Characterizing and Visualizing Predictive Uncertainty in Numerical Ensembles Through Bayesian Model Averaging

Abstract—Numerical ensemble forecasting is a powerful tool that drives many risk analysis efforts and decision making tasks. These ensembles are composed of individual simulations that each uniquely model a possible outcome for a common event of interest: e.g. the direction and force of a hurricane, or the path of travel and mortality rate of a pandemic. In this context, numerical ensemble data represent a complex exploration of a hypothesis space where each simulation explores a possible outcome based on distinct set of initial modeling parameters and internal processes. Paramount to ensemble analysis is quantifying and characterizing the ensemble’s predictive uncertainty, i.e. the ability for ensemble constituents to accurately and consistently predict an event of interest based on ground truth observations. This metric is key to defining what aspects of the hypothetical space are valid and how assumptions behind these hypotheses can be tuned to increase predictive accuracy.

This paper directly addresses the challenge of deriving insight from ensemble data and reducing predictive uncertainty in an ensemble through a new visualization strategy that is based on a Bayesian framework. We use the framework to first construct a statistical aggregate from the ensemble. We extend the information obtained from the aggregate with a visualization strategy that characterizes predictive uncertainty at two levels: at a global level, which assesses the ensemble as a whole, as well as a local level, which examines each of the ensemble’s constituents. Through this approach, modelers are able to better test and refine their models and derive greater understanding about the event of interest. We apply our method to two datasets to demonstrate its broad applicability.

Index Terms—uncertainty visualization, numerical ensembles, statistical visualization

1 INTRODUCTION

Visualization plays a powerful role in addressing many forms of information uncertainty. Applications that use visual cues and illustrative quantifiers to indicate information variability or depict levels of confidence in data values help analysts better understand and cope with uncertain information. The insight obtained from such applications is essential for risk analysis and decision making tasks. Visualizing uncertainty, however, is challenging; information uncertainty is a complex concept with many domain-specific definitions and interpretations.

Common across most scientific domains is the uncertainty associated with model conceptualization. Specifically, conceptual model uncertainty is the uncertainty associated with correctly selecting a set of domain-specific interpretations, processes, and mathematical systems to estimate - through numerical simulation - the behaviors of some event of interest (e.g. a weather system or the direction of subsurface aqueous flows). This uncertainty is arguably the greatest source of error and risk associated with modeling and forecasting, and is the primary uncertainty considered by risk analysis [1, 11, 29, 40].

One of the most powerful ways to address conceptual model uncertainty is with an aggregate prediction: a prediction that combines all ensemble constituents through a weighted average [2, 28, 29, 31, 41]. In the aggregation process, the numerical ensemble consists of a collection of individual numerical simulations, where each simulation is based on a unique conceptual model defined by a distinct set of initial conditions and modeling parameters. The premise behind aggregation strategies is that none of the ensemble members can perfectly model the event of interest because the specific processes and systems governing the event are largely unknown. The best model, therefore, must be estimated via a weighted combination of the ensemble’s constituents; estimates for these weights are obtained through statistical methods: e.g. bootstrap aggregation [4] and Bayesian Model Averaging (BMA) [22]. In practice the premise behind aggregated prediction strategies is valid as aggregate predictions are known to provide better overall predictive performance than the individual ensemble constituents [28, 31, 41, 45, 50].

Though aggregation strategies help ameliorate conceptual model uncertainty (by eliminating the need to rely on a single conceptual model) and can provide better predictive performance, they also have an inherent shortcoming: aggregate predictions don’t help modelers better understand the processes and mechanisms governing the target event of interest. More specifically, the weights used to combine the ensemble’s predictions don’t generalize to combine the underlying conceptual models in the ensemble that represent distinct and often conflicting hypothesis. To better address conceptual model uncertainty, modelers need visualization strategies that can augment the benefits of aggregate prediction with information feedback loops that can help modelers tune conceptual assumptions and converge on the correct set of domain-specific interpretations, processes, and mathematical systems. Only through such a process can modelers reduce an ensemble’s predictive uncertainty: a measure of conceptual convergence that measures the ability for a numerical ensemble’s constituents to consistently and accurately predict an event of interest based on ground truth observations.

This work presents a visualization strategy that leverages BMA to generate an aggregate prediction. BMA is used in this work because BMA-based predictions are highly accurate [22, 28, 50], are known to outperform predictions obtained from other statistical methods [9, 16, 22, 47], and are successfully applied across many domains [29, 38, 40, 45, 46]. We use BMA’s prediction to direct a two-step visual analysis. In the first step, the aggregate prediction helps quantify ensemble bias and variance (components of predictive uncertainty) based on ground truth measurements for an event of interest. These two metrics support a spatial segmentation strategy that identifies regions of low and high predictive uncertainty for the ensemble. In the second step, we analyze the value of individual ensemble members within the context of the ensemble to identify where specific constituents (and their conceptual models) help increase predictive accuracy. Combined, these visualization strategies help modelers identify and characterize both successful and unsuccessful trends in an ensemble. Through iterative application of this method during cycles of design and simulation, modelers can identify valid conceptualizations in the model design process and ultimately reduce conceptual uncertainties.

Overall the contributions of this work are as follows.

• We present a new analysis method to quantify and characterize predictive uncertainty in numerical ensembles. The strategy uses a Bayesian framework that combines sparse ground truth observations and numerical ensemble data to form an aggregate prediction. In Section 3, we use this approach to address an important sparse data challenge and to form the basis of our visualization method.
• We extend this framework with a visual segmentation strategy that highlights regions where models accurately and consistently model a target event (i.e. areas where the event’s behavior is well understood) and areas where more research is required to improve predictive accuracy.

