

Visualizing Numerical Uncertainty in Climate Ensembles

by

Xing Liang

A Thesis Presented in Partial Fulfillment
of the Requirements for the Degree
Master of Science

Approved October 2016 by the
Graduate Supervisory Committee:

Ross Maciejewski, Chair
Giuseppe Mascaro
Hessam Sarjoughian

ARIZONA STATE UNIVERSITY

December 2016

ABSTRACT

The proper quantification and visualization of uncertainty requires a high level of domain knowledge. Despite this, few studies have collected and compared the roles, experiences and opinions of scientists in different types of uncertainty analysis. I address this gap by conducting two types of studies: 1) a domain characterization study with general questions for experts from various fields based on a recent literature review in ensemble analysis and visualization, and; 2) a long-term interview with domain experts focusing on specific problems and challenges in uncertainty analysis. From the domain characterization, I identified the most common metrics applied for uncertainty quantification and discussed the current visualization applications of these methods. Based on the interviews with domain experts, I characterized the background and intents of the experts when performing uncertainty analysis. This enables me to characterize domain needs that are currently underrepresented or unsupported in the literature. Finally, I developed a new framework for visualizing uncertainty in climate ensembles.

To my loved ones for all the encouragement

ACKNOWLEDGMENTS

I would like to take this opportunity to thank Professor Ross Maciejewski for his continued guidance and supervision. I am thankful to Professor Giuseppe Mascaro and Professor Hessam Sarjoughian for being my Masters Thesis committee members and for their support throughout the completion of my research.

TABLE OF CONTENTS

	Page
LIST OF TABLES	vi
LIST OF FIGURES	vii
CHAPTER	
1 INTRODUCTION	1
2 RELATED WORK	3
2.1 Uncertainty Quantification and Visualization	3
2.2 Uses of Domain Knowledge	4
3 TAXONOMY OF UNCERTAINTY QUANTIFICATION AND VISU- ALIZATION	7
3.1 Defining and Quantifying Uncertainty	8
3.2 Descriptive Statistical Approaches	12
3.3 Inferential Statistical Approaches	15
3.4 Information Theoretical Approaches	18
4 INTEGRATING DOMAIN KNOWLEDGE INTO UNCERTAINTY ANAL- YSIS	22
4.1 Interviews	22
4.2 Online Questionnaire	23
5 VISUAL ANALYTIC TOOLS	29
5.1 System Architecture	29
5.2 Data Description and Uncertainty Definition	31
5.3 Exploring Model-Level Uncertainty	32
5.4 Exploring Ensemble-Level Uncertainty	35
5.5 Global Climate Assessment Models	38
5.6 Data Processing and Implementation	41

CHAPTER	Page
6 CASE STUDIES	43
6.1 Prediction of Water Scarcity in Niger River Basin	43
6.1.1 Data Description:	43
6.1.2 Visual Analytic View.....	45
6.1.3 Case Studies.....	47
6.2 Prediction of Malaria Spread in Western Africa	49
6.2.1 Data Modeling.....	49
6.2.2 Visual Analytic View.....	51
6.2.3 Case Studies.....	52
7 CONCLUSION AND FUTURE WORK.....	54
REFERENCES	56
APPENDIX	
A ONLINE QUESTIONNAIRE.....	62
B IRB APPROVAL FORM.....	75
C RESPONSES OF ONLINE QUESTIONNAIRE.....	78

LIST OF TABLES

Table	Page
3.1 Exemplar Visualizations in Each Category of Uncertainty Quantification.	11
3.2 Common Methods in Descriptive Statistics.	13
3.3 Common Methods in Inferential Statistics.	17
3.4 Common Methods in Information Theory.	21

LIST OF FIGURES

Figure	Page
3.1 Demonstration of Similar Distributions.	15
4.1 Examples of Other Visualizations in the Questionnaire.	24
4.2 Demonstrations of Some Visualization Issues.	26
5.1 Model-View-Controller Architecture.	30
5.2 The Single-Model Exploration Tool.	32
5.3 The Scatter Plots in Single-Model Exploration Tool.	34
5.4 The Multi-Model Exploration Tool.	36
5.5 The GCAM Line Chart View.	40
6.1 The Visualization Overview for the Case Study of Water Scarcity.	45
6.2 Water Scarcity Color Mapping Schemes.	48
6.3 The Case Study Showing Severer Water Scarcity in Future.	49
6.4 The Visualization Overview for the Case Study of Malaria Spread.	51
6.5 The Case Study for the Spread of Malaria under Climate Variability. ...	53

Chapter 1

INTRODUCTION

Computational advances are enabling the generation of massive amounts of model simulation results. However the complexity and high dimensionality of the raw simulation data brings challenges to traditional uncertainty analysis. Data analysis techniques largely focus on aggregations and summarizations to identify simulation patterns from which quantitative uncertainty analysis is performed. Visualization techniques encourage data explanation; however, the choice of visual metaphor (Potter *et al.* (2012b)) is critical for displaying uncertainty.

