuitive and commonly expressed 'horses for courses' notion that links model appropriateness to the question asked of a model. It is important to re-emphasize that we have developed this g function, not from intuition alone, but by mapping from well established information theory concepts. Table 1 overviews the nomenclature introduced so far.

[16]Furthermore, Tishby and Zaslavsky (2015) show how, with a finite sample of data, as is the case in our application, the true limit curves downward at some point as we move from left to right, conditionally allowing for an optimal β to be found at its peak.

5 Choosing the resolution of normative models

Due to the difficulty of unambiguously evaluating the accuracy of information a model provides, we cannot directly apply the concepts introduced so far to every modeling decision. We may, however, separate out a subset of modeling decisions where we can, at least partially, apply them due to the hierarchical structure identified in Figure 1. Our modeling decision on choosing resolution comprises one such subset of the decisions taken in formulating a normative model, and can be considered downstream of, and conditional upon, higher level modeling decisions relating to the question asked, modeling paradigm and associated model structure. In particular, for the case of choosing model resolution, we work around the shortcomings in our absolute knowledge by taking a relative perspective on how we assess formulations.

This section will deduce a number of cases where relative tests can be applied. The common trend through these cases is that the first term in the appropriateness function,[17] representing the accuracy of information the model provides, is fixed in either an absolute or relative sense. Assessment of the $g(f,Q)$ function is then possible as the 'model detail' term may be assessed analytically or numerically under certain conditions. We will also be able to reason somewhat about appropriateness when the first term is not fixed.

5.1 A relative test

To compare two models, one a compressed version of the other, assume model f_c is a compressed version of model f_0.[18] Applying the definition of appropriateness $g(f,Q)$ (4) to f_0 and f_c, we have:

$$g(f_0,Q) = (\texttt{Information}(f_0,Q) - \beta(\texttt{Model detail}(f_0)))$$
$$g(f_c,Q) = (\texttt{Information}(f_c,Q) - \beta(\texttt{Model detail}(f_c)))$$

Expressing these relationships in the notation of expression (5), and using f_0 as our datum of model detail, the binary choice of most appropriate model between f_0 and f_c is the model that returns the higher g value as follows:

$$\arg\max_{f\in\{f_0,f_c\}} g(f,Q) = \arg\max_{f\in\{f_0,f_c\}} \tilde{I}(f,Q) - \beta\tilde{I}(f,f_0)$$

We will now consider this choice under two cases, one where the first term is equal within a range of tolerance across the two models, and one where it is materially different.

5.1.1 When a compressed model returns same information

If the outputs from both the original model and the compressed model are the same for a given question, we can state that the first term of the respective $g(f,Q)$ functions are equal across both models, i.e. they are both as accurate or inaccurate as each other for the question Q. This then

[17] Statement (4); $g(f,Q) = \texttt{Information[f,Q]} - \beta(\texttt{Model detail[f]})$

[18] For the purposes of this section, we will treat aggregation and compression as interchangeable, and that the arguments here apply to any allowable compression process.

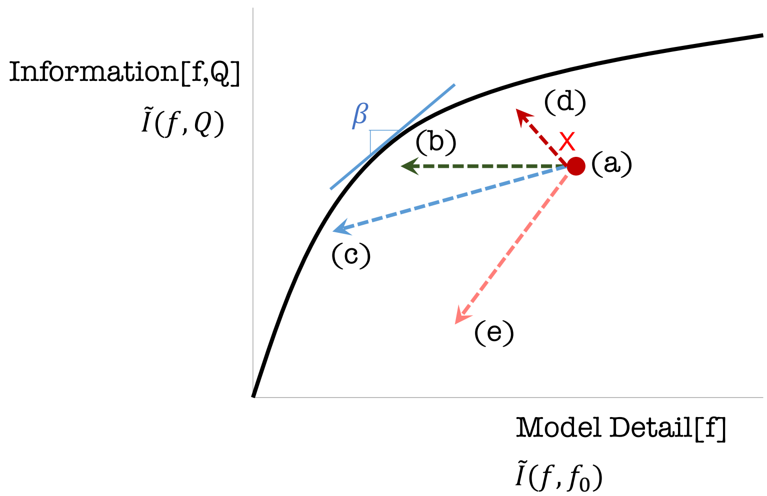

Figure 3: Assessing if a reduction in detail moves a model relatively closer to the appropriateness frontier. $(a) \rightarrow (b)$ is an easy-to-verify improvement by asking does compression change information returned by the model. If the returned information changes, it is necessary to judge if the path is the improving $(a) \rightarrow (c)$ or the unwanted $(a) \rightarrow (e)$. The path $(a) \rightarrow (d)$, meanwhile, is impossible by the definition of compression in the setup of our problem.

allows us to compare the models on the basis of model detail. As one model is a compressed version of the other, this comparison is unambiguous. Using our notation, we can state this logic as follows:

$$d_Q(\boldsymbol{x}_c^*, \boldsymbol{x}_0^*) \leq \epsilon_Q \implies \tilde{I}(f_0, Q) = \tilde{I}(f_c, Q) \implies g(f_c, Q) > g(f_0, Q) \tag{6}$$

Where ϵ_Q, a scalar, is the threshold of equivalence associated with the question Q, $d_Q()$ is a distance metric that compares outputs associated with the question Q, for example Euclidean distance, and $\beta > 0$. The final step above follows from $\tilde{I}(f_c, f_0) < \tilde{I}(f_0, f_0)$, by definition of compression.

