Statistical Methods for Effective Spatial and Temporal Scaling in Support of Probabilistic Performance Assessments – 17340

Megan Higgs, Paul Black, Amy Jordan, Paul Duffy, Tom Stockton, and John Tauxe

Neptune and Company, Inc.

ABSTRACT

Performance assessments (PAs) for the disposal of low-level radioactive waste have traditionally been performed in a deterministic manner, with inputs specified as single values as if they carry no uncertainty. Modern PAs, however, explicitly incorporate uncertainty through use of probability distributions for inputs instead of single values. While this is clearly an improvement, challenges remain in specifying appropriate distributions, many of which relate to capturing the appropriate spatial and/or temporal scale given the PA model, and matching the scale(s) of data and information to that specified by the model. More careful attention to scaling will lead to more defensible characterizations of physical variability and uncertainty in model inputs, thus leading to more defensible information regarding outcomes of interest to be used in decision making, and hence more defensible decisions. This work introduces general strategies and considerations for explicitly incorporating considerations of scale into distribution development. The strategies are explained as a series of steps and then presented with an example that starts to illustrate the new statistical concepts. This work frames the problem in terms of hierarchical statistical modeling, providing important connections to current statistical approaches to formally and rigorously account for temporal and spatial scaling. The larger goal is to build a more formal and unified approach to distribution development to help teams of modelers, subject matter experts, and statisticians work together to develop distributions and justify the choices made in the process, ultimately leading to more informed decision-making. Without this rigor, statistical distributions are not specified correctly and can have unexpected consequences for modeling and sensitivity analysis.

INTRODUCTION

Probabilistic performance assessments are an improvement over deterministic assessments because they explicitly incorporate uncertainty in inputs rather than rely on single values often deemed to be "conservative." A challenge in deterministic assessments is developing and justifying a single value to use as an input in the PA model. This requires decisions about how to use available information, and how to choose criteria for what is considered a conservative value. In fact, in order to devise a conservative value it is difficult not to envision a collection of possible values from which a large (or small) value is chosen; that is, a distribution is implicitly used in the process of deterministic modeling even if it is never formalized as a probability distribution. A conservative value is one in the tail of the distribution (right or left, depending on the context), and the tails of the distribution are often the most difficult to specify. In a deterministic setting, the rest of the distribution is ignored after the conservative value is chosen, thus

greatly restricting the information available to inform decision-making. The same value can be chosen to represent very different situations that should not be treated equally in an assessment. For example, the same value could come from a very wide distribution centered on a relatively small value or a very narrow distribution centered on a relatively large value. In contrast, probabilistic PA aims directly at formalizing the distribution and making the choices underlying it explicit and transparent, and then using the entire probability distribution in the assessment, thus propagating uncertainty to the endpoints (e.g., performance objectives), and ultimately to decision-making. Making the distribution development explicit, rather than implicit as in deterministic assessments, forces all information and decisions to also be made explicit, promoting transparency and critique, and in turn leading to more defensible assessments. Probabilistic PA may seem more complicated on the surface, but in reality it is simply formalizing and accounting for the things that are "swept under the rug" in deterministic PA.

Most PA models for radioactive waste disposal are now developed with at least some probabilistic flavor. The move towards fully probabilistic PA modeling appears to have momentum, but still a hybrid approach of deterministic and probabilistic approaches are used, despite the conundrums this creates. The hybrid approach often uses deterministic modeling for decision-making or compliance determination and probabilistic modeling only for sensitivity analysis. This is a "fool's errand" because it creates two models that are not fully compatible, in which case the decisions associated with each model are also not compatible, and are hence mutually incoherent.

Now that probabilistic PA modeling is more widely accepted, it has become more important to pay close attention to some critical—and commonly overlooked—details of distribution development in order to make the process explicit and transparent. One such detail is the challenge of capturing appropriate spatial and temporal scales of the input variables, particularly given the differences often seen in scale between the available data and that needed to match the spatial and temporal domain of a PA model. PAs must often use data representing very short and recent time scales, especially compared to the time frame over which the model will eventually be run. The same basic concerns exist over the spatial domain of the model compared to that of available data.

The process of developing probabilistic PA models requires statistical input at several stages, including site assessment, development of the conceptual site model (CSM), and planning for data collection and information gathering to inform the distributions. The CSM describes the relevant features, events and processes by which radionuclides can be transported from the disposal system to the accessible environment, and is used as the basis for developing the model structure for the system. The model consists of input variables (often called parameters) and relationships among these variables. A probability distribution is developed for each input to capture the current state of knowledge regarding the variable, but at the scale at which it will be used in the model. Distributions can be constructed based on site-specific data, data from other locations (secondary), literature review information/data, model abstraction, and/or expert elicitation. All these sources are referred to throughout this paper as information and/or data that need to be

evaluated for spatial and temporal scales, though an entire paper could be written on the challenges introduced by the use of different sources of information to inform distributions.

Information for each variable, regardless of source, should be used coherently with the CSM and modeling assumptions to develop probability distributions for the input variables. This should be done in consideration of the multivariate relationships, correlations, and spatial and temporal scales of the data in the context of a specific PA model. Depending on the source or type of the data/information, different statistical methods may be used to develop the probability distribution, and then to efficiently sample from the distributions in the subsequent simulation. Additional statistical analysis is then used to assess sensitivity of output distributions to the many different input variables [1]. Engagement of statisticians as part of the PA team is crucial to successful distribution development, as also indicated in [2]. The challenges in specifying appropriate distributions are often site- and variable-specific, although general considerations and strategies related to spatial and temporal scale apply across variables.

This paper focuses on building a more formal and unified approach to developing distributions for model inputs, focusing on the important step of recognizing the importance of spatial and temporal scales. This requires an understanding of spatial and temporal scales of the PA model relative to how values from the distributions are used in model simulations, and then requires strategies to scale the available data and information to match that of the PA model. That is, the distributions should reflect use of the data to inform a distribution at the scale driven by the model, and not be driven only by the scale of available data. Alignment of data scales to model scales for a PA model usually involves upscaling data to the spatial and temporal domain of interest for the model. By contrast, downscaling is often used in the spatial component of climate change modeling where the data represent integration over large spatial domains and need to be re-aligned to local levels of interest. Upscaling and downscaling both fall under the same general change-of-scale statistical framework. Recent developments in the field of statistics present Bayesian hierarchical models as a coherent and rigorous framework for addressing change-of-scale problems [3-5].