• We augment this visualization technique with a second visualization method that characterizes and segments the predictive strengths and weaknesses of individual ensemble constituents. This second method identifies the ensemble members (and therefore conceptual models) that are more accurate than other ensemble constituents.

In the next section we discuss work germane to our efforts. We then present a brief overview of uncertainty as it affects ensembles and ensemble analysis. This overview is followed by the presentation of our method (Section 3). We then apply our method to two datasets to demonstrate the method’s ability to characterize and help quantify predictive error and predictive uncertainty (Section 4). Finally, we make concluding remarks and discuss future work (Section 5).

2 PREVIOUS WORK

2.1 Visualization and Uncertainty

Much has been done to establish a basic foundation for uncertainty visualization. To address the complex, multifaceted nature of uncertainty, many visualization researchers have presented varying formal models and typologies. Pang et al. [34] present a model that delineates three types of uncertainty. These uncertainties—which correspond to sequential stages in the visualization pipeline—include the uncertainties associated with data measurements and collection, data transformations, and the uncertainties associated with visualizing the data itself. Plewe’s [35] complementary work defines the uncertainty associated with process and model conceptualization. This form of uncertainty, which we refer to as conceptual model uncertainty, arises when models are first being selected and designed, and occurs before the uncertainty stages defined by Pang et al., and is viewed as the greatest source for error in modeling and forecasting tasks [1, 11, 29, 40].

Thomson et al. [44] present a formal typology for uncertainty in the context of geospatial data. This work, along with MacEachren et al. [26], suggests that uncertainty is best addressed by jointly considering multiple categories of information that affect uncertainty, such as data error, precision of information, completeness of data, and consistency of information sources. Thomson et al. also provide statistical measures for each category to aid in quantifying uncertainty.

Gershon [17] provides a high-level taxonomy of causes for the state of imperfection in knowledge. Gershon stresses that the transformation of data to visualization can significantly increase uncertainty if the method for visualization is not carefully selected. Coulceis [8] presents a classification strategy that records the unknowable aspects in information to direct awareness to the knowable aspects. Johnson and Sanderson [23] provide an overview of uncertainty visualization methods and identify the need for a theoretical framework for error and uncertainty visualization.

Pang et al. [34] classifies uncertainty visualization strategies by matching data types (scalar, vector, tensor, and multivariate data) to discrete or continuous visualization techniques such as glyphs or stream lines. Others have proposed similar, domain-specific classifications that map quality of data to visualization methods. [5, 6]

Many researchers have addressed the perceptual issues of visualizing uncertain data. Several researchers have employed transparency to convey uncertainty in data values [10, 25, 26, 30, 33]. Color is a principal channel that has also been used to convey uncertainty information [10, 21, 30]. MacEachren [25] explores four aspects of color to convey information about uncertain data including color saturation, crispness, transparency, and image resolution. Hengle [21] explores mapping predictions and error into the hyper spectral imaging color space with a color mixture concept.

Several researchers have also conducted user studies to assess the best approaches for visualizing uncertainty [13, 30, 42]. Sanayal et al. [42] conducted a user study that focuses on validating uncertainty visualization techniques for 1D and 2D datasets. They observe that error bar visualization techniques are largely ineffective for conveying uncertainty information. Additionally, they note that the effectiveness of uncertainty visualization techniques is heavily task-dependent. Newman and Lee [30] examine eight techniques in the context of volumetric data. They conclude that multi-point glyphs and ball-and-arrow glyphs are highly effective in representing uncertainty in volumetric data. Edwards and Nelson [13] evaluate the effectiveness and accuracy of four methods for representing uncertainty on circle maps. They find that color-based strategies are superior to the legend-based techniques for conveying uncertainty. They also find that integrated display windows work better than separate display windows because separate displays create more perceptual and cognitive work for the reader.

Visualization techniques themselves have been the focus of uncertainty analysis with the goal of characterizing the bias that visualization techniques place on the data [12, 19, 37]. Prassni et al. [37] address the uncertainties associated with slice techniques as applied to volumetric data and propose a guided probabilistic volume segmentation strategy. This strategy directs the user’s attention to regions of high uncertainty to avoid misclassification of medical data. Grigoryan and Roe Jan [19] propose a point-based strategy to convey the uncertainty associated with isosurfaces. They combine point based rendering with other techniques such as pseudo-coloring and shading to convey trends in surface-based uncertainties.

Uncertainty visualization strategies have also been proposed for specific types of data including vector [3, 24, 32, 48, 52], tensor fields [14], and, more recently, numerical ensemble data [36, 43]. Several of these works employ glyphs [24, 48] or texture based flow [3] techniques to incorporate uncertainty into the visualization. Focusing on a global definition of uncertainty, Otto et al. [32] address the uncertainty of flow topology itself and define a process for obtaining an uncertainty-aware topological segmentation for 2D flow fields.

2.2 Visualization and Ensemble Uncertainty

Recent visualization efforts have addressed the notion of characterizing the uncertainty inherent in numerical ensembles. Potter et al. [36] present an application that addresses the uncertainty in numerical climate model ensembles. This work characterizes ensemble complexity through localized mean and variance properties. Specifically, for any spatial region \( b \in \mathbb{R}^d \) for the ensemble, there will be a mean and variance associated with the value obtained from the ensemble members. Potter et al. plot the mean and standard deviation of these values across a series of different visualizations including color maps, height fields, and iso-contour fields.