This expression (6) states that if we have a pair of models where one can be shown to be a compressed version of the other, and they produce the same output within a set range of tolerance for a given question, we can logically state, following from our initial modeling goals, that the compressed model is more appropriate for that question.

This comparison is graphically presented in Figure 3, comparing points (a) and (b), where the information returned at (b) is the same as that returned at (a). Indeed, rather than a simple pairwise comparison between a model and a compressed version of that model, we may search for the level of compression that minimizes the second term of the g appropriateness function, while producing the same output as the non-compressed model for the question asked. That is, moving from right to left across the horizontal line until the frontier is reached. This indicates a path to the appropriateness frontier with a conceptual guarantee - by specifying a model, and then compressing it as much as possible. It is important to note of course that the frontier is conditional on the upstream modeling decisions, which are difficult to judge in an absolute sense.

5.1.2 When a compressed model returns different information

In the case of comparing a model with a compressed version of itself, what can we say about appropriateness if model outputs associated with a question are materially different? There are then conditions where we say the compressed model is either more or less appropriate. Recall that, by definition, when a compressed model f_c produces different output for a question than its parent model f_0, then we can say that it produces strictly less accurate information about that question Q, $\tilde{I}(f_c, Q) < \tilde{I}(f_0, Q)$, within the constraints of the finite observations upon which the models are built. This stems from our discussion of compression in Section 4.1. Thus, referring to Figure 3, the move $(a) \rightarrow (d)$ is infeasible. The issue is then whether the benefit from reduction in detail is worth the loss of relevant information about the question Q.

For the compressed model to be *less* appropriate in this circumstance, i.e. have a lower g value, $g(f_c, Q) < g(f_0, Q)$ when $d_Q(\boldsymbol{x}_a^*, \boldsymbol{x}_0^*) > \epsilon_Q$, the following condition must hold:

$$\overbrace{\tilde{I}(f_0, Q) - \tilde{I}(f_c, Q)}^{\texttt{Reduction in accuracy}} \geq \beta \overbrace{(\tilde{I}(f_0, f_0) - \tilde{I}(f_c, f_0))}^{\texttt{Reduction in detail}} \tag{7}$$

That is, for the compressed model to be less appropriate, the loss in accuracy due to compression is greater than the benefits of reducing model detail. This is graphically expressed as $(a) \rightarrow (e)$ in Figure 3. An alternate path, where the benefits of reducing model detail are greater than the loss in accuracy is presented as path $(a) \rightarrow (c)$ in Figure 3.

We cannot generally evaluate this condition due to the aforementioned challenges in absolute evaluation of the accuracy and relevance of information provided by a model in our normative, and frequently long time horizon, context. Finding a way to do so, or at least estimate this condition given available data, would open up the problem of model formulation to a whole host of recent advances across a number of fields.

Challenges notwithstanding, our discussion so far allows us to make some informed observations about whether model compression that changes model output puts us on path $(a) \rightarrow (c)$ of Figure 3 or undesirable path $(a) \rightarrow (e)$. As an instructive case, we may observe that $\beta = 0$ allows us to evaluate statement (7). If $\beta = 0$, then the information about the question Q matters most, no matter the benefit in detail and we can say the compressed version is more appropriate if *and only if* it produces the same model output as its parent model. Equivalently, if the queried model output is materially different, then the compressed model is less appropriate. On the other hand, if we consider a high β value, where the benefits of reducing model detail are highly weighted, we may note graphically from Figure 3, that this high β value is associated with any acceptable $(a) \rightarrow (c)$ path. The consideration we then have when we compress a model and change the output it returns; are we squeezing out noise, or losing relevant detail?

For many practical applications, the reality is that somewhere between a very low or very high β will be desirable. For certain applications however, a low β may be plausible, thus ruling out any allowable material deviation due to compression:

- If the non-compressed model in this pairwise comparison is built upon small data availability and is perhaps already simplified, then the incentive to reduce detail, and associated benefits, is dampened.
- How the test for equivalency is applied can justify a low or zero β. The threshold of equiva-

lence, ϵ_Q, could be designed such that material and significant changes in model output would indicate that deviation is undesirable.

- We could apply additional assumptions on top of our system to extend these cases. For example, if we have some good reason that our non-compressed model is accurate, then generalizability is less of a concern. Certain physical or technical constraints could constitute such a good reason.[19] In such a case, we could justify a low value of β, that is a compressed model should not remove details that we know matter, and accordingly should not distort model output.

While the concepts introduced here do not give a direct answer to every case, they ideally provide value by clarifying the tradeoff through building on reasonable model formulation goals while drawing on established information theoretic ideas, and thus may augment human common sense.

5.1.3 What should be the starting point of comparison?

In this discussion we have largely focused on the properties of the compressed model, represented by points $(c), (d), (e)$ on Figure 3, relative to their 'parent' model. What should this model, represented by point (a) be, and how to choose it? An intuitive approach would be to write down the model that appears conceptually to be the best fit for the question Q, independent of computational concerns.

Compressing from this point has several indicators in its favour. First of all, it fits with the graphical picture of Figure 3, where it makes sense to try and pick the (a) starting point as far along the y-axis as possible, before compression moves us to the left, and does not allow any further movement further along the y-axis in the desirable direction. It aligns with the aforementioned Schneeweiss (1987) formalization of the process of quantitative model building - where the choosing of (a) corresponds to the abstraction phase, and the subsequent compression corresponds to the relaxation phase. Furthermore, it corresponds to Achille and Soatto (2018) discussion on training a deep neural network with good generalization properties, particularly the presence of a fitting phase and subsequent compression phase in the learning of weights. While some of these ideas can be tested more precisely in the context of, say, training a neural network, in the normative modeling context, and the choosing of model resolution within that context, they at the least provide useful analogies, while potentially providing deeper justification for the approach to choosing of resolution implied by our discussion.