From my literature study and collaborations with domain experts, what hinders effective analysis and visualization is a lack of tools that can couple a high level of domain knowledge with the uncertainty analysis. When a high level of domain knowledge, especially modeling knowledge, is required, most of the current visualization tools and results can only be understood by domain experts, which limits the use of such results in policy and decision making. Therefore, the gap among domain experts, analysts and stakeholders often serves as an obstruction to the effective uncertainty analysis and representation. In order to bridge the knowledge gap, many studies have focused on analyzing the visualization issues in other domains as well as the problem of integrating domain knowledge in visualization field (Goodwin *et al.* (2013); Dasgupta *et al.* (2015); MacEachren *et al.* (2012)). For example, Dasgupta *et al.* (2015) analyze a large set of static climate data visualizations for identifying defects with respect to the visualization design. At a higher level, some studies focus on how to enable human-in-the-loop analyses (Sacha *et al.* (2016); MacEachren (2015)) and avoid the pitfalls during the analysis (Sedlmair *et al.* (2012); Kwon *et al.*

(2011)).

This thesis focuses on addressing the gaps between uncertainty analysis and visualization. To this end, I have collaborated with several domain experts and conducted two types of studies: 1) a domain characterization study with general questions for domain experts from various climate related fields that utilize ensemble analysis and visualization, and; 2) a regular long-term interview with domain experts focusing on specific problems and challenges in uncertainty analysis and visualization. From the domain characterization, we identify the most common metrics applied for uncertainty quantification and discuss the application of these methods for visualization. Based on the interviews and surveys with domain experts, I characterize the background and intent of the experts performing uncertainty analysis. This enables me to characterize domain needs that are currently underrepresented or unsupported in the literature. Through collaborations with domain experts, I have developed an interactive web-based framework for visualizing numerical uncertainty in climate model ensembles. This framework not only enables uncertainty analysis at both the model-level and the ensemble-level uncertainty, but also helps users understand the uncertainty through visual comparisons among ensembles of models.

Chapter 2

RELATED WORK

In this section, I first review existing taxonomies of uncertainty quantification and visualization. Then I review work on integrating the role of domain experts in uncertainty analysis and visualization.

2.1 Uncertainty Quantification and Visualization

A variety of uncertainty metrics are defined, categorized, reframed or proposed (Buttenfield and Weibel (1988); Pang (2001); Plewe (2002); Kandlikar *et al.* (2005); Thomson *et al.* (2005); MacEachren *et al.* (2005); MacEachren (2015)). The related visualization techniques and challenges have been well documented. For example, Sanyal *et al.* (2009) compare uncertainty visualizations in 1D and 2D datasets. MacEachren (2015) applies concepts from visual semiotics to characterize the visual significance of different categories of uncertainties. Potter *et al.* (2010) focus on the use and adaptation of box plots in uncertainty visualization. Potter *et al.* (2012b) and Bonneau *et al.* (2014) summarized multiple visualizations in terms of their data dimensions and demonstrated their practical uses, and Potter *et al.* (2012b) illustrated the use of color maps for 2D uncertainty (Potter *et al.* (2012a)) and the use of glyphs, color maps, isosurfacing and volume rendering for 3D uncertainty (Potter *et al.* (2008)).

However, there are few studies characterizing how uncertainty quantification methods are understood by domain users. Potter *et al.* (2012b) argued that communicating uncertainties is a task often left to visualization without any connection between the quantification and visualization. Klir and Wierman (1999) evaluated uncertainty

from the fuzziness of the data and generalized applications of information theory for quantifying uncertainty. Thomson *et al.* (2005) expanded the typology for uncertainty from past frameworks in scientific computing and presented some basic quantitative models. Potter *et al.* (2010) reviewed a narrow set of summary statistics, from which uncertainty is represented as a single value, and later work by Potter *et al.* (2012b) focused on typical measures for two types of uncertainty: epistemic and aleatoric. Epistemic uncertainty is caused as the information we will lose due to the lack of knowledge or data. Aleatoric uncertainty is expressed as the randomness of mutable data values. Bonneau *et al.* (2014) generally reviewed three types of most utilized uncertainty theories (classical probability theory, Dempster-Shafer Theory, and possibility theory) but failed to go into deeper analysis. Therefore a more thorough taxonomy and problem diagnosis for uncertainty quantification measures is needed.

2.2 Uses of Domain Knowledge

Though there have been many endeavors in developing automatic algorithms and workflows, how to engage humans in the analysis loop and make use domain knowledge for uncertainty analysis have drawn more discussions in recent work through theoretical frameworks and practical experiments. For example, Sacha *et al.* (2016) proposed a knowledge generation model for visual analytics which proposes a pipeline for how humans' perceptual and cognitive biases influence the user's awareness of uncertainties. Sedlmair *et al.* (2012) proposed a methodological framework consisting of nine stages: learn, winnow, cast, discover, design, implement, deploy, reflect, and write. Each stage involves different levels of the participation from different parties (e.g. writers, tool builders, or project coordinators) in the analysis pipeline. Endert *et al.* (2014) argued for a shift from "human in the loop" philosophy for visual analyt-

ics to the “human is the loop” which focuses on recognizing analysts’ work processes and seamlessly fitting analytics into existing interactive processes. However, another issue in engaging humans in the analysis process is how to prevent the pitfalls from the participation of human. In the nine-stage framework proposed by Sedlmair *et al.* (2012), more than 30 pitfalls are defined based on the analysis of previous work and, more importantly, solutions to avoid the pitfalls are also outlined. Kwon *et al.* (2011) also identified some “visual analytic roadblocks” for novice users in an investigative analysis scenario.