Sanayal et al. [43] also address ensemble uncertainty and focus on numerical weather ensembles. This work, like that of Potter et al. [36], also relies upon localized statistical measures. However, Sanayal et al. correctly note that values in an ensemble do not follow a normal distribution and so calculating localized mean and variances cannot proceed according to standard methods for normal distributions. To address this challenge, Sanayal et al. apply a technique referred to as bootstrapping that approximates the underlying distribution for the ensemble without requiring knowledge or assumptions regarding the true, unknown distribution. Sanayal et al. then apply their bootstrapped mean and variance values to various 2D visualization techniques such as color maps, ribbons, and glyphs, to convey ensemble uncertainty.

Our work also addresses uncertainty in numerical ensembles. However, in contrast to existing literature which quantifies uncertainty through statistical variance, this work incorporates ground truth observations into our analysis to account for both predictive error (i.e. statistical bias) and predictive variability (i.e. statistical variance). Resultantly, unlike previous methods, this work distinguishes between low variance predictions that are accurate from those that are inaccurate (consensus and accuracy are two different attributes of an ensemble). Our work is also distinct from existing literature in that we analyze ensemble data from two standpoints: we analyze the ensemble as a whole, and we analyze the value of individual members within the context of the ensemble to identify where specific constituents (and
their conceptual assumptions) help to increase predictive accuracy.

3 Method

This work presents a new method for quantifying and characterizing predictive uncertainty in numerical ensembles. The framework we present directly supports a visual segmentation strategy that highlights regions of uncertainty for numerical ensembles and the ensemble constituents. We begin by presenting the formalisms necessary to discuss our method (Section 3.1). We then introduce Bayesian Model Averaging (BMA) as an aggregation strategy for numerical ensemble data and show how BMA can be used to estimate behavior for a target event of interest (Section 3.2). We then use this estimate to quantify predictive uncertainty in a numerical ensemble and characterize this uncertainty with a two-part visualization strategy that assesses uncertainty for both the ensemble and individual constituents (Section 3.3).

3.1 Formalism

The principle task in predictive modeling is to estimate the behavior of an unknown function \( f_D : \mathbb{R}^3 \rightarrow \mathbb{R} \) that defines some processes of interest \( \mathcal{D} \) on a domain \( D \subset \mathbb{R}^3 \). Since \( f_D \) is unknown, it cannot be evaluated. This work addresses the case where \( f_D \) can be measured at a finite set of points \( S := \{ s_i | s_i \in D, i = 1, \ldots, n \} \), called the sample set. In measuring \( f_D \) at these points, we obtain the response \( y : n \times 1 \)

\[
y := < f_D(s_1), \ldots, f_D(s_n) >
\]

(1)

This case also includes a set of unique conceptual models \( C := \{ c_j | j = 1, \ldots, p \} \) that each distinctly define a likely scenario that postulates how to estimate \( f_D \). Through computer-assisted numerical simulation, each conceptual model, \( c_j \), generates a predictive response function, \( f_j : \mathbb{R}^3 \rightarrow \mathbb{R} \) on \( D \). We define \( F := \{ f_j | j = 1, \ldots, p \} \) as the collection of all predictive response functions and refer to \( F \) as the numerical ensemble.

Predictive uncertainty for \( F \) is defined through joint consideration of the mean predictive error (i.e. statistical bias) and the predictive variability (i.e. statistical variance) of the \( f \) in \( F \) [26, 44]. Thus, for ground truth observation \( y_i \) at location \( s_i \in S \), we quantify predictive uncertainty for \( F \) through the mean squared error (MSE) [20]:

\[
MSE_k(s_i) = \frac{1}{p} \sigma^2(s_i) + \eta(s_i)^2
\]

(2)

\[
= \frac{\sum_{j=1}^{p} (f_j(s_i) - \bar{f}(s_i))^2}{p} + (\bar{f}(s_i) - y_i)^2
\]

where \( \bar{f}(s_i) = \frac{\sum_{j=1}^{p} f_j(s_i)}{p} \) denotes the mean estimate, \( \sigma^2 \) the statistical variance, and \( \eta \) the statistical bias of the ensemble at point \( s_i \).

Note the quantifying predictive uncertainty in Equation 2 requires ground truth measurements, \( y_i \), to determine statistical bias (\( \eta \)). Unfortunately, due to factors associated with cost and time, ground truth for many domains is often very limited both spatially and temporally. This sparsity of ground truth often hinders the ability to quantify and characterize predictive uncertainty for analysis and visualization.

3.2 Bayesian Model Averaging

BMA addresses this sparse data challenge by combining all \( f \in F \) through a weighted coefficient, \( \beta_{BMA} : p \times 1 \) that can be used to estimate \( f_D(b) \). Importantly, this aggregation strategy also directly addresses conceptual model uncertainty: by estimating \( f_D \) through a combination of all ensemble members, the risks associated with relying on a single model to estimate \( f_D \) are avoided.

To form this aggregate, BMA forms a linear relationship between all ensemble members, \( f_j \), and the observations in \( y \). Formally, let \( x^{(j)} : n \times 1 \) be an estimate for \( y \) based on the sample space \( S \) such that \( x^{(j)} \) contains the value of \( f_j \) evaluated at all sample points \( s_i \):

\[
x^{(j)} := < f_j(s_1), \ldots, f_j(s_n) >
\]

(3)

The combination of all \( x^{(j)} \) forms the numerical ensemble estimate matrix, \( X : n \times p = [x^{(1)}, \ldots, x^{(p)}] \). This matrix, along with \( y \), defines a linear regression model

\[
y = X\beta + \varepsilon
\]

(4)

where, each \( x_{i,j} \) corresponds to the \( j^\text{th} \) constituent in the numerical ensemble, the parameter vector \( \beta : p \times 1 \) defines the unknown relationship between the ensemble’s constituents, and \( \varepsilon \) is the vector of random errors associated with the observed values in \( y \).