A general strategy for explicitly acknowledging and incorporating the issue of scaling is introduced and explained within the context of a tangible PA input variable: soil temperature. The hierarchical statistical modeling framework is introduced for the change-of-scale problem, with a focus on the conceptual ideas behind it, rather than the theoretical statistical details. Some statistical notation is introduced to facilitate communication across disciplines, but the main points of the paper do not depend on the notation.

UNDERSTANDING THE CHANGE-OF-SCALE PROBLEM

In the field of statistics, the problem of moving between different spatio-temporal scales is typically called the *change-of-support* (COS) problem, although in other disciplines it is often called change-of-scale (as used in this paper), downscaling and/or upscaling, or in geography, the modifiable areal unit problem (MAUP). It is a

problem often not adequately addressed, in part because the statistical methods for addressing COS are still evolving. PA modeling is usually associated with data collected in relatively short time periods and over relatively small spatial scales, but those data are needed to inform models that are structured to address large spatial and temporal scales. Consequently, PA modeling typically has a need for effective upscaling. If upscaling is not performed, then input distributions may carry greater uncertainty than is reasonable for the scale of the model, possibly leading to disposal decisions that are unnecessarily conservative and potentially wasteful of resources. If upscaling is not performed, or is not performed properly, then the subsequent probabilistic simulation and sensitivity analysis are compromised in a classic example of "garbage in, garbage out".

Conceptually, upscaling requires some form of integration across time, space, or both, often leading to the use of averaging that addresses assumptions of the PA model such as linearity, homogeneity, and stationarity. Under assumptions that justify averaging, upscaling is not difficult mathematically. However, the degree of upscaling relative to the data and potential lack of trust in the assumptions given the degree can create real challenges, both conceptually and from a practice perspective when trying to quantify such uncertainty. In general, the greater the degree of upscaling (the larger the difference between the data scale and the model scale), the more uncertainty is introduced through extrapolation and assumptions used to justify the extrapolation, and translating this uncertainty quantitatively into added variance in the distribution is the challenge. An approach is desired that is flexible enough to not depend on strict, and possibly unrealistic assumptions such as homogeneity over space and/or time or linear relationships between model inputs and outputs. Therefore, the focus should first be on general strategies and frameworks, rather than mathematical details of how to do the scaling that depend on the specific assumptions being made. Hierarchical statistical modeling provides such a conceptual structure for formally aligning spatial and temporal scales of the data with those of the model.

A PROPOSED STRATEGY

In this section, a general strategy is proposed as a starting point for distribution development that explicitly addresses scale from the beginning. The strategy is consistent with the hierarchical statistical modeling framework described later in the paper, but is presented first to provide a more accessible introduction and context for the presentation of the hierarchical modeling information.

Steps of the Strategy

The proposed strategy requires a thorough understanding of the PA computer model and how a random value taken from the distribution of the input variable will be propagated through the PA model, which depends on the CSM and assumptions made in translating the CSM into the PA model. This information is used to inform the rest of the distribution development process by defining the ultimate goal for what should be captured in the distribution. This is a different approach than starting first from data and not explicitly defining the scale(s) of the PA model.

The proposed strategy is presented in a series of seven steps that are briefly explained and then revisited in the context of an example. The first three steps should be considered independently of any data and/or information collected to inform the distributions. The last three steps involve connecting the information from the first three with available data or information.

Step 1: Identify scales of the PA model (temporal and spatial)
Step 2: List spatial and temporal sources of variability
Step 3: Describe distributional goal in words
Step 4: Assemble data/information
Step 5: Scale data to align with scales of PA model from Step 1
Step 6: Develop statistical model to upscale and incorporate uncertainty
Step 7: Identify and incorporate additional sources of uncertainty if necessary

It is difficult to describe distribution development issues as they apply to input variables in general, because there are fundamental differences among different types of input variables that necessitate different considerations (e.g. soil temperature versus geochemistry inputs like K_ds). Therefore, the details in the steps are not exhaustive for all inputs and additional thought may be required to identify input-specific considerations.

Step 1: Identify the spatial and temporal scales at which a specific input variable is used (or modeled) within the given PA model. This involves asking two questions: (1) what spatial region is a single random draw from the distribution representing in the model, and (2) how long is the time period over which the random draw will be used (i.e. held constant)? At this point, the information should simply be conditional on the current modeling assumptions.

Step 2: Create a list of the spatial and temporal sources of physical variability that should be captured in the distribution given the scale identified in *Step 1*. It may be helpful to start with a list of as many physical sources as can be thought of and then select only those that are still relevant at the model scales and under the assumptions of the model. Physical sources of variability are those that would persist even if full knowledge were available about the system at that scale.

Identification of physical sources of variability should take into account what is included or excluded in the structure of the model. For example, different elements or species may be specified to have their own distributions as part of the model structure, or groups of elements may share a distribution when they are similar enough and not treated differently within the model structure. For distributions meant to cover multiple elements, those elements should be explicitly stated and any differences described. Such information may help in decisions about how missing model structure should, or should not, be captured in distributions by increasing the variance of the distribution, and therefore is part of the more general problem of considering when it is justifiable to include physical variability in an input distribution as opposed to incorporating further structure into the model.

Identification of physical constraints on the input variable is also important as part of this step. Physical restrictions on possible values should be explicitly described, and can involve numerical bounds (e.g. saturation can range from 0 to 1), or can involve relationships with other input variables (e.g. saturated water content cannot exceed porosity).

Step 3: Combine the information from *Step 1* and *Step 2* to describe, in words, the goal for the distribution being developed. The words should describe the target spatial and temporal scales, as well as sources of physical variability that are relevant at those scales. The specific distributional form and values of the parameters (e.g. mean and variance) are *not* specified, but the goal is described in words. Obvious sources of uncertainty related to obtaining the distributional goal can also be described at this point, such as uncertainty related to estimation of the mean at the identified scale(s) of interest. Incorporation of additional uncertainty is discussed in *Step 6*.

The goal, or target, for the distribution is now identified and the next step is to align available data with the scale of the distributional goal identified in words in *Step 3*. Collecting or finding data representing the scale of interest is ideal, though not often realistic; hence scaling is needed to transform the data and other information to a scale that is relevant to the goal described in *Step 3*.