In practice, the major approach for extracting domain knowledge is to interview domain experts directly. For example, Goodwin *et al.* (2013) carried out several workshops which invited energy analysts and modelers to investigate the requirements, design concepts and give feedback of visualizations in energy field. Dasgupta *et al.* (2015) held both in-person and teleconference meetings and three workshops to exchange knowledge in respective domains. They conducted a visualization use and design study by transforming the visualization design problems created by climate scientists to the challenges for visualization researchers. For example, when domain experts intend to compare the temporal variability of multiple models in a line chart, Dasgupta *et al.* (2015) characterized the visual clutter problem in the line chart and solved it through a series of small multiples. Following the guidelines in Sedlmair *et al.* (2012), Quinan and Meyer (2016) characterized their research problems through a series of contextual interviews. Another way of investigating visual analysis problems is to review and extract problems from cross-domain publications (Sedlmair *et al.* (2012)). From the view of domain experts, for example in climatology of atmospheric fields, Gleckler *et al.* (2008) employed various graphical tools to visualize, analyze and compare climate models, in which multiple visualization problems may be exposed. For example, when Gleckler *et al.* (2008) encoded different models within the

orientations of triangles, users had difficulty in distinguishing each model due to the visual clutter.

TAXONOMY OF UNCERTAINTY QUANTIFICATION AND VISUALIZATION

A wide array of taxonomies and typologies (e.g. Pang (2001); Sanyal *et al.* (2009); MacEachren *et al.* (2012)) focusing on uncertainty visualization have been published to help researchers in the design and development of uncertainty visualization tools. As an essential step before uncertainty visualization, especially for numerical uncertainty, quantifying the uncertainty plays an important role in delivering uncertainty information. Many taxonomies on uncertainty visualization (e.g. Thomson *et al.* (2005); Potter *et al.* (2012b); Bonneau *et al.* (2014)) have mentioned some quantification approaches. However, due to the growth of simulation results in climate research, it is increasingly necessary for domain experts to choose proper uncertainty quantification approaches under different uncertainty analysis requirements. With this need, and the lack of taxonomies centering on uncertainty quantification, I conducted a literature review by focusing on the use of uncertainty quantification approaches. The scope of this literature review is limited to the climate research and visualization fields. Furthermore, I restrict the literature review of uncertainty quantification approaches to: 1) quantification approaches that do not require reference data; 2) quantification approaches in low dimensional space; 3) quantification approaches commonly seen in the visualization domain. The first restriction is due to the fact that domain experts rarely have access to ground truth data, especially in the climate modeling field. In the second restriction, I find that the biases from high-dimensional approaches are hard for domain experts to understand in low dimensional space. In the third restriction, I focus on methods that visualization designers frequently adopt.

In this taxonomy, I categorize three expressions of uncertainty: 1) uncertainty portrayed as the data divergence, 2) uncertainty portrayed as the estimation results of unknown parameters, and; 3) uncertainty portrayed as the chaos in categorical data. I then extract three types of quantification approaches based on the above three forms and analyze their defects and visualization issues. Exemplar uncertainty visualization techniques along with their uncertainty quantification approaches are summarized in Table 3.1. While more systematic uncertainty visualization taxonomies can be found in past works (e.g. Pang (2001); Sanyal *et al.* (2009), and MacEachren *et al.* (2012)), in contrast, our main goal throughout the taxonomy is to shift the focus of uncertainty analysis from visualization to quantification and mitigate the concerns of domain experts in choosing quantification approaches.

3.1 Defining and Quantifying Uncertainty

In the climate research field, one of the earliest discussions about uncertainty and visualizing uncertainty can be found in the work of MacEachren (1992). MacEachren (1992) argued that, compared to the term data quality, uncertainty might be a better description for the data quality of geographic information. More often, uncertainty is understood as a composition of different concepts (Pang *et al.* (1997); Pang (2001); Thomson *et al.* (2005)) such as error, imprecision in measurements, accuracy, noise, non-specificity, etc. The ambiguity in defining the uncertainty is caused by its various sources and applications. For example, uncertainty can be explained as imprecision because of the limitation of instruments, or uncertainty can be explained as an accuracy issue due to data conversion or resampling. For a modeler in the climate research field, the source of uncertainty is defined with respect to the behavior of models in different scenario analyses. After generalizing the functionalities of the uncertainty analysis tools presented in the past works (e.g. Potter *et al.* (2010); Sanyal *et al.*

(2009); Potter *et al.* (2013); Xu *et al.* (2010); Chen and Jaenicke (2010)), I categorize three major expressions of uncertainty in scenario analyses:

- Portrayed by the data divergence
- Portrayed by the estimation results of unknown parameters
- Portrayed by the fuzziness in clustering or classifying data