Aggregation strategies estimate \( f_D \) by first solving for \( \beta \), and then using this parameter vector to combine all \( f \in F \) through a weighted average. In practice, the ensemble estimate matrix \( X \) is rarely composed of sample vectors from all ensemble members, but may be constructed from a subset of ensemble member estimates. More specifically, approximations for \( \beta \) can be derived from any combination of the \( p \) estimates in \( X \). As there are \( K = 2^p - 1 \) distinct combinations of these estimates, there are (at least) the same number of \( \beta \) that may be considered for Equation 4. We let \( M \) define the space of these \( \beta \)

\[
M := \{ \beta(k) | \beta(k) \in \mathbb{R}^p, k = 1, \ldots, K \}
\]

where \( b \in \mathbb{R} \) is some non-zero weight on \( x^{(j)} \). We define \( X^{(k)} : n \times p \) as the matrix of columns that correspond to \( \beta^{(k)} \)’s non-zero components

\[
x_{i,j}^{(k)} := \begin{cases} x_{i,j}, & \text{if } k \text{’s } j^\text{th} \text{ bit is } 1 \\ 0, & \text{otherwise} \end{cases}
\]

(7)

where \( 0 : n \times 1 \) is the zero column. We denote \( m_k \) as the statistical model based on \( \beta^{(k)} \) and \( X^{(k)} \). In this context we use \( m_k \) for simplicity to refer to a given statistical model during the aggregation process (see Equations 8 – 12).

As an example of the relationship between Equations 5 through 7, consider the case where \( p = 3 \). In this example there are 7 distinct \( \beta^{(k)} \) and \( X^{(k)} \) pairs:

\[
\begin{aligned}
&k = 1 \quad \beta^{(1)} := < b_1, 0, 0 > \\
&k = 2 \quad \beta^{(2)} := < 0, b_2, 0 > \\
&k = 3 \quad \beta^{(3)} := < b_3, b_4, 0 > \\
&k = 7 \quad \beta^{(7)} := < b_{10}, b_{11}, b_{12} >
\end{aligned}
\]

\[
X^{(k)} = \begin{bmatrix}
x^{(1)} & x^{(2)} & x^{(3)} \\
\end{bmatrix}
\]

where \( b_{10}, \ldots, b_{12} \in \mathbb{R} \).

Aggregating Statistical Models

Most strategies rely on a single \( \beta \in M \) to describe the linear system in Equation 4 [9, 15, 16, 18, 47]. Unfortunately, while many of these \( \beta \) may accurately fit this system for a given \( y \) and \( X \), there is often great variability in the accuracy of these \( \beta \) when they extrapolate from \( y \) to estimate \( f_D \). Due to the inherent bias and variance of each statistical model, there is significant risk in relying on an individual \( \beta \in M \) to estimate \( f_D \) [22, 39, 50]. Analogous to how aggregation strategies address conceptual model uncertainty by combining ensemble members, BMA addresses this challenge by aggregating all \( \beta \in M \) through a weighted sum. This process generates an aggregate-based parameter vector, \( \beta_{BMA} \), that provides better overall predictive performance than any \( \beta \in M \) [22, 27, 39, 46, 50].

\[
\beta_{BMA} \text{ is based on the expected value of the coefficient terms for all } \beta \in M.
\]

The posterior mean of this set of coefficients is given by

\[
\hat{\beta}_{BMA} \text{ is based on } \beta_{BMA} \text{’s ability to balance the excessive bias and variance associated with individual } \beta \in M \text{ that impact } f_D \text{ estimates [20].}
\]
\[ \beta_{BMA} = E[\beta | y] = \sum_{k=0}^{K} E[\beta_k | y, m_k] Pr(m_k | y) \] 

where \( E[\beta_k | y, m_k] \) is the expected value of the posterior distribution of \( \beta^{(k)} \) that is weighted by the posterior probability \( Pr(m_k | y) \) (i.e. the probability that \( m_k \) is the “true” statistical model given \( y \)).

For linear regression models, the distribution of \( \beta^{(k)} \) is approximated [39]

\[ E[\beta_k | y, m_k] = \hat{\beta}^{(k)} = (X^{(k)} X^{(k)})^{-1} X^{(k)} y \] 

which is the least squares solution to \( y = \beta^{(k)} X^{(k)} + \epsilon \). The posterior probability in Equation 8 is given by

\[ Pr(m_k | y) = \frac{Pr(y | \beta^{(k)}, m_k) Pr(\beta^{(k)} | m_k)}{\sum_{i=1}^{K} Pr(y | \beta^{(k)}, m_k) Pr(\beta^{(k)} | m_k)} \]

Under the non-informative prior belief that no model \( m_k \) is preferred over any other model, the marginal distribution of the data is written

\[ Pr(y | m_k) = \int Pr(y | \beta^{(k)}, m_k) Pr(\beta^{(k)} | m_k) d\beta^{(k)} \]

where \( Pr(y | \beta^{(k)}, m_k) \) is the likelihood of \( y \) under model \( m_k \), \( Pr(\beta^{(k)} | m_k) \) is the prior density of \( \beta^{(k)} \) under model \( m_k \), and the evaluation of the integral is done through a closed form expression [39].

When \( K \) is large, there is considerable computational cost associated with evaluating Equation 8. Madigan and Raftery address this challenge through a modified greedy-search algorithm called Occam’s Window that operates on a graphical representation of the model space [27]. In this method, all models \( m_k \) are pruned from the graph if they fail to fit the data in \( y \) far less well than the best models. This pruning produces a subset of the model space \( A \subset M \). By constraining Equation 8 to \( A \) and only aggregating over models that fit \( y \) well, Madigan and Raftery both accelerate the evaluation of Equation 8 and help to markedly improve BMA’s predictive capability [22,27].