Step 4: Collect data and other information regarding the input variable. Explicitly identify and record the spatial and temporal scales associated with available data or other information. Different information sources will not necessarily be on the same scale, so scale identification must occur separately for each source. The scale can just be described in words for each source. If expert elicitation and/or professional judgment are used, care should be taken to explicitly identify the scale at which information is elicited, and if *Steps 1* through *3* are undertaken first, it can improve the quality of information obtained in the elicitation.

Step 5: Under the assumptions and scales of the PA model, upscale (or downscale) the available information to align with the scales identified in *Step 1*. Under common assumptions, such as homogeneity, stationarity, and linearity, it is typical to employ averaging as the mathematical means to achieve the scaling; making methods more accessible when these assumptions are not in effect is on-going work. Under the simplest scenario where enough site-specific data exist to inform the appropriate scale, this step might result in a distribution of averages at the appropriate scale (e.g. annual averages). A normal distribution might be appropriate as the form of this distribution, particularly for large scales, with mean and variance estimated from the data on the relevant scale. A value from this distribution should represent a realistic value of the input variable to use in a model run. Sources of uncertainty are formally incorporated in *Step 6*, though it is often helpful to combine *Steps 5* and *6* at least to account for basic uncertainty in estimating a mean.

Step 6: If not already incorporated, account for uncertainty due to estimation of the mean and variance of the distribution in *Step 5*. The easiest source of uncertainty to incorporate using standard statistical techniques is that due to sample size. That

is, assuming a random sample from the spatial area and time period of interest, the magnitude of uncertainty depends on the sample size and this can be easily quantified using standard statistical methods. This source of uncertainty can easily be incorporated into the distribution in *Step 5* through the framework of predictive distributions, which are commonly used in statistics for making predictions about a value of a variable at a new location or new time, beyond the specific context of the collected samples. The idea is to account for both uncertainty in the mean, and variability of potentially new measurements of the mean. Therefore, the variance of the predictive distribution is larger than the estimated distribution at the appropriate scale (e.g. annual averages).

Step 7: Given the information gathered in the previous steps, identify additional sources of uncertainty that should be incorporated into the ultimate distribution as added variance. For example, this might be uncertainty in the assumption of homogeneity over the site related to the assumption that a limited sampling is representative of the entire site, or uncertainty in the assumption that the input variable will remain constant into the future under an assumption of stationarity. In general, this step requires assessing the degree of trust in the assumptions of the model for that scale, which are often based on the temporal or spatial "distance" between the scale of the data and information available to inform the distribution and the scale of the input in the PA model. After identifying the sources, consider the magnitudes of the uncertainty of each relative to the others, and relative to the physical variability described in *Step 2* and captured in *Step 5*. It is difficult to quantify the adjustment to variance, such as through the use of an uncertainty multiplier, and this adjustment is usually based largely on expert opinion and professional judgment. Therefore, the variance of the ultimate distribution will likely be larger than the predictive distribution specified in *Step 6* unless the additional sources of uncertainty are judged to be small relative to the variance in the predictive distribution, which may be the case for variables that are not expected to vary much over time or space.

Not identified in the above steps is the potential iterative process of updating the model based on the considerations within the steps. That is, the process of explicitly laying out the assumptions can actually lead to revisions in the PA model if assumptions are called into question, in which case distribution development would go back to *Step 1*. The revisions would mostly likely be in the form of added structure to the model, which could be related to space (e.g. volume represented by a compartment), physical properties (e.g. different elements or species), or time (length of time steps).

Soil Temperature Example

The steps of the previous section are now revisited in the context of the soil temperature example. The goal for this section is not an exhaustive discussion of the distribution development, but an illustration of what can be considered for each of the steps described above. Soil temperature is an input variable to some PA models that is inherently both spatial and temporal, and is understandable across disciplines. The PA model in this example is assumed to be a 1-D column discretized with depth into compartments for vertical fate and contaminant

transport modeling. The example initially covers much smaller scales than typically encountered in PA models to illustrate the main points on a scale that can be directly informed by data. This serves as a path to considering extremely large temporal scales and the additional challenges that come with larger scales. Therefore, assume the example "PA model" will be run for one year and the random value of the soil temperature input from the distribution will be held constant over that time frame, and will be used in the computational model compartment representing the first two meters of soil below ground surface over the entire site (assume the site is a waste disposal region of area one hectare, or 10,000 m^2). One year is clearly much smaller than the time scales typically considered for PA, but this is purely for illustration.

Step 1: First, the distributional goal, or target, is identified from information about the conceptual PA model and how draws of the input will be used in running the model. The distributional goal is described in words in this step, before data are ever used (or even need to be collected). In the case of soil temperature, a single random draw from the distribution will be used to represent the soil temperature of a compartment of the PA model representing the first two meters of soil over the entire site, and held constant over one year. Therefore, the spatial scale of interest is the first two meters of soil over the entire 1-ha site, and the temporal scale of interest is one year. For another example, change in soil temperature as a function of climate change could be considered.

Step 2: Examples of physical sources of spatial variability causing vertical and lateral heterogeneity are changes in materials (bedrock or unconsolidated soils), mineralogy, porosity, and saturation (factors that affect bulk thermal conductivity); differences due to aspect (angle of the surface relative to solar radiation); and ground cover types (e.g., plants, rock fragments, asphalt) with differing densities and albedo. These identified sources of variability should be evaluated for their relevance based on the scales identified in *Step 1*, and in doing so, it is clear that all of these sources of physical variability over space occur at a scale smaller than the first two meters with depth and laterally over the entire site.

Spatially, the values captured in the distribution should be realistic to represent the entire site to a depth of two meters, making the mean soil temperature of that volume of soil the quantity of interest. The focus on the mean comes from the assumption in the modeling that one random value from the distribution can be used to represent the soil temperature of that volume for a run of the model. Data from the volume of interest, if available, can be used to estimate the mean over the site, with a distribution of the mean (or a sampling distribution of the average) used to quantify uncertainty in knowledge of the mean. The uncertainty depends on the amount of data available, whether the data can be treated as if they are a random sample from the area of interest, and how representative the data are judged to be of the entire volume of interest. The decision to use the mean should be supported by knowledge of the site or model assumptions, such as believing soil temperature is fairly homogenous across the site within the depth range and time frame of interest, that the response is linear, and that soil temperature is stationary both spatially and temporally. Note, if data are not available from the site of

interest, then the context of predicting the mean soil temperature for the site (at the identified spatial scale) from data collected elsewhere also becomes relevant.