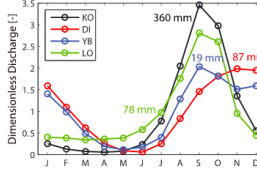
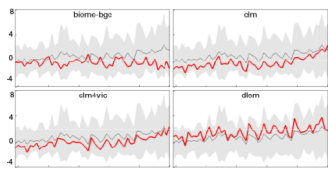
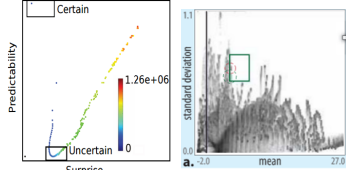
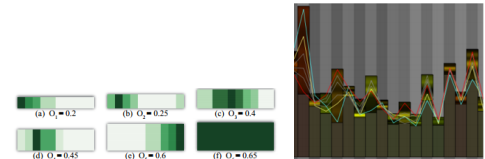
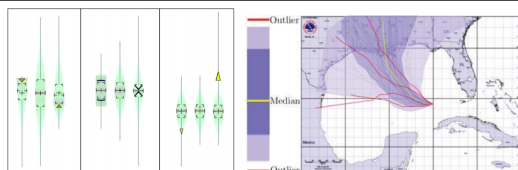
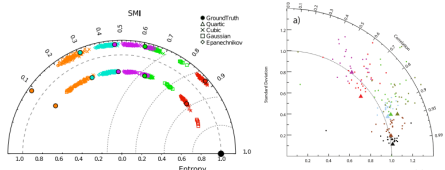
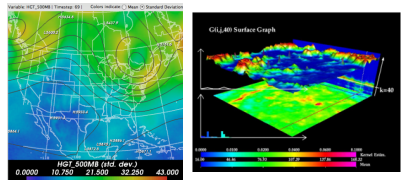
The first expression of uncertainty is often manifested as the inconsistent behaviors of models, such as the variability of the simulation results from multiple runs or from different model settings. For instance, Gleckler *et al.* (2008) defined uncertainty as the variation of the difference between the simulated results and the reference data of each model. In the second expression, one of the examples is weather forecasting, in which the estimated result is either a confidence interval or taken as the largest probability value in a probability distribution function (PDF). For example, Smith *et al.* (2009) propose a Bayesian analysis that estimates a posterior distribution of future temperature increase from multiple models, in which the uncertainty of the predicted temperature increase is expressed as probabilities in the posterior distribution. The third expression of uncertainty often occurs in clustering or classifying data. For example, when segmenting an image into different classes, we may define a membership function for each group and each pixel in the image may belong to different classes with different probabilities. In this case, the uncertainty is defined as the fuzziness among different membership functions. Potter *et al.* (2013) illustrated this problem with the segmentation of brain tissues, in which each voxel belongs to 11 different tissues with different probabilities.

In response to the above three major expressions of uncertainty, a lot of approaches are adopted in the quantification step. From our literature review, three major types of approaches are categorized:

- Descriptive statistical approaches (e.g. moments)
- Inferential statistical approaches (e.g. hypothesis test)
- Information theoretical approaches (e.g. Shannon entropy)

Descriptive statistics involve the measurements of central tendency (e.g. median and mean) and variability (e.g. standard deviation). These methods are considered to be simple and intuitive summaries about the data and are used extensively. Inferential statistics are used for estimating statistical properties from observed data, in which the estimated properties could be seen as the numerical expression of uncertainty. These methods include point estimates (e.g. estimating PDF), interval estimates (e.g. confidence interval), testing hypothesis (e.g. modality test), or clustering data points. For example, one may estimate the confidence interval, which belongs to inferential statistics, as a numerical expression of the uncertainty for an unknown parameter. For the third category, information theory has an intrinsic advantage over the above two categories, descriptive statistics and inferential statistics, in characterizing the spread of categorical data. If there are more categories in the data, each data object may belong to more categories and hence has larger uncertainty. In classification, if a data object belongs to different classes with different probabilities, uncertainty can be expressed as the disagreements of the probability values. However, it is noteworthy that these three categories are not exclusive to each other. For example, information theory requires knowledge of probability distribution functions from inferential statistics. In the following sections, each type of quantification approach will be discussed with respect to its definitions, advantages, disadvantages, and accompanying visualization approaches.

Table 3.1: Exemplar visualizations in each category of uncertainty quantification. D.S. represents the works of descriptive statistics. I.S. represents the works of inferential statistics. I.T. represents the works of information theory.

Visualization	Quantification	Exemplar Work
Line Chart	D.S.	  <p>(Mascaro <i>et al.</i> (2015); Dasgupta <i>et al.</i> (2015))</p>
Scatter Plot	D.S./I.S./I.T.	 <p>(Biswas <i>et al.</i> (2013); Kehrner <i>et al.</i> (2010))</p>
Bar Chart	D.S./I.S.	 <p>(Chen <i>et al.</i> (2015); Demir <i>et al.</i> (2014))</p>
Box Plot	D.S.	 <p>(Potter <i>et al.</i> (2010); Mirzargar <i>et al.</i> (2014))</p>
Taylor Diagram	D.S./I.T.	 <p>(Correa and Lindstrom (2013); Gleckler <i>et al.</i> (2008))</p>
Map/Original Data Space	D.S./I.S./I.T.	 <p>(Potter <i>et al.</i> (2009); Kao <i>et al.</i> (2002))</p>

3.2 Descriptive Statistical Approaches

Descriptive statistics are measures that quantitatively summarize features of a collection of information (Mann (1995)). Generally, there are two major types of measures in descriptive statistics: measures of central tendency and measures of dispersion. Measures of central tendency, which are frequently represented as mean, mode and median, describe the central value among a collection of data. Measures of dispersion, such as standard deviation, kurtosis and skewness, describe how data values are stretched. In descriptive statistics, central tendency and data divergence are used most often in climate research as measures of uncertainty. For example, Potter *et al.* (2010) present a new hybrid summary plot that incorporates a collection of descriptive statistics, including mean, standard deviation, skewness, and kurtosis, to highlight the salient features of temperature ensembles. Zehner *et al.* (2010) encode the mean and maximum deviation of weather prediction ensembles with colors and sizes for visually comparing the ensembles. More measures of the data dispersion are adopted to compare the differences between climate models. Mascaro *et al.* (2015) evaluate the spatial uncertainty in precipitation, evaporation and runoff models through the coefficient of variation. Taylor (2001) binds three measures (root mean square, covariance, and correlation coefficient) together to compare the performance of climate models. The most common descriptive statistics are listed in Table 3.2.