Additionally, note that these indicators are one directional, in that they support comparing a model with its compressed form, but not the other way around, comparing a model with a more detailed version of itself. Furthermore, this alternate approach would introduce increased parameterization requirements, with the associated possibility of introducing errors.

[19] For example a model of electricity supply needs to recognize the economic importance of matching supply and demand at each instant.

5.2 Practicalities

This section will attempt to bring the concepts introduced in the previous sections closer to being actionable, and useful in the design and evaluation of models. In so doing, this section explores how to implement the test of relevancy of model detail numerically and analytically, and how to test for global relevance locally. Throughout this section, an underlying objective is to address the following key question regarding the ideas that have been introduced previously: is the only way to test the appropriateness of a less detailed model, relative to a more detailed model, to build both?

5.2.1 Numerical testing of detail relevancy

In our discussion so far, the test on whether compressing a model changes the model output is key. The coarsest and direct implementation of this test is to simply compare the outputs of a model with the compressed version of itself and compare outputs. This approach involves gathering the required data, before building and solving both models.

For multiple questions we would have to repeat the test for each question. Similarly, for a variety of model parameterizations, or even models, we would have to apply the test repeatedly.

Clearly this numerical testing is computationally challenging, but with ongoing improvements in computing power and in software stacks to manage it, along with increasing data availability, such numerical testing of models is becoming ever more possible, and will continue to do so. Additionally this numerical testing does offer a principled way of harnessing these new computational opportunities.

Through systematic running of such tests we can develop ‘operating conditions’. These conditions would establish a range of parameterization and questions for which a compressed model structure is appropriate by our definitions, which, in practice, frequently means that the compression does not materially change model outputs. As a corollary, we note that though the full pre-compression model may have to be solved for testing, the computational burden of solving it is not required for day-to-day operation of the model.

To design compressed models that maintain their appropriateness across multiple parameterizations and across multiple questions, and do not require the running of such tests every time, sometimes an infeasible proposition in practice, we need a more general result based on an analytic guarantee. The analytic approach, which may be a substitute or complement to the numeric approach, is discussed in the following section.

5.2.2 Analytical assessment

There are cases where differences in output by a lower resolution model relative to a higher resolution model can be determined analytically by considering the mathematical structure of the problem. In such cases, it may be analytically proven that more detail along a particular dimension will not change a model’s output to a question of interest. Computational tests, as described in the previous section, could then be foregone. This analytical approach is also useful to provide guarantees across all possible questions, or a defined subset, as opposed to only those questions for which the computational tests have been run.

In an example from the trade policy literature, Arkolakis et al. (2012) show that under certain conditions for a certain class of models, additional micro-level model features do not provide any further information in the model outputs on the macro question of what welfare gains are associated with trade. This approach can be considered a part of the broader 'sufficient statistics' approach discussed in Chetty (2009), where analytical assessment of a problem guides how much detail is necessary to address a question of interest, before the computational work is undertaken. Interestingly, there is a mathematical connection between the information bottleneck idea and minimal sufficient statistics.[20]

In the case of model aggregation, an analytical approach may take advantage of theoretical results bounding the objective function value of the unsolved disaggregated problem given a specified aggregation scheme. As already mentioned, Zipkin (1980a,b) develop a priori and a posteriori bounds on how far the objective value of the aggregated model is from the objective value of the original model. Importantly, calculation of the bounds do not require the solution of the original model. This paper overlays a broader conceptual framework around such bounds, providing a context for the appropriateness of any errors introduced by an aggregation scheme.

As previously noted, clustering algorithms are helpful for model aggregation through the aggregation of similar entities. There are a variety of options for implementation of clustering algorithms, with numerous approaches applied in the literature. While this paper draws upon the information bottleneck ideas to the broader conceptual formulation problem, it can be applied to clustering itself, guiding associated implementation choices, as discussed by Slonim and Tishby (2000).

5.2.3 Testing for global relevance locally

Another path to reduce the burden of computational tests to assess the relevance of model detail is to harness any hierarchical structure in problem formulation. When there is hierarchical structure in a problem formulation, we can decompose the problem into subcomponents. This can allow us to test for the global relevance of a decomposed subcomponent locally.[21] Referring to Figure 4, our tests for the relevance of model detail could be applied locally to either or both the subcomponent model or the master model, but not necessarily to the integrated model, reducing the computational and/or analytical burden. In conducting the tests, we will want to test the sensitivity of $x|y$ and $y|z$. For example, if the sensitivity of x to y is low, we are able to justify a coarser representation of the subcomponent model, allowing a greater threshold of equivalency when assessing model outputs from compressed versions of the subcomponent model.

This intuitive method of testing the input/output performance of a subcomponent model has been seen in other fields,[22] while assessing the required detail of subcomponents of integrated models is a relatively common modeling task. For example, electricity capacity planning models, with their associated resolution issues, are frequently subcomponents of larger integrated assessment models.

[20] For further discussion, see Shamir et al. (2010).

[21] An early example of discussion of model decomposition in an energy-economic model context is Weyant (1978). The abstract example here is similar to the Hogan and Manne (1977) metaphor of how much a rabbit changes the taste of an elephant stew. Merrick et al. (2018) discuss decomposing electricity planning problems and solving them with a decentralized ADMM algorithm.