Using a distribution that includes physical sources of variability smaller than the scale of the model implies that when a random value is taken from the distribution, it will be used as if it represents the entire spatial region when really it only represents a potentially small part; thus the model is run at unrealistic values relative to the chosen scale. If there is justification to include smaller scale sources of variability, such as differences by depths within two meters for this example, then this may be an indication of a need for additional structure in the model. The appropriateness of using a wider distribution in place of additional model structure is context dependent and should be made by the team of subject matter experts and statisticians. That is, the decision to represent the upper two meters over the entire site as a single compartment implies that either the site is homogeneous enough to assume temperature is constant over that volume, or that identified sources of physical variability are not judged to be important enough to the system to model (such as differences in soil temperature due to aspect). For other inputs, such as K_ds or K_Hs (air/water partition coefficients), it may be important to carefully consider the potential effects of lumping elements or species into a single distribution (e.g. Jordan et al. [6]).

Sources of temporal physical variability include the daily and seasonal cycles through minimum and maximum ranges, as well as even longer potential episodes of periodic climate trending and/or climate change. From a temporal perspective, a random draw should be realistic to be held constant for one year, and therefore physical variability in soil temperature for scales less than one year (daily, weekly, monthly, seasonal variation) does not need to be captured in the distribution (assuming linearity in smaller scale responses). Values meant to represent an entire year (because they are held constant over an entire year) should come from a distribution of yearly averages. If, for example, seasonal variability is included as extra variability in the distribution, then the model will be run for the entire year at winter values or summer values when it is not realistic for an extreme seasonal temperature to be held constant for an entire year. If ignoring such variability raises flags, then the model structure of how individual draws are meant to represent a time period can be revisited. For example, it may be possible to construct the model so that new draws can be taken to cover smaller time steps over the run of the model, though this may cause challenges computationally and for the subsequent sensitivity analysis.

For the soil temperature example, variability in annual averages does need to be captured because that is the identified scale of interest; this is a distribution capturing physical variability in average soil temperatures that would still exist even if all information were available for millions of years about average annual soil temperatures. The distribution will, however, be estimated from data and therefore there will also be uncertainty in the mean and variance of the distribution, and that uncertainty should be included as additional variance in the distribution.

Step 3: The goal for this step is to combine the information from *Step 1* and *Step 2* to explicitly describe the goal for the distribution (i.e., the distributional target). A

challenge for some input variables at this step, such as soil temperature, is to combine the spatial and temporal considerations into a single goal. For this example, the end distribution should capture uncertainty in estimating the mean annual soil temperature of the volume of soil in the upper two meters across the site, along with the physical variability expected in annual averages of the mean soil temperature. A predictive distribution for annual averages would allow a new realistic annual average to be drawn for each run of the model, while simultaneously accounting for uncertainty in estimating the mean. However, final decisions about how to capture the necessary information in a specific distribution are made after collecting relevant data or assessing the information available, which happens in the following steps.

Step 4: In this step, data and information are gathered to inform the distribution. It is possible data could be designed and collected explicitly to inform site-specific distribution development, but it is more common that relevant data and information already available are gathered to inform initial distribution development. For the soil temperature example, suppose 25 years of data from a single sensor are available at a depth of one meter below the surface, at a single location within the site. Soil temperature is recorded from the sensor every 12 hours (in the afternoon and early morning) to capture the daily extremes. Temporally, data are available on a scale smaller than the annual scale identified in previous steps, but they can easily be aggregated to annual averages to match the scale of the model. Spatially, data are available from only a single location and a single depth, clearly representing a small spatial area relative to the spatial region the distribution should represent in the model. With one location, there is no physical averaging over space as would be considered with multiple soil samples, so a distribution for the mean cannot be developed purely over space. Figure 1 shows some illustrative simulated data.

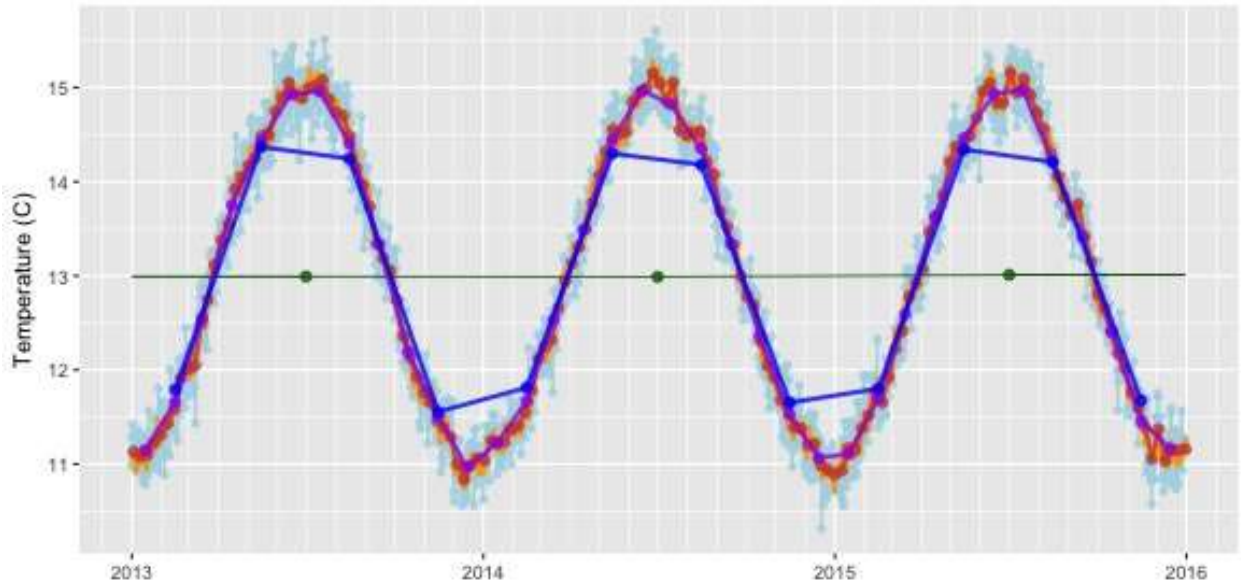

Figure 1. Example soil temperatures reported at different scales over three of the 25 years of simulated data (orange=true soil temperatures every 12 hours, light-blue=measured soil temperature every 12 hrs, red=weekly averages, purple=monthly averages, dark-blue=3 month averages, green=yearly average).