One of the biggest advantages of descriptive statistics is that these methods are considered to be very intuitive to end users and only require a small set of parameters (Potter *et al.* (2010)) regardless of the data complexity. This is also the reason why descriptive statistics have become the most prevalent means of summarizing data features and portraying uncertainty information in many domains. However,

Table 3.2: Common methods in descriptive statistics. x_i represents the i th observed values in a random set of values X . y_i represents the i th observed values in another random set of values Y . Q_3 and Q_1 represents the third and first quartile.

Categories	Equations
Mean	$\mu = \sum_{i=1}^N x_i / N$
Standard Deviation	$\sigma = \sqrt{\frac{1}{N} \sum_{i=1}^N (x_i - \mu)^2}$
Root Mean Square	$x_{rms} = \sqrt{\frac{1}{N} \sum_{i=1}^N x_i^2}$
Covariance	$\sigma(X, Y) = E[(X - E[X])(Y - E[Y])]$
Interquartile Range	$IQR = Q_3 - Q_1$
Coefficient of Variation	$C_v = \frac{\sigma}{\mu}$
Correlation Coefficient	$R = \frac{\frac{1}{N} \sum (x_i - \mu_x)(y_i - \mu_y)}{\sigma_x \sigma_y}$

most descriptive statistics, such as mean, standard deviation, interquartile range and coefficient of variation, are based on the assumption that data follows a Normal distribution. Only when the data distribution is Normal, do measures of central tendency make credible summaries about the data (Vogt (2011)). However, in practice, there is no empirical distribution that can precisely match the Normal distribution. Due to this problem, many works have adopted different solutions for avoiding the normality constraint. For example, Sanyal *et al.* (2010) implement an ensemble mean bootstrap (Efron and Tibshirani (1994)) which resamples the data and generates an estimated distribution with no assumptions on the types of source distribution. Kehrer *et al.* (2010) employ robust estimates of four statistical moments (mean, variance, skewness and kurtosis) and compare them with the traditional moments in an iterative visual analysis process. Another limitation of descriptive statistics is that they are too weak to differentiate some distributions within one or two values. For example, shown as Figure 3.1, three very different distributions have similar mean and standard deviation value. Bensema *et al.* (2015) and Chen *et al.* (2015) both point out this problem

by giving the example that two distributions with similar mean and standard deviation may have different modalities or other characteristics. To solve this problem, Bensema *et al.* (2015) focus on the modality test of data distributions and Chen *et al.* (2015) project the data into another space where two data distributions can be better differentiated.

To visualize the quantification results of descriptive statistics, central tendency and dispersion are often visualized together for providing users with a contextual analysis environment. When only two or three types of descriptive statistics are involved, traditional visualization methods are sufficient, such as scatter plots (Kehrer *et al.* (2010)), bar charts (Chen *et al.* (2015)), Talyor diagrams(Taylor (2001)), and line charts (Sanyal *et al.* (2010)). However, visualization becomes problematic when users want multiple descriptive statistics in one graph. One of the solutions is to encode the extra information with more visual variables. For example, Potter *et al.* (2010) present an advanced summary plot in which multiple statistics are represented as different symbols and put at different places in a density histogram. Sanyal *et al.* (2010) encode the uncertainty statistics with sizes and colors in an ellipse and compose a ribbon by connecting the ellipses on the map. Inspired by these encoding strategies, Chen *et al.* (2015) take the sum of the standard deviation as the overall uncertainty and visualize these sums within a discrete color bar chart. Another possible solution is to combine multiple visualization charts into one chart. Demir *et al.* (2014) present a new chart which combines bar charts and line charts together to show multiple summary statistics. Mirzargar *et al.* (2014) plot the curve box plot onto the map to show the uncertainty in weather forecasting ensembles.

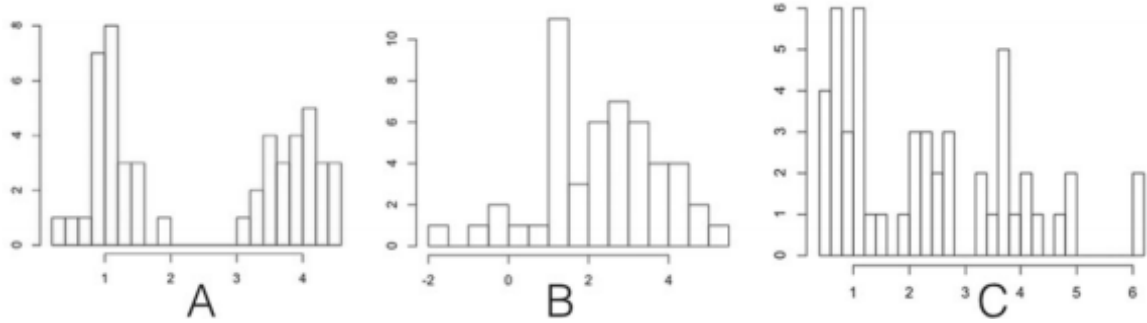


Figure 3.1: All three distributions (A, B, C) have similar mean and standard deviation values, but different modalities. Distribution A is bimodal. Distribution B is unimodal. Distribution C is multimodal (Bensema *et al.* (2015)).