[22] For example, Zeigler (1984) presents a system similar to that presented in Figure 4 for designing simulation models.

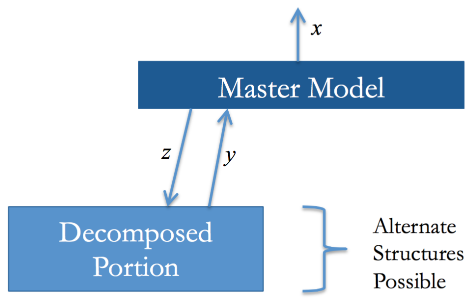

Figure 4: Decomposition allows checking of local sensitivity of $x|y$ and $y|z$

6 Illustrative applications

The problem of choosing resolution relates to a large number of modeling decisions in the normative modeling context. This section illustrates a sample of these relations. In addition, the appendix contains a toy example relating to model aggregation.

6.1 Clock toy example

We now will use the analogy of a clock to further illustrate the concepts, in particular to address some potential concerns. This illustration is expressed in the form of two thought experiments, one where a clock is stopped, and one where it is functioning but is set to the wrong time.

6.1.1 The stopped clock

Imagine a broken clock, with the hands stuck at a certain time. It will tell the correct time twice per day. If this clock represents a model, what do our tests say about it? First of all, we can consider the broken clock a compressed version of the fully functioning clock. For certain questions, for example asking what time it is at the instant the broken clock happens to be stuck at, the test would say the broken clock was more appropriate. For other questions, our test would show that this compressed clock was inappropriate,[23] as it would produce different output than the fully functioning clock. This example shows the role of asking the right question. Our absurd outcome where our test showed the broken clock to be more appropriate stemmed from the narrow question asked, and is avoided for more relevant questions for determining how well a clock is functioning. Such questions could ask if the clock is correct for each instance, or what are the dynamics by which the time changes. Choosing the question remains in the domain of wisdom.

6.1.2 The functioning clock that is fast

Consider, as our next malfunctioning clock, a clock that is fully functioning except that it is running one hour fast. This case highlights another potential concern with our tests. Because this clock is producing different output than that produced by the properly calibrated clock, will it be deemed inappropriate even though a simple adjustment in interpretation of the clock's outputs will produce

[23] Assuming the gain in simplicity was not worth the loss of accuracy.

the correct time? An answer to this concern is that our fast clock does not meet our condition of being a compressed model, as a relaxation or simplification does not move us from the original clock to our fast clock. In this sense, our core test cannot be applied, however higher level judgment and analysis could identify the relationship between the clocks.

Additionally, the discussion of this case assumed we knew the background correct time as we assessed the clock. In the broader set of models to which this paper relates, we do not always know the true and correct model, so we are reduced, at least in direct application of our tests, to assuming a model and assessing compressed models relative to that model.

6.2 Temporal and spatial resolution

As outlined in the introduction, temporal and spatial resolution is an important consideration in the design of normative models. We here briefly discuss each in the context of the ideas presented in this paper. As shown in our goals of model formulation, and supported by our information theoretic arguments, the principle that will guide us is that for a given range of model output, we want to find the most compressed representation that delivers this output. Similarly, for a given level of model detail, we want to find the representation that maximizes relevant information. That is, we would like to choose our resolution such that we are on the 'appropriateness frontier' of Figure 2.

6.2.1 Temporal resolution

The concepts in this paper underlie the approach taken by Merrick (2016) to consider the representation of temporal resolution in models of optimum electric sector investment. Merrick (2016) showed that, with the introduction of variable renewables, an increase of two orders of magnitude, to the order of 1000 representative time periods of the hourly 8760, from the order of 10, was required to guarantee that no error was introduced by a clustered aggregation for any model question.

Choosing a resolution using such an approach is implicitly a demonstration of the attempt to find a model that approaches the 'appropriateness frontier' through compression that does not change model output. Merrick (2016) also highlighted how a subset of hours of the resolution may be more influential than others, and how prior information about these hours may be harnessed to find an appropriate greater level of aggregation.[24] An interesting topic for further research is to conduct the aggregation directly using the clustering algorithm that can be derived from Tishby et al. (1999), with the challenge to reasonably codify how different aspects of the non-aggregated resolution affect model outputs.

6.2.2 Spatial resolution

As a spatial resolution example, we will consider the Dynamic Interactive Vulnerability Assessment (Diva) database (Vafeidis et al., 2008; Hinkel and Klein, 2009), a primary input to numerous studies relating to sea level rise, including the Coastal Impact and Adaptation Model (Ciam) of Diaz (2015, 2016). Figure 5 presents an assessment of the geographic similarity of coastal regions in the Diva database, indicating a high degree of similarity across regions. For questions that relate

[24]For more on an implementation of the idea in this specific context, see Blanford et al. (2018).