Steps 5 and 6: The temporal and spatial components represent different problems, the former where there are data to directly inform the needed scale, and the latter where the upscaling must be justified based on assumptions. For the temporal component, there are 25 annual averages available to directly inform the distribution at the appropriate scale. Therefore, the 25 observations of annual averages can be used to estimate a predictive distribution of annual averages to combine the physical variability in annual averages with uncertainty in estimating the parameters of that distribution. This can be accomplished using standard statistical tools in most cases, and note the scale of the predictive distribution should match the identified scale(s) used in the model.

For the spatial component, the upscaling to the entire site must be based completely on assumptions, the simplest of which is to assume the site is homogenous in terms of soil temperature. Another possibility is to assume that variability in annual averages over time can be used as a surrogate for variability in the annual average over space; this is convenient because the distribution of the mean in space and time is then estimated with the same data, but it also requires assumptions that are difficult to check and justify. There is no single correct choice here, but the assumptions used to justify the choice should be stated and defended as much as is possible, based on expert judgment, knowledge of the site, and information gathered during development of the conceptual site model. In this example, data exist from the site of interest so predicting at a new (i.e. different) location is not of interest; however, uncertainty due to extrapolation (predicting values for a site with no data of its own) should be a consideration when data are only available from other locations.

Step 7: As described in the previous paragraph, having data from only a single location when the model scale is an entire volume over the site forces reliance on assumptions to go ahead with distribution development and the assumptions should be justified based on expert knowledge and professional judgment. A lack of trust in an assumption should be translated into additional variance in the distribution; a major challenge of distribution development is quantifying the extent to which the variability identified in Steps 5 and 6 should be increased to account for sources of uncertainty in assumptions. For this example, there is uncertainty in the assumption of homogeneity over space — is the site homogeneous enough in terms of soil temperature to use a single location to estimate the mean? Or, is the variability in the annual averages over time similar enough to cover the variability that would be expected if multiple locations were monitored? Even in the temporal context, are the 25 years of data of annual averages trusted to represent annual variability in the far future, if that is necessary for the model? Again, there is no single correct choice, with the decision ideally being made by a team including subject matter experts and statisticians. A productive attitude toward distribution development is not finding *the* correct distribution, but a distribution that is justified based on the information and expertise available — it is not realistic for every team to represent the current state of knowledge about an input variable with the same distribution.

At this point, it is worth realizing that often distribution developers find themselves facing a single spatial location or perhaps few locations, as well as data from a much shorter time period than the time scale of PA models (i.e. it is not possible to have multiple 10,000-year averages with which to inform a distribution). How much trust is there in assumptions that must be made to obtain a distribution of 100-yr or 1000-yr averages in soil temperature in the future? There is nothing inherently wrong with making the needed assumption, but uncertainty inserted by doing so should be reflected in the distributions to the extent possible. Additionally, site data might not even be available and distribution development must rely on data from other sites, additional information found in the literature, or expert opinion. While the challenges are many and will often force considerations beyond those explicitly addressed in this paper and example, the ideas presented here are meant to help provide some structure to approaching the development. This more formal path to distribution development is technically defensible, and is hence superior to arbitrarily using data distributions that do not attend to the spatial and temporal scaling issues.

Scaling for Distributions and Simpson's Paradox

This section provides additional detail to illustrate issues related to scale and why they are important for distribution development. The information builds on discussions presented in *Step 2,* and ties them to a well-known result within Statistics called Simpson's paradox (a.k.a ecological fallacy or ecological effect). Cressie and Wikle [3] discuss the connections between change-of-scale problems and Simpson's paradox, and suggest the usefulness of the connection in understanding the problem.

Simpson's paradox describes a situation where different relationships exist, or are observed, at different scales. Simpson [7] demonstrated the "paradox" with a contingency table, showing that a positive relationship between two variables displayed in a 2-way table can actually reverse to a negative relationship when a third variable is used to further divide the tables. A famous example discussed by Meng [8] and Cressie and Wikle [3] describes results from a comparison of open surgery and ultrasound in the treatment of kidney stones. In data reported by Charig et al. [9], open surgery was successful in 78% of the cases and ultrasound was successful in 83% of the cases. However, when the size of the stone was taken into account, the success rate for open surgery on small stones was 93% compared to 87% for ultrasound, and likewise for large stones it was 73% compared to 69%, respectively. The "paradox" name is given because people intuitively expect the results to follow the aggregated pattern. This example highlights the general importance of considering the appropriate scale, or what should be conditioned on (like size of stone), for any statistical analysis, and this should be dictated by the question of interest or goal of the analysis. Cressie and Wikle [3] further illustrate the point in an environmental context:

> "…*day trading on stock markets, based on economic relationships estimated from quarterly trade figures, would probably lead to financial ruin. In a spatial setting, regional climate data may warn, correctly, of a future drought in the Northwest United States (states of Washington and Oregon). However, local orographic effects may favor certain parts of the Willamette Valley in Oregon to the point where above-average rainfall is consistently received there*."

How does this relate to change-of-scale problems for PA modeling? It is helpful to continue with the soil temperature example, where data are available over time at small scales. Clearly, annual averages hide all smaller-scale fluctuations (e.g. seasonal differences – see Figure 1), and therefore tell a different story about soil temperature than summaries at smaller scales. It is helpful to actually look at distributions of averages at four different scales (daily, weekly, monthly, and yearly) using the 25 years of data from the soil temperature example, without the added complications of attempting to account for sources of uncertainty (Figure 2). The averages of the distributions are the same (about 13°C), but the change in shape as scale increases is clear (note the number of observations going into the histograms is also changing), and the variance is far smaller in yearly averages than it is in the smaller scale distributions that capture seasonal fluctuations.