3.3 Inferential Statistical Approaches

Estimating statistical properties from observed data is common in ensemble modeling. Smith *et al.* (2009) propose a Bayesian analysis that estimates a posterior distribution of future temperature increase from multiple models. Adamowski (2008) predict the peak daily water demand by different statistical models, such as linear regressions and artificial neural networks. In such cases, the estimated properties are often seen as the numerical expression of uncertainty. If a parameter is unknown, domain experts may use the confidence interval to show the possible range of that parameter, in this example, the confidence interval is defined as uncertainty. Also, when domain experts conduct parameter estimations in predictive analyses, the posterior distribution of the parameter can be portrayed as the prediction uncertainty. In such cases, inferential statistics are often adopted over descriptive statistics. The reason is that, in contrast to descriptive statistics, inferential statistics can infer new statistical properties, such as PDF, from a larger population in which the observed data is assumed to be sampled from the larger population. From our literature review in the climate research field, these methods are frequently seen in:

- Deriving estimates, including point estimates and interval estimates

- Testing hypotheses
- Clustering or classifying data objects into groups

Deriving estimates can be divided into point estimates and interval estimates (See Neyman (1937) for further explanation). Typically, point estimates refer to the process of estimating a parameter from a probability distribution. For example, Smith *et al.* (2009) quantify the uncertainty of climate model projections as a Bayesian posterior distribution in which the estimation of the posterior distribution belongs to the point estimates. Interval estimates refer to the process of finding the possible range of an unknown parameter. For example, Sanyal *et al.* (2010) portray uncertainty using the width of the 95% confidence interval in which the confidence interval is a result of interval estimates. For the second form, testing hypotheses refer to the process of checking if a formulated hypothesis should be accepted or not, which often involves a p-value or significance level test. For instance, Bensema *et al.* (2015) use Hartigan's dip test to test the unimodality of the data, in which the p-values are used for validating hypotheses. Clustering or classifying data objects is often applied on large ensembles of climate models. For example, twelve seasonal forecasting models are clustered by Yuan and Wood (2012). Other typical methods and related work in different types of inferential statistics are listed in Table 3.3.

There are multiple advantages in using different inferential statistics. First, when using the confidence interval in interval estimates, the uncertainty is explained as a bounded range, which is more understandable than a single value (e.g. standard deviation)(Potter *et al.* (2012b)). Second, when selecting the best model or parameter using Bayesian inference in point estimates, the uncertainty could be described as a probability value, which is more straightforward in decision-making (Spiegelhalter *et al.* (2011)). However, in order to infer the statistical properties beyond the ob-

Table 3.3: Common methods in each type of inferential statistics.

Categories	Measures used in Visualization
Point Estimates	Bayesian Inference (Kniss <i>et al.</i> (2005); Smith <i>et al.</i> (2009); Saad <i>et al.</i> (2010); Gosink <i>et al.</i> (2013)), Maximum Likelihood (Najafi and Moradkhani (2015))
Interval Estimates	Confidence Interval (Sanyal <i>et al.</i> (2010))
Testing Hypotheses	Dip Test (Bensema <i>et al.</i> (2015))
Clustering or Classification	Hierarchical clustering (Yuan and Wood (2012))

served data, most of the inferential statistics, such as point estimates, will need to estimate the probability distribution function. From our literature review, I find both parametric and non-parametric methods are used in density estimations. Among the non-parametric methods, kernel density estimation (KDE) is frequently used (Hall and Manabe (1997); Feng *et al.* (2010); Maciejewski *et al.* (2010); Chen *et al.* (2015)). The advantage of KDE is that it only requires configuring two values: the kernel function and the bandwidth. Particularly, the choice of the bandwidth is critical to the bias-variance trade-off. If the bandwidth is too large, the bias will be large, especially for heavily tailed data. However, if the bandwidth is too small, the bias is small but the variance is large. To avoid such issues, many works (e.g. Chen *et al.* (2015); Pöthkow and Hege (2013)) choose automated selection methods. For parametric methods, the choice of the parametric model is dependent on analysts' knowledge of the data. A typical example for such a parametric model is the Normal model where only two parameters, $\vec{\theta} = (\mu, \sigma^2)$, need to be estimated. The methods of estimating the parameters of the assumed parametric model can be maximum likelihood (e.g. Najafi and Moradkhani (2015)), Bayesian estimation (e.g. Gosink

et al. (2013)) or expectation maximization (e.g. Liu *et al.* (2012)). As opposed to the descriptive statistics, inferential statistics need more knowledge of the data and more expertise in statistics.

Visualization of uncertainty information from inferential statistics is also more complex than descriptive statistics. While the uncertainty is typically expressed as a PDF, few visualization approaches can directly visualize the PDF of each point in the original data space (Potter *et al.* (2012b)). More often, visualization designers will summarize a few features from the PDF of each pixel within a single value or an interval to mitigate such visualization issue. For example, rather than showing the whole PDF of each point on the map, Bensema *et al.* (2015) categorize the PDF into three types of modality (unimodal, bimodal and multimodal) and map each modality with a unique color. Zehner *et al.* (2010) extract the major divergence of the PDF as an interval, and then visualize it as a small line segment over the data point. However, this solution can often misrepresent the characteristics of the actual data (Potter *et al.* (2012b)). At the cost of space, another solution is to plot more properties of the PDF in another graph and then link the graph with the original data space. For example, Gosink *et al.* (2013) visualize the bias and variance of the PDF in a scatter plot that maps the color to the original data space.