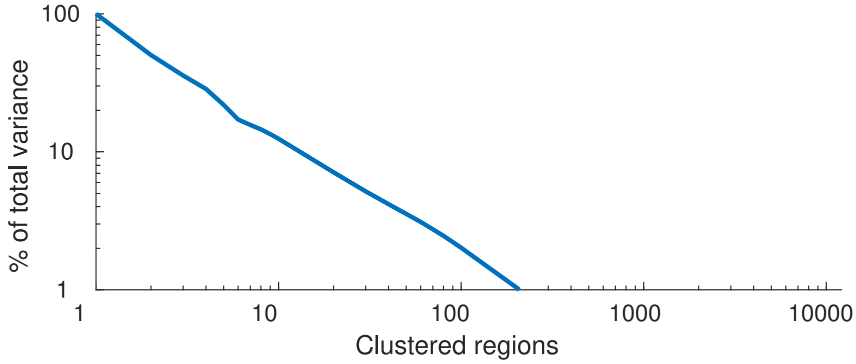

Figure 5: Assessment of similarity of regions in DIVA dataset. Zero on the y-axis indicates regions are duplicates of each other.

to the underlying parameters (as opposed to questions about regional differences not present in this dataset), Figure 5 indicates that aggregation from 12,148 to 1,000 regions will not distort model outputs due to the possibility of grouping extremely similar regions together. For other questions, there are additional data to include such as the value of capital stock in each region, requiring the assessment to be undertaken again with the addition of the new data.

The broad point here is that aggregation of similar entities intuitively allow reduction in model detail with minimal effect on model outputs. If we wish to reduce resolution further, the introduced framework for model appropriateness demonstrates the tradeoffs involved, while the underlying ideas encourage squeezing the most relevant information 'through the bottleneck' when aggregating. The value provided by such assessments in choosing model resolution will either be the ability to reduce the detail and associated computational burden of the model, or a firm justification for not doing so.

Another spatial resolution example is a comparison of the global DICE integrated assessment model with its sister regional RICE model (Nordhaus and Boyer, 2000). For regional questions, RICE can be shown to be more appropriate, as DICE does not return information on sub-global questions, and thus would receive a negligible appropriateness ranking for such questions. For global questions, the test of relevance of model detail allows the assessment of whether the global outputs differ under aggregated global resolution compared to under multi-region resolution. The work of Schumacher (2015) may be considered an application of this test, by providing analytic conditions for when selected global outputs are changed by aggregation from regional to global spatial resolution, and when they are not. More broadly, DICE is an excellent / canonical example of the compression of model detail in order to address a well chosen question. Many, many details about humankind's activities are compressed into a simplified representation, and coupled with a model of core climate system behavior, in order to understand, at a high level, optimal global responses to the challenges posed by climate change.

6.3 Treatment of uncertainty

Uncertainty is prevalent when modeling systems at the scales considered in this paper. Indeed, the treatment of uncertainty is comprehensively linked to our goals of model formulation, particularly relating to model detail. The intuitive desire to limit model detail when necessary can be shown to be linked to the generalizability of information produced by a model, and its use in an uncertain world. There are also various methods of explicitly treating uncertainty in model formulation, and this section will briefly discuss how the concepts introduced in this paper could be extended to guide their design.

Scenario analysis is a way of aiding decision making under uncertainty (Schwartz, 1991). Assessing a range of scenarios is a common approach in co-ordinated model inter-comparison projects (for example, see the EMF27 scenarios in Luderer et al. (2014)). In choosing a scenario from the universe of possible scenarios, it is desirable to choose the scenarios most relevant to the decision at hand. To do so, we could apply the concepts of choosing model resolution introduced in this paper to a more meta level, by squeezing out the scenarios that are most relevant while incentivising the choice of as few as possible. Like before, direct implementation is challenging, as it requires a priori mapping between the scenario space and the most useful information in outputs. However, conceptually the approach outlined in this paper may be useful to move scenario design beyond seemingly ad hoc choices, in addition to providing a base for future research into systematic choices of scenarios with perhaps formal learning approaches taking the place of the unknown a priori distributions. Similarly, the approach can be applied to managing the curse of dimensionality in stochastic optimization by helping to prune the space included in the optimization. Again, we wish to consider only the most relevant uncertainties.

Like temporal and spatial resolution, treatment of uncertainty is a core modeling decision that trades off a potential increase in the desired relevant information with an increase in the unwanted extra detail. This paper attempts to provide a framework for augmenting human expertise when making such a modeling decision.

7 Roadmap

At this point, we will attempt to distill our discussion into key points, that will ideally provide a 'roadmap' for the modeling community. Additionally, we lay out these points as tenets that may be tested rigorously by future research.

(a) Define the question to be addressed and implicit assumptions to be made in addressing it. The art of asking the right question, with a model appropriate for addressing it, to help decision makers seeking information and insights regarding the challenges they face, remains supreme.

(b) Conceptualize an 'ideal' model that aligns best with the question, particularly paying attention to relevant physical and economic structural dimensions of the problem. In reality, there will naturally be some degree of joint selection in choosing a question and model given what is feasible to model and what data is available.

(c) Compress the model resolution as much as possible without materially changing what is concluded about a question. If compression does change model outputs, be aware of the increased weight that places on model parsimony in assessing appropriateness, with associated loss in relevant information about the question. Compression can mean aggregation of resolution along a given dimension, or alternatively approximations like convex relaxation of constraints.

(d) Analytic assessment to check whether a less detailed model resolution materially affects model outputs can provide a general guarantee that, and is thus likely preferable to, numeric assessment. However, numeric testing can complement the analytic assessment, or as a substitute, be used to work out the appropriate operating conditions of the model in a relatively blunt, but potentially effective, fashion.

(e) Be careful to operate the model within the derived operating conditions that indicate where model/question alignment holds.

(f) Interpret the model with humility, providing appropriate caveats, remembering that the quest for why the results are the way they are is key in the normative modeling context. To gain an understanding of what aspects of the modeled system drive results is a worthy goal of modeling.