3.3 Visualizing Predictive Uncertainty

3.3.1 Visualizing Uncertainty in the Ensemble

Equation 14 quantifies predictive uncertainty for \( F \) based on the numerical ensemble’s variance, \( \sigma^2 \) and bias, \( \eta \). We characterize this uncertainty by thresholding \( \sigma^2 \) and \( \eta \) with user-specified constants: \( c_{\sigma^2} \) and \( c_{\eta} \). These constants differentiate between low and high values of variance and absolute bias (i.e. \( |\eta(b) - f_{BMA}(b)| \)). In Figure 1, \( c_{\sigma^2} \) and \( c_{\eta} \) segment all \( b \in D \) into four types of predictive uncertainty:

- **Type I - Low predictive uncertainty.** Locations \( b \in D \) where \( F \) has low variance and low absolute bias indicate regions where model estimates converge around \( f_{BMA} \) with a high degree of accuracy. The lower left quadrant in Figure 1 depicts an example of this type of behavior. These regions are often indicative of common, valid conceptualizations behind the \( f \in F \). Identifying these regions can help to isolate the conceptualizations and choices that are important to preserve in future design iterations.

- **Type II - Moderate predictive uncertainty.** Regions where \( F \) has high variance and low absolute bias indicate regions where only some of the \( f \in F \) are accurate. In these cases there is usually a lack of complete agreement across the ensemble for how to model the target event of interest. The upper left quadrant in Figure 1 depicts an example of this type of behavior. When \( F \) exhibits Type II predictive uncertainty, a key secondary step is to identify which \( f \in F \) possess more predictive accuracy. We address this task in Section 3.3.2 with a second strategy that contrasts the performance of a given \( f \) with the ensemble as a whole.

- **Type III - High predictive uncertainty with ensemble agreement.** Regions where \( F \) exhibits Type III predictive uncertainty indicate locations where all estimates converge inaccurately (i.e. low variance but high absolute bias). This type of distribution is shown in the lower right quadrant of Figure 1 and is similar to Type I predictive uncertainty. Distinguishing between Type I and Type III predictive uncertainty is important as both types indicate the possibility of common design choices for ensemble members. Unlike Type I predictive uncertainty, which indicates
common sets of valid design choices, Type III uncertainty indicates a possible common set of flaws. Differentiating these regions helps identify design flaws and accelerate tuning efforts.

- **Type IV - High predictive uncertainty with divergent estimates.** The final type of predictive uncertainty is defined by traits of high variance and high absolute bias in $F$. As shown in the top right quadrant in Figure 1, regions exhibiting Type IV predictive uncertainty for $F$ indicate regions where all models both conflict in their estimates, and almost all estimates are inaccurate. Like Type III, Type III regions represent areas where the target event’s behavior is the least understood.

We use this classification to produce a visual partition of the ensemble’s estimates. We use this partition to segment and visualize spatial regions by color which enables us to determine identify regions of high and low predictive uncertainty for the ensemble.

### 3.3.2 Visualizing Uncertainty in Ensemble Constituents

We extend Section 3.3.1 by contrasting the predictive performance of individual $f_j$ to that of the ensemble, $F$. This contrast helps identify the more accurate models in regions where the ensemble exhibits degrees of moderate or high predictive uncertainty. The goal of this approach is to classify $f_j$ into one of four cases:

- **Case I: Accurate with Consensus** The ensemble member $f_j$ is an accurate estimator and its estimate is close to the mean prediction obtained from the ensemble. This case indicates that the ensemble member is estimating in a manner that is very similar to the majority of ensemble members. The left-most image in Figure 2(a) depicts this case.

- **Case II: Accurate and Outlier** The ensemble member $f_j$ is an accurate estimator and its estimate deviates from the mean prediction obtained from the ensemble. In these cases, the conceptualizations behind $f_j$ may be more accurate than the majority of other conceptual models represented in the ensemble. This case, as seen in the right-most image in Figure 2(a), helps to identify $f_j$ that may be modeling the target event more effectively than other ensemble members.

- **Case III: Inaccurate with Consensus** The ensemble member $f_j$ is an inaccurate estimator and its estimate is close to the mean prediction obtained from the ensemble. In these cases, the estimate obtained from $f_j$ appears to be performing worse than most models. This may indicate that portions of the conceptualization in the underlying model are flawed (especially if this classification for $f_j$ is predominant for $b \in D$). The right-most image in Figure 2(b) depicts this case.

- **Case IV: Inaccurate and Outlier** The ensemble member $f_j$ is an inaccurate estimator and its estimate deviates from the mean prediction obtained from the ensemble. In these cases, the estimate obtained from $f_j$ appears to be performing worse than most models. This may indicate that portions of the conceptualization in the underlying model are flawed (especially if this classification for $f_j$ is predominant for $b \in D$).

### 3.4 Addressing Cases Where Ground Truth is Absent

During early experimental stages when ground truth is not yet available, predictive modeling tasks often rely on a base-line model (e.g. $BMA$). As shown in the left-most image in Figure 2(b), indicates that the ensemble member, $f_j$, is an outlier with poor accuracy; if this classification for $f_j$ is predominant, it can indicate that the underlying conceptual model for $f_j$ has significant flaws.