3.4 Information Theoretical Approaches

Categorical output (e.g. types of land uses) is very common in the outputs of climate model simulations. To describe the uncertainty in these data, information theoretical methods have an intrinsic advantage over descriptive statistics and inferential statistics. From the literature review, as shown in Table 3.4, Shannon entropy and mutual information are broadly applied in climate research, such as climate model ensembles (Correa and Lindstrom (2013)), multi-dimensional data (Chen *et al.* (2015)),

and 2D flow data (Xu *et al.* (2010)). Shannon entropy quantifies the amount of uncertainty in a set of random variables. Larger Shannon entropy values indicate larger uncertainty in the data. Mutual information is a measure of quantifying the shared information between two variables, in which larger values indicate two variables are more similar.

The great advantage provided by information theory lies in the discovery that there is a unique and unambiguous criterion for the amount of uncertainty represented by a discrete probability distribution (Jaynes (1957)). It is very useful in comparing uncertainties of different sets of models. Typically, the unit of the information is decided by the number of categories in the data. However sometimes only part of the categories will show up in some data. Therefore, in order to unify the unit of the Shannon entropy globally, domain experts may need to find all possible categories by scanning the whole data first. From this, domain experts can also infer the maximum Shannon entropy if every data value is unique, and the minimum Shannon entropy can be inferred if all data values are the same. To apply information theory to continuous data, a typical strategy is to discretize the numerical data with a set of bins. For example, Chen *et al.* (2015) discretize the numerical data into 256 bins before computing the relative entropy. Other concerns in information theory focus on the clarification of information and uncertainty. Biswas *et al.* (2013) have discussed their differences and usages through the generation of streamlines in flow visualization. For example, for a visualization resulting in visual clutter, visualization users will obtain less information and be faced with larger uncertainty. But for the visualized subject, its original intention is that adding more visual variables may deliver more information and eliminate the uncertainty from misunderstanding.

In the accompanying visualization approaches, quantification results of information theoretical measures are either encoded as visual representations to portray the

uncertainty or taken as the indicators for generating a more effective visualization result. In portraying uncertainty, one common method is visualizing the quantification results in the original data space, such as maps, with values encoded by color lightness, hues or saturations. Van der Wel *et al.* (1998) compute the entropy values on a set of remote sensing data and encode them within different saturations of gray. Potter *et al.* (2013) compute the entropy values of classified brain tissues and encode them with different color lightness. From another perspective, Correa and Lindstrom (2013) built a new coordinate system, namely Taylor diagrams, by taking advantage of the triangle equality among Shannon entropy, mutual information and the variation of information. The axes in the Taylor diagram, shown in Table 3.1, represent the three measures and each dot represents a model. When taking the information theory methods as the indicators for generating a more effective visualization result, the visualization results are often updated in an iterative process. For example, Xu *et al.* (2010) and Ji and Shen (2006) evaluate the communication effectiveness through the information entropy in the visualization results. If a visualization result involves more data categories, that is larger entropy values, the visualization result is regarded as being more effective. Therefore, the visualization can be improved by iteratively looking for larger entropy values. Xu *et al.* (2010) also demonstrated that a flow visualization with highlighted features has a larger entropy value. Likewise, Ji and Shen (2006); Wang and Shen (2011) conclude some similar uses of information theory in scientific visualization, such as the selection of viewpoints. A better view point is defined as larger entropy in the visualization result.

Table 3.4: Common methods in information theory. $p(x, y)$ is the joint probability distribution function of variable X and Y. $p(x)$ and $p(y)$ are the marginal probability distribution function of X and Y respectively.

Categories	Equations
Shannon Entropy	$H(X) = - \sum_X p(x) \log_2 p(x)$
Mutual Information	$I(X; Y) = \sum_Y \sum_X p(x, y) \log_2 \left(\frac{p(x, y)}{p(x)p(y)} \right)$
Variation of Information	$VI(X; Y) = H(X) + H(Y) - 2I(X, Y)$

Chapter 4

INTEGRATING DOMAIN KNOWLEDGE INTO UNCERTAINTY ANALYSIS

The importance of integrating domain knowledge into uncertainty analysis has been broadly discussed (Kwon *et al.* (2011); Sedlmair *et al.* (2012); Goodwin *et al.* (2013); Endert *et al.* (2014); Dasgupta *et al.* (2015); MacEachren (2015); Sacha *et al.* (2016)). Previous works described three major approaches to extract domain knowledge: literature reviews, interviews and questionnaires. In the previous chapter, I have conducted a literature review and summarized the work into a taxonomy. In this chapter, I will introduce a long-term interview and an online questionnaire from which I characterize domain knowledge.

4.1 Interviews

The long-term interview occurred over one and a half years with weekly face-to-face and email meetings. The goal of the interview is to thoroughly investigate and elicit the role of domain experts in uncertainty analysis. Participants of the meetings include two parties: domain experts in hydrology and domain experts in visual analytics. Domain experts provide the source data, abstracted domain problems and domain insights for the new tools. Visual analytics experts convert the domain problems into visual analytical problems and look for solutions in the visualization field. Though the participants changed over time, which refers to pitfall three in Sedlmair *et al.* (2012), roles of each party were continued. In the regular interview, the continuous participation of each party helps prevent any disjointing from the research cycle. Also, the direct engagement of experts helps maximally eliminates tool builders' recognition biases during the implementation. To improve the structure

of the long-term interview, the meetings were composed of multiple small iterative research cycles and generally followed the structure of the nine-stage framework proposed by Sedlmair *et al.* (2012) which include learn, winnow, cast, discover, design, implement, deploy, reflect, and write.