(g) Overall, we want to choose resolution so as to develop the lightest model that answers the most relevant question in a fashion that efficiently harnesses available data.

A contribution we provide in this paper is that this roadmap is not only based on empirically induced best practice, but has a grounding in, and can be related to, information theoretic concepts. Finally, machine learning and broader artificial intelligence methods will allow us to understand available data better, and find more efficient compressed representations, consistent with the steps above, to model the questions we are interested in. The broader question for the modeling community is, however, how these methods might, or might not, change the whole paradigm of normative modeling, a topic we leave for ongoing research.

8 To conclude

This work, through a novel approach, has attempted to inform the choosing of model detail in normative models with inherent validation difficulties. The starting concept is that there are modeling decisions that are conditionally separable, allowing us to isolate situations, where systematic evaluation is possible, from others where evaluation rests with judgment and expert option alone. We have shown how two familiar modeling goals can be based not only on wisdom but on information theoretic ideas, and develop how they may be applied to the choosing of model resolution. Furthermore, based upon this journey, we have developed a roadmap to aid modelers think about this topic, a topic that can have first order impact on model results

This work may be applied to modeling challenges relating to, for example, temporal and spatial resolution, treatment of uncertainty, technological details, and representation of heterogenous demographic groups. While associated modeling decisions can be made in an ad hoc fashion in practice, there are also more sophisticated approaches in the literature to choosing resolution in these cases, such as the use of clustering methods. This paper aims to provide principled guidance on when more sophisticated methods are necessary, and on their implementation. While models that include increased resolution will be able to say more about the disaggregated entities, this work is most valuable in thinking about whether the disaggregation, and the interaction between the disaggregated components, affect more aggregate model outputs of interest.

From a broader perspective, as computational power continues to develop through hardware and algorithm advances, and as availability of data continues its recent increasing trend, having a logical basis to address how much model detail is required to address certain questions may be increasingly useful. Data and computation will less frequently be the binding constraints of an analysis. Additionally, the derivation of simple rules from the abstract concepts introduced in this paper could be beneficial to those who 'consume' model outputs.

This paper is an attempt to open up a new frontier in model formulation, particularly the choosing of model resolution, largely following a discursive approach, and will ideally inspire further research in this domain. More rigorous results can follow, and also more direct numerical applications of the information theoretic ideas for choosing model resolution. Systematic methods to evaluate formulation of the models discussed in this paper are not common. A vision would be for model formulation, parameter estimation, scenario development, uncertainty characterization, and model diagnostics to be all considered comprehensively and systematically together during the whole modeling process. This vision may be conceptual at first, followed by more formal systems, all with a goal of increasing the utility of, and confidence in, normative models.

There is an analogy to computational tasks such as image recognition where programs were coded to identify specific features selected by experts. This paradigm has recently been displaced by programs that search the problem space themselves, learning the relevant features. The challenge of extending the analogy to the domain of the models discussed in this paper is the challenge of how to label the appropriateness of different formulations. For now, we focus on the more constrained, yet useful, problem of choosing model resolution.

Acknowledgments

The authors would like to thank Professor James Sweeney, Dr. Melanie Craxton, and anonymous reviewers, for feedback that has greatly improved this paper. This paper has also benefited from many conversations about modeling that the authors have enjoyed with collaborators on numerous applied projects. Primary funding for this project came from the U.S. Department of Energy, Office of Science PIAMDDI (DE-SC0005171) and MultiSector Dynamics (DE-SC0016162) grants to Stanford University.

A Aggregation toy example

Appropriate aggregation has played a central role in this paper. This toy example will walk through some of the concepts presented. Firstly, let us define model f as follows:

$$\max_{\boldsymbol{x}} 2x_1 + 3x_2 + x_3$$

subject to:

$$x_1 + x_2 + x_3 \leq 3, \quad x_1 - x_3 \geq 1, \quad \boldsymbol{x} \geq 0$$

The optimum value of the function maximized is 8, and the corresponding $\boldsymbol{x}$ values are $(1;2;0)$. Now assume model h is a compressed version of model f with variables x_1 and x_2 aggregated such that the coefficients associated with the aggregate variable are an average of the original coefficients. We will assign these aggregated entities the subscript $\bar{12}$. Model h can be represented thus:

$$\max_{\bar{\boldsymbol{x}}} 2.5x_{\bar{12}} + x_3$$

subject to:

$$x_{\bar{12}} + x_3 \leq 3, \quad 0.5x_{\bar{12}} - x_3 \geq 1, \quad \bar{\boldsymbol{x}} \geq 0$$

The optimum value associated with model h is 7.5, and the corresponding $\bar{\boldsymbol{x}}$ values are $(3;0)$. Models f and h help illustrate some of the concepts introduced about testable model aggregation. As h is a compressed version of f, to determine its relative appropriateness for different questions, we simply compare the outputs of the two models.

To apply the test we first have to define the question. For this first example, let our question be what is the value of x_3 at the optimal solution. Assuming L1 distance as our distance metric, d_Q compares x_3 across the two solved models:

$$d_Q(\boldsymbol{x}^*, \bar{\boldsymbol{x}}^*) = |x_3^{(f*)} - x_3^{(h*)}| = |0 - 0| = 0 \leq \epsilon_Q$$

We see that for this question, the compressed model, h, produced the same output as the original model f. We can then say that model h is more appropriate than model f for this question. The test is simply whether the aggregation that took place from model f to model h distorted the output of interest, x_3.