The proportion of daily, weekly, and monthly averages less than 12°C is close to 0.33, whereas it is zero for yearly averages. There are relatively few weeks, and especially months, that actually have a soil temperature near the values in the yearly soil temperature distribution. That is, soil temperature aggregated to the yearly scale does not tell the same story as soil temperature at the daily or monthly scale. Therefore, it would be a mistake to attempt to capture what is going on at the monthly scale using yearly summaries; and a mistake to use yearly averages to capture a behavior of monthly temperatures. This may seem intuitive and obvious for this tangible example, but it gets more complicated as the scales of interest are farther from those of the data. The degree to which the stories will match at

different scales depends on how homogeneous the process is. For the example, soil temperature is *not* homogenous within a year in the example, though if soil temperatures were taken from a location near the equator, the monthly and annual story might be similar and the assumption of homogeneity is clearly more justified. The usefulness of a model, and associated distributions, depend on identifying and acknowledging appropriate scales.

The take-home message for PA distribution development is that averages, or relationships among variables, observed within a particular group (or at a particular scale), do not necessarily remain the same when groups are aggregated or disaggregated (or scale is changed). Therefore, it takes creativity and statistical work to carefully assess the implications of the use of different spatial and/or temporal scales for the particular scientific endeavor. Instead of assuming the relationships within subgroups (or smaller scales) are relevant to larger scales, or vice-versa, explicit statistical modeling should be used to account for such change of scale differences. And, in the absence of site-specific data to support such formal modeling, literature review, expert opinion and professional judgment must be relied on to incorporate additional information and its associated uncertainty.

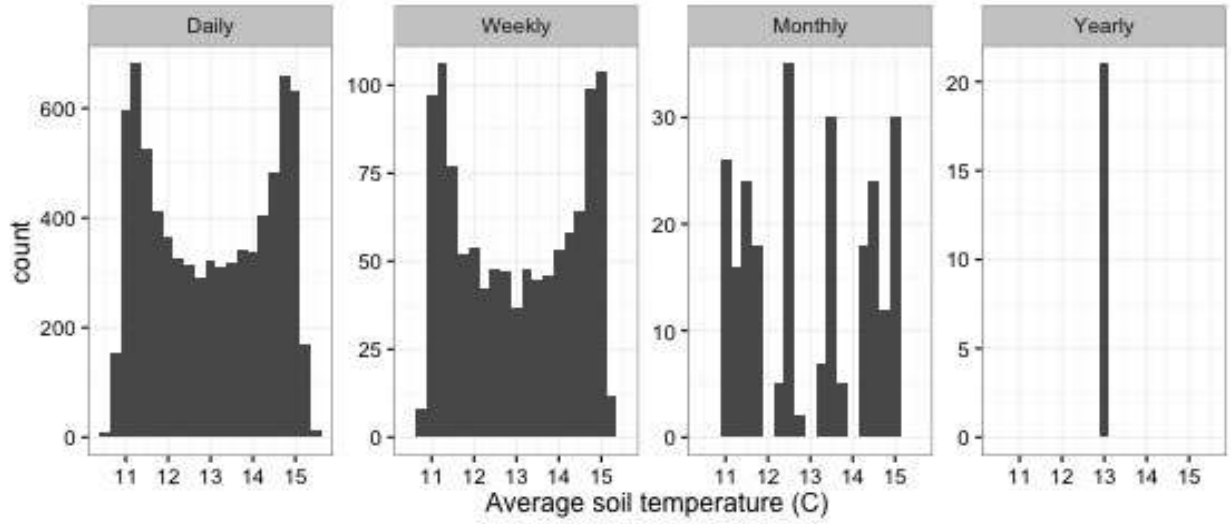

Figure 2. Histograms of daily, weekly, monthly, and yearly averages from the entire 25 year data set where three years were shown in Figure 1. The bin width is 0.25°C for all histograms and the x-axis scale is held constant.

A FRAMEWORK FOR MOVING ACROSS SCALES

Statisticians have been working toward developing flexible statistical methods to deal with the change-of-scale problem, and hierarchical modeling provides a natural framework for addressing the problem [3, 4]. This section presents hierarchical modeling concepts in a very general sense and connects these to the needs in PA modeling. References and connections are made to the soil temperature example when relevant, and while the information in this section is presented with alternate language, the ideas are consistent with those underlying the strategy presented in the previous sections.

Thinking in Terms of Hierarchical Statistical Models

Hierarchical models provide a very general framework for incorporating natural levels of model structure and the uncertainty associated with each. The "hierarchical" refers to the simultaneous modeling of the different levels by taking advantage of the hierarchy to model a level conditional on knowledge of other levels. The conditioning simplifies complicated problems by breaking them down into manageable steps by modeling one level at a time, after conditioning on others. Conceptually, it can be thought of as simplifying the process of modeling by breaking it down into a sequence of manageable steps, rather than attempting to tackle it as one large problem.

Berliner [5] introduces a relatively simple idea in the context of hierarchical models that has led to many statistical and scientific developments. The idea is to efficiently combine the model for the observable data $[data]$ with the model for some underlying and unobservable "process" $[process]$ using hierarchical modeling. The term "process" is commonly used in the Statistics literature as a very vague term simply meant to capture a higher conceptual level of a model that the collected data and information are meant to inform. In the context of the soil temperature example, the "process" of interest is defined, for the current PA model, as the true distribution of annual averages over the spatial area and going out thousands of years into the future. The problem is placed in a hierarchical framework by combining conditional distributions in a way that follows the natural hierarchy. The model for the data is considered conditional on the model for the process, which greatly simplifies the problem (conceptually and mathematically) when modeling potential outcomes of data assuming the process generating the data is known. Then, the process is modeled, again by taking advantage of conditioning to consider the process conditional on unknown parameters defining it. Finally the unknown parameters used to define the process part of the model need to be assigned distributions reflecting current knowledge in the parameters. To simplify notation, the square bracket notation for models and distributions is adopted here; for example, the distribution of the data is denoted $[data]$, and the distribution of the data given some assumed process is denoted as $[data|process]$ (read "the distribution of the data given the process"). Using this notation, the ideas in this paragraph are formalized in Eq. 1,

$$[data, process, parameters] = [data \mid process, parameters][process \mid parameters][parameters]. \quad \text{(Eq. 1)}$$

The elegant idea of conditioning in this way has been used to tackle very difficult problems in many disciplines, and can involve many more levels than described here. To simplify notation, letters are typically used in place of the words, with $data = Z$, $process = Y$, and $parameters = \theta$, so Eq. 1 is rewritten as

$$[Z, Y, \theta] = [Z|Y, \theta][Y|\theta][\theta]. \quad \text{(Eq. 2)}$$

Usually, the goal is to make conclusions about the process model and its parameters, which is accomplished using the distribution conditional on (or "given")

the observed data, denoted $[Y, \theta|Z]$. This is termed a posterior distribution, and can be obtained (up to a proportionality constant) using the right side of Eq. 2.

$$[Y, \theta|Z] \propto [Z|Y, \theta][Y|\theta][\theta]. \qquad \text{(Eq. 3)}$$

Another distribution often of interest is that of $[Y|Z]$, obtained after integrating the posterior distribution in Eq. 3 over θ. This distribution is termed a posterior predictive distribution; the goal is to use the distribution of Y *given* all available information in Z.