Fig. 2: This figure depicts four distinct cases that characterize a given ensemble member’s predictive performance within the context of the numerical ensemble $F$. This characterization is based on the constituent’s residual error, $\tilde{g}_j(b)$, and the ensemble’s mean residual error, $\bar{g}(b)$. In these subfigures, the distributions are based on the residual error obtained from $F$ for a given $b \in D$. In these distributions, $c_{\tilde{g}}$ indicates the window (centered around zero error) where an estimate is considered accurate (Figure 2(a)) or inaccurate (Figure 2(b)). $c_{\bar{g}}(b)$ indicates the window around the ensemble’s mean error estimate, $\bar{g}(b)$ that distinguishes between estimates that are in consensus with $\bar{g}(b)$ (e.g. distributions in the left of Figures 2(a) and 2(b)) and those that are outliers (e.g. distributions in the right of Figures 2(a) and 2(b)).

Mean predictive performance for the ensemble, $F$, is defined as

$$\bar{g}(b) = \frac{\sum_{j=1}^{p} g_j(b)}{p}$$

(16)

To characterize $g_j$ as either an outlier or following with the consensus of estimates in $F$, we contrast $g_j(b)$ to the distribution of $g$ values in Equation 16. Here $g_j$ is defined as an outlier if its value is outside one standard deviation of the estimate mean, $\sigma(g_j)(b)$. Formally, we define $\tilde{\delta}_j(b)$ as the absolute distance between $g_j(b)$ and $\bar{g}(b)$

$$\tilde{\delta}_j(b) = |g_j(b) - \bar{g}(b)|$$

(17)

If $\tilde{\delta}_j(b) > \sigma(g_j)(b)$, then $f_j$ is an outlier. Based on Equation 15 and Equation 17, our segmentation is defined in Table 1. For illustrative purposes, we also depict examples of these cases in Figure 2.
Table 1: This table shows distinct classifications of uncertainty for individual ensemble members. Membership is based on a combination of metrics: individual constituent performance ($c_g$) and consensus ($\sigma_g(b)$). Full descriptions of the categories are in Section 3.3.2.

| $|g_j(b)| \leq c_\eta$ | $|g_j(b)| > c_\eta$ |
|-------------------------|------------------|
| Case I: low error consensus | Case III: high error consensus |
| Case II: low error outlier | Case IV: high error outlier |

The slices shown in Figure 4 are estimates from two ensemble contours that characterize predictive uncertainty in a numerical ensemble. This analysis characterizes predictive uncertainty in a numerical ensemble that estimates a plume of bromine (a chemical tracer) in a bioremediation field study in Rifle, Colorado [49]. The objective of the study is to better understand the underlying subsurface so that bioremediation efforts in this area will be more effective.

To help meet this objective, a series of injection pumps (see the distribution wells in Figure 3) were used to pump a chemical tracer into the regions' ground water. For four months, the concentration of the tracer was measured weekly at 12 sample wells. Each well's measurements were made at three different depths (36 measurements made weekly). During the field study, a numerical ensemble of twenty models made estimates for the concentration of the chemical tracer throughout the region. Each constituent was based on a distinct conceptual model for the region, and all models were based on a small, irregular hexahedral grid (56 X 43 X 48 nodes).

We characterize the ensemble’s predictive uncertainty for the set of measurements collected during the last week of the field study’s third month. We begin by estimating $F_{BMA}$ (Equation 8). Ground truth for the model forms the response vector, $y$, and the ensemble’s estimates (based on 36 measurements in sample set $S$) form the matrix $X$.

The slices shown in Figure 4 are estimates from two ensemble constituents, for Models 1 and 2, and a BMA-based estimate for the tracer plume. In these figures, units are in moles of tracer per liter of water. The slices all show concentrations for a cross-section taken into the region where the first set of observation wells are approximately located (see Figure 3). The color map in this figure (far right) indicates the concentration of the tracer. In all slices, isocontours for 4.0e-3 (outmost contour), 7.5e-3 (middle contour), and 1.8e-2 (inner contour) moles per liter. In these images, BMA’s estimate is distinctly segmented into two regions. In the field study, the pumps broke for approximately one week and no tracer was injected while the pumps were down. The reduction of tracer concentration during pump maintenance is captured by well measurements. BMA’s estimate accounts for this event and may thus provide a more accurate tracer estimate.

Next we quantify and characterize the ensemble’s predictive uncertainty (Section 3.3.1). Values for maximal variance, $c_\sigma$, and absolute bias, $c_\eta$, were set at 2.2e-4 and 1.5e04 moles per liter respectively. From these thresholds we partition all estimates (and their corresponding nodes in the ensemble’s underlying grid) based on predictive uncertainty (Section 3.3.1). From this partition, we show three visualizations that characterize the ensemble’s predictive uncertainty; all images are color-coded by the colormap.

The “Scatter Plot” in Figure 5 shows the nodes for the ensemble’s underlying grid ($56 \times 40 \times 43$ nodes); here each node is plotted by its ensemble-based variance and absolute bias. The contours indicate the increasing density of nodes. The four colors in this image correspond to the types of uncertainty. In general there is a high degree of Type IV uncertainty indicating divergence in the ensemble and a great deal of error in estimates with respect to ground truth.

The “Cells” image in Figure 5 shows all nodes in the “Scatter Plot” images based on their location in the region. Each node is rendered as a single hexahedral cube and colored according the node’s classification for predictive uncertainty. The image shows a cut-away view of the region to better see the coherent patterns of the various types of predictive uncertainty. Note that three light green channels that run through the center of the region; these Type III predictive uncertainty regions indicate locations where the ensemble as a whole is estimating in a consistent, though erroneous way with respect to ground truth. Purple regions neighboring these regions indicate low predictive uncertainty. Refinement of common conceptualizations in the ensemble may help to lower the error for the Type III regions.