In contrast, if the question asks what is the objective function value returned by the model, the difference between model outputs can be expressed as follows:

$$d_Q(\boldsymbol{x}^*, \bar{\boldsymbol{x}}^*) = |8 - 7.5| = 0.5$$

The test is then, firstly, whether this level of deviation is material, depending on the choice of the ϵ_Q parameter, and if so, secondly, then depends on the tradeoff between model detail and model information discussed in Section 5. While for this toy example, the comparisons may seem obvious, the idea is to demonstrate a logical basis of to model evaluation, so that the reasoning can be extended to large-scale models or cases where the comparisons are not so obvious, improving transparency of modeling decisions.

Bibliography

Achille, A. and Soatto, S. (2018). Emergence of Invariance and Disentanglement in Deep Representations. *Journal of Machine Learning Research*, 19:1–34.

Arkolakis, C., Costinot, A., and Rodriguez-Clare, A. (2012). New Trade Models, Same Old Gains? *American Economic Review*, 102(1):94–130.

Bishop, R. (2017). Chaos. In Zalta, E. N., editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, spring 2017 edition.

Blanford, G. J., Merrick, J. H., Bistline, J. E., and Young, D. T. (2018). Simulating Annual Variation in Load, Wind, and Solar by Representative Hour Selection. *The Energy Journal*, 39(3):189–212.

Chetty, R. (2009). Sufficient Statistics for Welfare Analysis: A Bridge Between Structural and Reduced-Form Methods. *Annual Review of Economics*, 1(1):451–488.

DeCarolis, J., Daly, H., Dodds, P., Keppo, I., Li, F., McDowall, W., Pye, S., Strachan, N., Trutnevyte, E., Usher, W., Winning, M., Yeh, S., and Zeyringer, M. (2017). Formalizing best practice for energy system optimization modelling. *Applied Energy*, 194:184–198.

Denton, B. (2017). A rose by any other name ... *ORMS Today*, October:8–9.

Déry, R., Landry, M., and Banville, C. (1993). Revisiting the issue of model validation in OR: An epistemological view. *European Journal of Operational Research*, 66(2):168–183.

Diaz, D. B. (2015). *Modeling Uncertain Climate Impacts and Adaptation for the Integrated Assessment of Carbon Policy*. PhD thesis, Stanford University.

Diaz, D. B. (2016). Estimating global damages from sea level rise with the Coastal Impact and Adaptation Model (CIAM). *Climatic Change*, 137:143–156.

Filar, J. A. (2002). Mathematical Models. In *Knowledge for Sustainable Development, An Insight into the Encyclopedia of Life Support Systems*, chapter Vol. 1, pages 339–354. UNESCO/EOLSS, Johannesburg.

Gabaix, X. and Laibson, D. (2008). The Seven Properties of Good Models. In Caplin, A. and Schotter, A., editors, *The Foundations of Positive and Normative Economics*, pages 292–299. Oxford University Press, Oxford.

Geoffrion, A. (1976). The Purpose of Mathematical Programming is Insight, Not Numbers. *Interfaces*, 7(1):81–92.

Gigerenzer, G. and Brighton, H. (2009). Homo Heuristicus: Why Biased Minds Make Better Inferences. *Topics in Cognitive Science*, 1(1):107–143.

Hämäläinen, R. P., Luoma, J., and Saarinen, E. (2013). On the importance of behavioral operational research: The case of understanding and communicating about dynamic systems. *European Journal of Operational Research*, 228(3):623–634.

Hamilton, S. H., El Sawah, S., Guillaume, J. H., Jakeman, A. J., and Pierce, S. A. (2015). Integrated assessment and modelling: Overview and synthesis of salient dimensions. *Environmental Modelling & Software*, 64:215–229.

Herrera, A. O., Scolnik, H. D., Chichilnisky, G., Gallopin, G. C., Hardoy, J. E., Mosovich, D., Oteiza, E., de Romero Brest, G. L., Suárez, C. E., and Talavera, L. (1976). *Catastrophe or New Society? A Latin American World Model.* International Development Research Centre, Ottawa.

Hinkel, J. and Klein, R. J. T. (2009). Integrating knowledge to assess coastal vulnerability to sea-level rise: The development of the DIVA tool. *Global Environmental Change*, 19(3):384–395.

Hogan, W. and Manne, A. S. (1977). Energy-Economy Interactions: The Fable of the Elephant and the Rabbit. In Hitch, C. J., editor, *Modelling Energy-Economy Interactions: Five Approaches.* Resources for the Future, Washington DC.

Howard, R. A. (1968). The Foundations of Decision Analysis. *IEEE Transactions on Systems Science and Cybernetics*, 4(3):211–219.

Howard, R. A. and Abbas, A. E. (2015). *Foundations of Decision Analysis.* Prentice Hall, Upper Saddle River, N.J.

IPCC (2014). *Climate Change 2014: Mitigation of Climate Change. Contribution of Working Group III to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change.* Cambridge University Press, Cambridge, UK and New York, NY, USA.

Jakeman, A., Letcher, R., and Norton, J. (2006). Ten iterative steps in development and evaluation of environmental models. *Environmental Modelling & Software*, 21(5):602–614.

Kann, A. and Weyant, J. (2000). Approaches for Performing Uncertainty Analysis in Large-Scale Energy/Economic Policy Models. *Environmental Modeling & Assessment*, 5:29–46.

Kasparov, G. (2010). The Chess Master and the Computer. *The New York Review of Books.*

Landry, M. and Oral, M. (1993). In search of a valid view of model validation for operations research. *European Journal of Operational Research*, 66(2):161–167.