Hierarchical Models and Change-of-Scale

Change-of-scale discussions in a statistical context typically refer to the *support* of a random variable. Support can be thought of as a more formal mathematical term of *scale* and both terms are used in this section. A crucial addition to the notation presented in the previous subsection is the incorporation of a letter (M) to denote the scale at which the variable is being considered [3, 4]. That is, in the hierarchical $data - process - parameter$ framework, the initial focus should be on identifying the scale M at which Y is needed and then on the data and information available to inform the identified scale for Y. The added notation explicitly acknowledges the scale of both the desired Y and the available data Z, and easily incorporates the fact that different data sources can have different supports (or scales) that can be finer or coarser than M (e.g. $Z(A)$ and $Z(P)$ can be used to inform $Y(M)$). This is in contrast to simply thinking about Y and Z without reference to scale, as done in the previous subsection.

The stochastic process $Y(M)$ with support M describes the general $process\ Y$ after it is integrated over region M: $\frac{1}{|M|}\int_M Y(u)du$. Practically, the support is the spatial and/or temporal resolution at which the process is being investigated (i.e. the level of aggregation or the scale). For example, in the context of time, M might be one year (as in the soil temperature example) so that $Y(M)$ is defined by integrating the temporal process annually to match the scale M of the model. The end goal depends on M. For example, changes in the scale of the PA model will change the desired support from M to something different, such as $P = 1000$ years, and then $Y(P)$ would be of interest instead of $Y(M)$.

The scale at which the data are available is usually different from the desired scale for the model (i.e. $Z(M)$ is not available, but $Z(A)$, $Z(D)$, and $Z(H)$ may be). The importance of accounting for scale really means that attention should be paid to extracting information from $Z(A)$, for example, to inform $Y(M)$, rather than mistakenly using $Y(A)$ in the model simply because data are available with support A. Accounting for scales should be an explicit step within the distribution development process, as described in earlier sections of this paper.

Connections to PA Modeling

Cressie and Wikle [3] describe the practical and statistical aspects of the problem in a contemporary way. By approaching the problem from a coherent probabilistic viewpoint through use of formal statistical hierarchical modeling, they are able to

build spatio-temporal statistical models allowing for scientific inferences at various spatial and temporal scales. They also see the possibilities of better meshing of spatial and temporal statistical models with physical models, which they term scientifically-based dynamic models. The approach is clearly useful in data-rich settings with strong scientific buy-in to underlying physical models. The PA setting, however, is typically data-poor instead of data-rich. Therefore, the immediate benefit of these methods to PA is not to replace the current PA modeling strategy, but instead to improve the stochastic, or probabilistic, part of the model by formalizing the strategy of developing probability distributions for inputs. There is also value in becoming familiar with basic language and ideas, because it is possible that PA modeling may adopt more of the principles underlying "scientifically-based dynamic models" as it evolves. For now, borrowing the general framework to work on the change-of-support problem within the current PA modeling framework provides a convenient and needed started point.

To make the connection to PA modeling, notation is extended to include numerical subscripts on Y and Z to distinguish among different input variables and different sources of information for a particular input variable. For example, $Y_1(M)$ refers to input variable 1 at scale M, and $Z_{1,2}(A)$ refers to the second source data, available at scale A, to inform the distribution of $Y_1(M)$ (the distributional goal or target as described in previous sections).

The observable data and information available at different scales is used to inform the input distributions at the PA model scale, as shown in Figure 3. In the figure, Zs represent different sources of information with associated scale given in parentheses (letters higher in the alphabet represent larger scales) and Ys represent the variables whose distribution we desire to use in the PA model. The goal is to pay attention to scale so that the distributional goal for the input Y is on the scale matching the PA model (i.e. $Y(M)$ in this case), and then the data and information are scaled to match M as well. The distribution for a Y is conditional on the data and information potentially available at different scales.

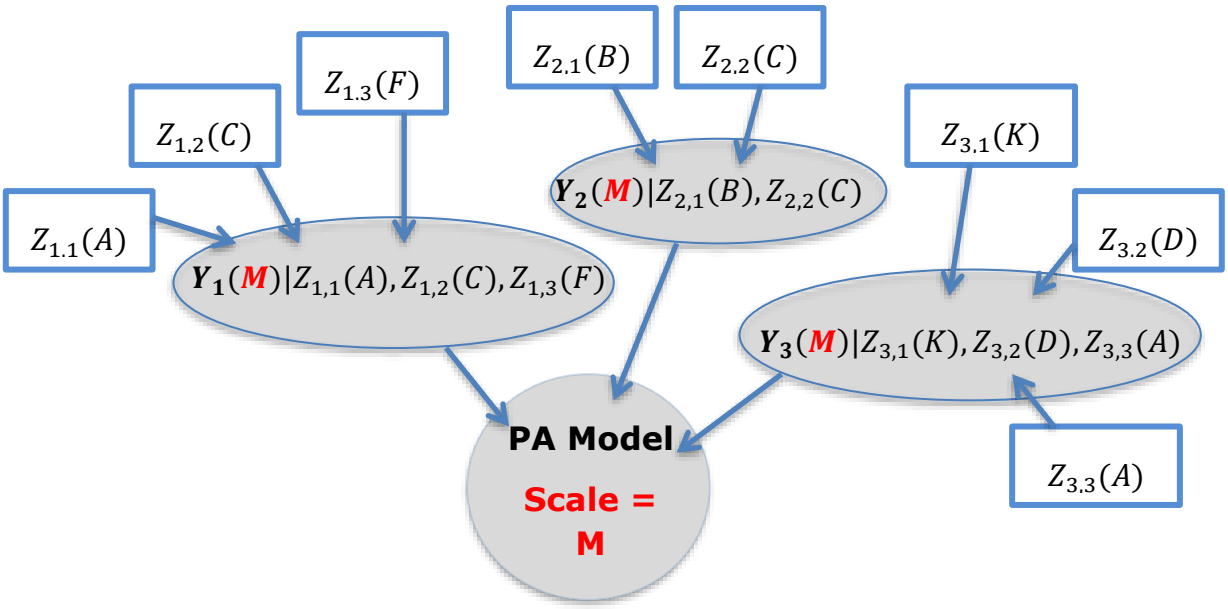

Figure 3. Conceptual diagram connecting probabilistic PA models to the hierarchical modeling framework.