The “Slice” image in Figure 5 shows a cross-section through the region. As with the slices in Figure 4, this image is based on a slice made 30 meters below the subsurface. The isocontours are the contours made for the BMA-based estimate in Figure 4. Contrasting isocontours with colors for characterized predictive uncertainty, we identify that most of the higher concentration regions of the tracer correspond to regions of higher, Type III and IV predictive uncertainty.

We characterize predictive uncertainty for two of the ensembles constituents based on the methods provided in Section 3.3.2. Values for absolute bias, $c_\eta$ and absolute error, $|g_j(b)|$ (both 2.2e-4 moles per liter) segment the estimates of both models. From this partition, we construct three visualizations to characterize each model’s predictive uncertainty; all images are color-coded by the colormap.

The “Scatter Plot” images in Figure 6 help to distinctly characterize and contrast predictive uncertainty for Model 1 and Model 2. Based on density contours, estimates from Model 1 are significantly more accurate and consistent than Model 2. In Figure 6’s “Cells” images, we see estimates from Model 1 and 2 based on spatial location in the domain. Here each model’s estimate is rendered as a hexahedral cell and colored by its classification of predictive uncertainty. Note that both Models have three yellow channels of Case III predictive uncertainty similar to channels in Figure 5. These channels support previous evidence that there is a common set of erroneous estimates in the ensemble constituents. Identifying conceptual commonalities in Models...
Model 1 | Model 2 | BMA
--- | --- | ---

Concentration of chemical tracer (µmoles per liter).

Fig. 4: These slice-based images show three different estimates for a chemical tracer used in a bioremediation field study (Section 4.1): Model 1 (left), Model 2 (center), and the BMA-based estimate (right). All estimates are for the tracer’s concentration after three months of drift via aqueous transport. In the images, three concentrations of the tracer are shown as white isocontour lines: 4.0e-3 (outmost contour), 7.5e-3 (middle contour), and 1.8e-2 (inner contour) moles per liter.

Fig. 5: Ensemble Uncertainty Characterization

These three figures depict results of our method’s application to an ensemble that estimates the path of a chemical tracer in a bioremediation field study. In these figures, we analyze the ensemble’s estimates for a chemical tracer based on a three month duration for drift. All images regions are color-coded based on predictive uncertainty (Section 3.3.1): dark purple (Type I); light purple (Type II); light green (Type III); and dark green (Type IV). This characterization process is based on thresholds for $\sigma$ and $\eta$ of 1.5e-4 and 2.2e-4 respectively. The Scatter Plot view characterizes all nodes in the numerical ensemble’s underlying grid (56 X 40 X 43); nodes are plotted by ensemble std deviation and absolute bias. The Cells view shows the same nodes rendered spatially in the subsurface region; here each node is rendered as a colored hexahedral cell. The cutaway shows the interior of the ensemble’s segmentation. The Slice view shows various types of predictive uncertainty for a vertical cross-section cut into the region (see the slice in Figure 3 for reference). Isocontours correspond to the same contours in Figure 4: 4.0e-3 (outmost contour), 7.5e-3 (middle contour), and 1.8e-2 (inner contour) moles per liter.

Fig. 6: Uncertainty Characterization of Individual Models

These three figures depict results of our method’s application to two models (top and bottom row) that estimate the path of a chemical tracer in a bioremediation field study. In these figures, we analyze the estimates for a chemical tracer based on a three month duration for drift. In all images regions are colored based on predictive uncertainty classification (Section 3.3.2): blue (Case I); green (Case II); yellow (Case III); and red (Case IV). The classification is based on $\delta$ and an absolute error threshold of 2.2e-4 moles per liter. Scatter Plot views show the characterization of all nodes in the numerical the underlying grid (56 X 40 X 43 nodes) used by both models; cells are plotted $\delta$ vs. absolute error to the BMA estimate. The Cell shows the same nodes rendered spatially in the subsurface region; here each node is rendered as a colored hexahedral cell. The Slice view shows the various types of predictive uncertainty in the ensemble for a vertical cross-section cut two meters into the region (see the slice in Figure 3 for reference). Isocontours correspond to the same contours in Figure 4: 4.0e-3 (outmost contour), 7.5e-3 (middle contour), and 1.8e-2 (inner contour) moles per liter.
CCL event began after rising groundwater flooded a set of compromised containment tanks. Significant amounts of CCL were removed from containment once the ground water began to subside (Figure 7).

The field study uses a parsimonious model as a ground truth surrogate while monitoring data is collected. The parsimonious model, which is defined on a sparse, irregular grid of 5×5×5 nodes per liter respectively. From these thresholds we determine ensemble-based bias and standard deviation for each node in the dense regular grid. We partition the nodes based on their classification of predictive uncertainty (Section 3.3.1). From this partition, we construct three visualizations to characterize the ensemble’s predictive uncertainty; in contrast Model 2 is characterized as Case III.

We begin by estimating $F_{BMA}$ (Equation 8). In this task, the parsimonious model’s estimate forms the response vector, $y$, and ensemble’s estimates (based on sample set $S$) form the ensemble matrix $X$. The slices shown for Models 1 and 2 in Figure 8 are the estimates from two ensemble constituents for the plume’s concentration. The third image Figure 8 is the BMA-based estimate for the plume. The slices show all concentration at 30 meters beneath the subsurface (see Figure 7). This figure’s color map indicates CCL4 concentration in μg per liter. In all images, isococontours for 500 (outmost contour), 1000 (middle contour), and 2000 (inner contour) are shown.

Next we quantify and characterize the ensemble’s predictive uncertainty (Section 3.3.1). Values for maximal variance, $c_\eta$, and absolute bias, $c_{|\eta|}$, were set at 450 and 500 μg per liter respectively.