Lempert, R. J. (2003). *Shaping the Next One Hundred Years: New Methods for Quantitative, Long-Term Policy Analysis.* Rand Corporation, Santa Monica.

Luderer, G., Krey, V., Calvin, K., Merrick, J., Mima, S., Pietzcker, R., Van Vliet, J., and Wada, K. (2014). The Role of Renewable Energy in Climate Stabilization: Results from the EMF27 Scenarios. *Climatic Change*, 123(3-4):427–441.

Luenberger, D. G. (2006). *Information Science.* Princeton University Press, Princeton.

Merrick, J., Ye, Y., and Entriken, R. (2018). Assessing the System Value of Optimal Load Shifting. *IEEE Transactions on Smart Grid*, 9(6):5943–5952.

Merrick, J. H. (2016). On representation of temporal variability in electricity capacity planning models. *Energy Economics*, 59:261–274.

Morgan, M. G., Henrion, M., and Small, M. (1998). *Uncertainty: A Guide to Dealing with Uncertainty in Quantitative Risk and Policy Analysis*. Cambridge University Press, Cambridge [England], repr. [with corr.] edition.

Murphy, F. H. (2005). ASP, The Art and Science of Practice: Elements of a Theory of the Practice of Operations Research: A Framework. *Interfaces*, 35(2):154–163.

Nestler, S. (2011). Reproducible (Operations) Research. *ORMS Today*, 38(5).

Nordhaus, W. D. and Boyer, J. (2000). *Warming the World: Economic Models of Global Warming*. MIT Press, Cambridge MA.

Oreskes, N., Shrader-Frechette, K., and Belitz, K. (1994). Verification, Validation, and Confirmation of Numerical Models in the Earth Sciences. *Science*, 263(5147):641–646.

Pfenninger, S., Hawkes, A., and Keirstead, J. (2014). Energy systems modeling for twenty-first century energy challenges. *Renewable and Sustainable Energy Reviews*, 33:74–86.

Pidd, M. (1999). Just Modelling Through: A Rough Guide to Modeling. *Interfaces*, 29(2):118–132.

Rogers, D. F., Plante, R. D., Wong, R. T., and Evans, J. R. (1991). Aggregation and Disaggregation Techniques and Methodology in Optimization. *Operations Research*, 39(4):553–582.

Sargent, R. (2013). Verification and validation of simulation models. *Journal of Simulation*, 7:12–24.

Schneeweiss, C. A. (1987). On a formalization of the process of quantitative model building. *European Journal of Operational Research*, 29:24–41.

Schumacher, I. (2015). The Aggregation Dilemma in Climate Change Policy Evaluation. In *Beiträge zur Jahrestagung des Vereins für Socialpolitik 2015: Ökonomische Entwicklung - Theorie und Politik - Session: Environmental Economics IV, No. D13-V3*.

Schwartz, P. (1991). *The Art of the Long View: Planning for the Future in an Uncertain World*. Doubleday Currency, New York.

Shamir, O., Sabato, S., and Tishby, N. (2010). Learning and generalization with the information bottleneck. *Theoretical Computer Science*, 411:2696–2711.

Silberschatz, A. and Tuzhilin, A. (1996). What makes patterns interesting in knowledge discovery systems. *IEEE Transactions on Knowledge and Data Engineering*, 8(6):970–974.

Simon, H. A. (1981). *The Sciences of the Artificial*. MIT Press, Cambridge, Mass., 2nd edition.

Slonim, N. and Tishby, N. (2000). Agglomerative Information Bottleneck. *Advances in Neural Information Processing Systems (NIPS-12)*, pages 617–623.

Takeda, H., Veerkamp, P., Tomiyama, T., and Yoshikawa, H. (1990). Modeling Design Processes. *AI Magazine*, 11(4):37–48.

Tishby, N., Pereira, F. C., and Bialek, W. (1999). The information bottleneck method. In *The 37th Allerton Conference on Communication, Control, and Computing*.

Tishby, N. and Zaslavsky, N. (2015). Deep Learning and the Information Bottleneck Principle. In *IEEE Information Theory Workshop*.

Vafeidis, A. T., Nicholls, R. J., McFadden, L., Tol, R. S. J., Hinkel, J., Spencer, T., Grashof, P. S., Boot, G., and Klein, R. J. T. (2008). A New Global Coastal Database for Impact and Vulnerability Analysis to Sea-Level Rise. *Journal of Coastal Research*, 24(4):917–925.

Volkenstein, M. (2009). *Entropy and Information*. Birkhauser, Basel.

Wang, Z., Glynn, P. W., and Ye, Y. (2016). Likelihood robust optimization for data-driven problems. *Computational Management Science*, 13(2):241–261.

Weyant, J. P. (1978). Combining Energy Models: Conventions and Examples. Technical report, Stanford University, Stanford CA.

Weyant, J. P. (2017). Some Contributions of Integrated Assessment Models of Global Climate Change. *Review of Environmental Economics and Policy*, 11(1):115–137.

Willemain, T. R. (1995). Model Formulation: What Experts Think About and When. *Operations Research*, 43(6):916–932.

Zeigler, B. P. (1984). *Multifacetted Modelling and Discrete Event Simulation*. Academic Press, London.

Zipkin, P. H. (1980a). Bounds for Row-Aggregation in Linear Programming. *Operations Research*, 28(4):903–916.

Zipkin, P. H. (1980b). Bounds on the Effect of Aggregating Variables in Linear Programs. *Operations Research*, 28(2):403–418.