While M is used to broadly represent the scale of the input as used in the PA model, it may not be the same for all inputs, and the notation could be further developed to differentiate between the temporal and spatial scales (some inputs may share a temporal scale, but differ in their spatial scale given the structure of the PA model). Related to the strategies presented earlier in the paper, a $Y_j(M)$ alone can be thought of as the distributional goal or target, where the scale of interest is defined by the underlying PA model. A specific distributional form and values of the parameters (e.g. normal with specific mean and variance) is not specified until assessing available data and other information. In reality, $[Y_j(M)]$ is not attainable, so instead it is developed *given* available data and/or other information (e.g. $Z_{j,i}(A)$ and $Z_{j,k}(B)$). That is, $[Y_j(M)|Z_{j,i}(A),\ Z_{j,k}(B)]$ captures, as best possible, the current state of knowledge regarding the input variable at the desired scale. Even more generally, the distributional goal can be denoted as $[Y_j(M)|$*state of knowledge*$]$. Therefore, as described previously, the challenge is to combine information from multiple scales to inform the input variable distribution on a yet another potentially different scale, and to do this in a rigorous and defensible manner.

The objectives of this paper are to clarify the goal for PA distribution development, and to make the goal more explicit and reachable by explicitly acknowledging and accounting for scale through appropriate statistical modeling.

DISCUSSION

While scale is the focus of this paper, there are certainly other challenges in distribution development that deserve mention and future attention, and many are related to the problems of scale.

Types of information available to inform PA input distributions fall on a continuum of relevance in information, from site-specific data to possibly obscure literature sources, and also cover expert elicitation and professional judgment. Data from different literature sources or experts will differ regarding overall quality of information, quantity of information, and relevance of the information to the site [6], each of which should be viewed in the context of scales. Formal incorporation of such characteristics of data and information is a needed area of attention for PA distribution development.

As mentioned previously, the idea of using, or at least conceptualizing, predictive distributions at a scale that matches the model may be a way to proceed if values used in the PA model represent times or locations not directly informed by available data. Predictive distributions allow formal combining of distributions describing physical variability (from data after scaling) and the distributions describing uncertainty in the mean, or other unknowns (given available data). It also easily extends to formal incorporation of prior distributions reflecting the current state of knowledge in inputs, or the mean of inputs, before incorporating new data. A challenge for many inputs comes from the fact that PA models are currently run at large time scales, and predictive distributions are harder to conceptualize at large time scales with no data to directly support them at the identified scale.

CONCLUSIONS

This paper highlights the importance of considering scale and introduces statistical hierarchical modeling as a conceptual framework to formalize the scaling problem in the context of performance assessments. The goal is to describe useful strategies and help create a coherent framework that can naturally account for changing scales to facilitate work and discussion regarding the role of scaling in distribution development. An additional goal is to document distributional development considerations and strategies, as an important step toward a unified and consistent approach (e.g. [6]). A conceptual framework is a first step in making progress in addressing scale explicitly across the PA community. A strong background in statistics is not needed to benefit from the conceptual approach, but is needed to implement this statistical strategy. What is critical is that statisticians and scientist or subject matter experts work together to develop a probabilistic PA model that addresses the upscaling needs given the data and the temporal and spatial domain and structure of the PA model.

The example of soil temperature illustrates challenges commonly arising in distribution development and serves to provide a tangible context within which to discuss much larger challenges. This work raises awareness of timely and important issues that must be addressed to continue improving risk and performance assessments. Although some simple efforts at scaling have been attempted in the past for probabilistic performance assessment model inputs, these examples show how a hierarchical Bayesian structure can be used to more formally account for temporal and spatial scaling. This is critical for improving the way performance assessment models are used to inform decision-making.

REFERENCES

1. Stockton, T., P. Black, and P. Duffy, *Modern Sensitivity Analysis.* WMS Journal, 2016. **1**(3).
2. EPA, *Risk Assessment Guidance for Superfund, Volume I Human Health Evaluation Manual (Part A)*. 1989, United States Environmental Protection Agency: Washington DC.
3. Cressie, N. and C.K. Wikle, *Statistics for Spatio-temporal Data*. Wiley Series in Probability and Statistics. 2011, Hoboken NJ: John Wiley & Sons Inc.
4. Wikle, C.K. and L.M. Berliner, *Combining Information Across Spatial Scales.* Technometrics, 2005. **47**(2005): p. 80–91.
5. Berliner, L.M., *Hierarchical Bayesian Time Series Models*, in *Maximum Entropy and Bayesian Methods*, K. Hanson and R. Silver, Editors. 1996, Kluwer Academic Publishers: Dordrecht, Netherlands. p. 15–22.
6. Jordan, A., et al., *Realistic Geochemical Parameter Uncertainty for Performance Assessment Modeling—17066*, in *WM2017 Conference, March 5–9*. 2017: Phoenix AZ.
7. Simpson, E.H., *The Interpretation of Interaction in Contingency Tables.* Journal of the Royal Statistical Society, Series B, 1951. **13**(1951): p. 238–241.
8. Meng, X.L., *Statistics: Your Chance for Happiness (or Misery).* The Harvard Undergraduate Research Journal, 2009. **2**(2009): p. 21–27.
9. Charig, C.R., et al., *Comparison of Treatment of Renal Calculi by Open Surgery, Percutaneous Nephrolithotomy, and Extracorporeal Shockwave Lithotripsy.* British Medical Journal (Clinical Research Edition), 1986. **292**(6524): p. 879–882.