The “Cells” image in Figure 9 shows all nodes in the “Scatter Plot” images based on their location in the region. Regions with Type I predictive uncertainty are shown in purple. These are regions where, despite conceptual complexities, the ensemble’s estimates strongly agree with each other, and with the parsimonious model. In contrast, green regions, which indicate Type IV predictive uncertainty, are regions where the ensemble’s estimates strongly vary internally and deviate from the parsimonious model. These regions indicate where more complex assumptions for geological parameters significantly alter estimates from parsimonious assumptions for the target event. The other types of predictive uncertainty are shown in light purple (Type II) and light green (Type III).

The “Slice” image in Figure 9 shows a slice through the region. As with the slices in Figure 8, this image is based on a slice made 30 meters below the subsurface. The isocontours are the contours made for the BMA-based estimate for 500 (outmost contour), 1000 (middle contour), and 2000 (inner contour) μg per liter. By contrasting these isocontours with colors for characterized predictive uncertainty, we can identify that most, but not all, of the higher concentration regions of CCL4 (see the inner contour) correspond to regions of higher predictive uncertainty (i.e. Type III and IV).

We characterize predictive uncertainty for two of the ensemble’s constituents based on the methods provided in Section 3.3.2. In this analysis, we use values of 500 μg per liter for both absolute bias, and absolute error. Figure 10 contains image sets for Models 1 and 2 that contain images for “Scatter Plots”, “Cells”, and “Slice” techniques. The color map in this figure characterizes predictive uncertainty for each constituent-based estimate in these images.

Figure 10’s “Scatter Plot” images help to distinctly characterize and contrast predictive uncertainty for Model 1 and Model 2. Based on density contours, Model 1 has more nodes in Case II than Model 2. This trend indicates many of Model 1’s nodes consistently estimate the way the parsimonious model does when other ensemble constituents don’t (e.g. Model 2). While Model 2 also has some nodes classified as Case II, the density contours for Model 2 indicate more of its nodes are classified as Case III and IV. Such classifications indicate that many of Model 2’s estimates deviate from the parsimonious model.

The “Cells” images in Figure 10 show the spatial representation of the scatter plot distributions for Model 1 and 2. The distinct green regions for Model 1 indicate the locations where Model 1’s estimate is close to the parsimonious model’s estimate. In contrast, Model 2 has a predominance of yellow regions that indicate most of Model 2’s estimate for the plume are significantly different from the parsimonious estimate. The red regions in both models indicate regions where the models’ estimate deviate from the parsimonious model and deviate from the majority of the ensemble estimates. These regions are most likely related to the unique conceptualization of the models’ designs.

Figure 10’s “Slice” images convey similar information to the cells images; slices are based on a depth of 30 meters beneath the subsurface. The iso-contours of CCL4 concentration (500, 1000, and 2000 μg per liter) in these images help correlate uncertainty with increasing CCL4 concentration. For example, much of Model 1’s estimate within the 2000 iso-contour line is characterized as Case II or III predictive uncertainty; in contrast Model 2 is characterized as Case III.

5 Future Work

This work presents visual methods for characterizing predictive uncertainty in numerical ensembles. We demonstrate our approach using two datasets to demonstrate the broad utility of our techniques. However, there is much yet to address in this field. Our future work will look at extending the characterization presented in this paper to the time-varying aspect of these datasets. In this effort, we hope to characterize predictive uncertainty spatio-temporally by identifying how regions stay, or transition between categories of uncertainty. We also are looking to extend our methods to address multivariate data.

References

Fig. 8: These images show slices for three different estimates at approximately one year during the CCl₄ event. CCl₄ concentration ranges between 0 and 3315 µgrams per liter. Three concentrations of CCl₄ are shown in this image as isocontour lines (i.e. the white lines): 500 (outmost contour), 1000 (middle contour), and 2000 (inner contour) µgrams per liter.

Fig. 9: Ensemble Uncertainty Characterization These three figures depict results from our method (Section 3.3.1) for characterizing uncertainty and segmenting the CCl₄ ensemble. In all three images regions are color coded based on predictive uncertainty (Section 3.3.1): dark purple (Type I); light purple (Type II); light green (Type III); and dark green (Type IV). This characterization process is based on thresholds for σ and η of 450 and 500 respectively. The Scatter Plot view shows the nodes where nodes are plotted by ensemble std deviation and absolute bias. The Cells view shows the nodes rendered and colored by ensemble classification; the cutaway shows the interior of the ensemble’s segmentation. The Slice view shows the various types of predictive uncertainty in the ensemble at 30 meters beneath the subsurface. Isocontours correspond to the same contours in Figure 8 for 500 (outmost contour), 1000 (middle contour), and 2000 (inner contour) µgrams per liter.

Fig. 10: Uncertainty Characterization of Individual Models These images depict results from our method (Section 3.3.2) for characterizing and segmenting individual models based on their predictive uncertainty. In this figure, the estimates of Model 1 (top) and Model 2 (bottom) are analyzed at approximately one year during the CCl₄ event. In all images regions are colored based on predictive uncertainty (Section 3.3.2): blue (Case I); green (Case II); yellow (Case III); and red (Case IV). The classification itself is based on δ and an absolute error threshold of 500. Scatter Plot views show the characterization of all nodes; here cells are plotted δ vs. absolute error to the BMA estimate. The Cell views show the model’s estimates colored by classification; the cutaway shows the interior of each model’s segmentation. The Slice view shows the various types of predictive uncertainty in the ensemble at 30 meters beneath the subsurface. Isocontours correspond to the same contours in Figure 8 for 500 (outmost contour), 1000 (middle contour), and 2000 (inner contour) µgrams per liter.