4.2 Online Questionnaire

In reviewing the knowledge gaps between domain experts and visualization experts, most of the domain experts did not have systematic principles in uncertainty analysis and often failed to recognize their roles (Sedlmair *et al.* (2012)). With respect to this problem, I have conducted a broad survey through an online questionnaire (See the Appendix II for the IRB approval form). As shown in Appendix I, the questionnaire is composed of four parts (Background, Visualization, Quantification and Analysis) and 30 questions. Most of the questions are more general than the discussions in the long-term interview. The goal of the questionnaire is to collect details about domain uses of uncertainty quantification and visualization problems that the experts commonly encountered. The targeted participants of the questionnaire are domain experts with strong expertise in climate research. Eight participants responded to our invitation and only one of them submitted an incomplete questionnaire.

Background: Out of eight participants, five came from hydrology and the other three came from the emissions field, the integrated assessment modeling field, and the land use field. Six of the participants have related experience in uncertainty analysis. Some of them shared their concerns in uncertainty analysis. For example, one of them said that they typically did uncertainty analysis through scenario analysis rather than formal quantifications. Another said that they do not know how to properly visualize uncertainty. When asking their goals in uncertainty analysis, most of respondents replied with supporting the decision making process and providing

model comparisons.

Visualization of Uncertainty: Typically, the visualization methods are closely tied to specific types of data. The goal of this part is to investigate how domain experts use visualization methods for data analysis and exploration, and then elicit what kind of visualization problems they have encountered. In the fields of the participants, most of their data are related to climate, hydrology, gas emissions and energy, and the dimensions range from one dimensional data to multi-dimensional data. They were given different types of visualization methods, such as coloring schemes, visualization charts, visual variables, and then were asked to select visualization problems they have encountered.

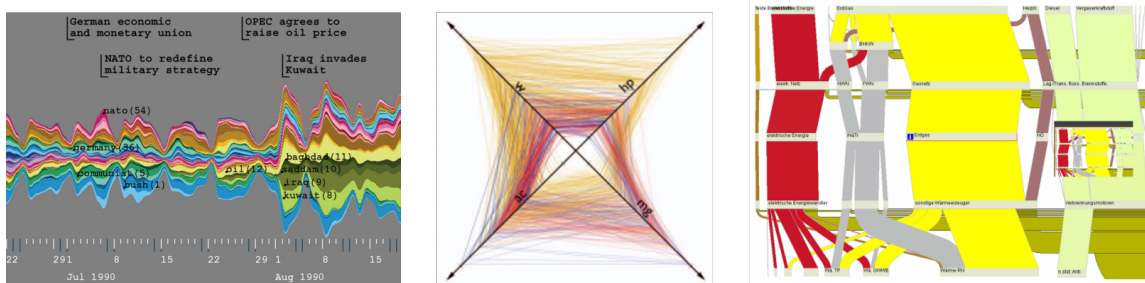


Figure 4.1: Visualization examples for the visualization methods specified by the respondents of the questionnaire. From left to right, it shows the River Flow visualization (Havre *et al.* (2002)), radar charts (Claessen and Van Wijk (2011)) and Sankey visualization (Riehmman *et al.* (2005)).

Most of respondents has clear goals and principles in selecting proper color schemes (sequential, divergent, and qualitative). From the responses, the qualitative color scheme is used to differentiate the categorical variables, the divergent color scheme is used to reflect the data fluctuations, such as the Normal distribution and changes from negative to positive values, and the sequential color scheme is only used for strictly positive (or negative) data and showing the magnitude. Traditional visualization charts, including line charts, bar charts, scatter plots and box plots, are used most often among the respondents. Another popular visualization method is to directly

visualize the data in its original data space, such as on a geographical map. However, Taylor diagrams, shown in Table 3.4, have never been used by any respondents. From this, one may infer that visualization improvements on traditional visualization charts, such as line charts, might be more helpful due to their frequent usages. One of the respondents also specified many other visualization methods (shown as Figure 4.1) outside the given list, such as River flow, Sankey visualization, radar charts, etc. Eleven visual variables (see as the Appendix I) were given to the participants and we found sizes and hues are used most frequently in their experiences. However, it is noteworthy that the orientation, grain and arrangement were never used among any respondents.

Participants were also asked to select the visualization issues they encountered. The goal of these questions were to identify which visualization problems in other domains can be avoided by adapting more advanced visualization techniques, and which problems can be mitigated by strengthening the collaborations between different parties. First, the respondents were asked to choose the visualization problems they have encountered. The top two problems in the response include visual clutter due to overlapping or color mixing and comparison complexity due to superposition overload. Stemming from these issues, a complicated legend involving too many symbols becomes another problem for experts. Typically, the major cause of these problems come from the overwhelming information to visualize. Possible solutions to mitigate visualization problems include using different levels of details and focus + context techniques. The least voted visualization problem is the comparison complexity issue due to lack of explicit encoding schemes or missing annotations. This problem is typically caused by the knowledge gaps between two domains. For example, while one may encode the dissimilarity among a set of models as the Euclidean distance in a 2D space, the convention in climate research is to encode the dissimilarity with colors

so that they can compare the data values. In this case, we assume that the participants had close collaborations with visualization designers and hence the knowledge gap across domains was mitigated. Participants also indicated some other visualization problems based on their experience, part of which are reproduced in Figure 4.2. For instance, the scatter plot in Figure 4.2, one respondent said that the large values mapped with more salient colors in the scatter plot dominate the visualization effect, which will distract people’s attention from the uncertainty area in the middle of the plot. Another example is shown in the map in Figure 4.2 where each data interval is too hard to be distinguished if too many colors hues are used .

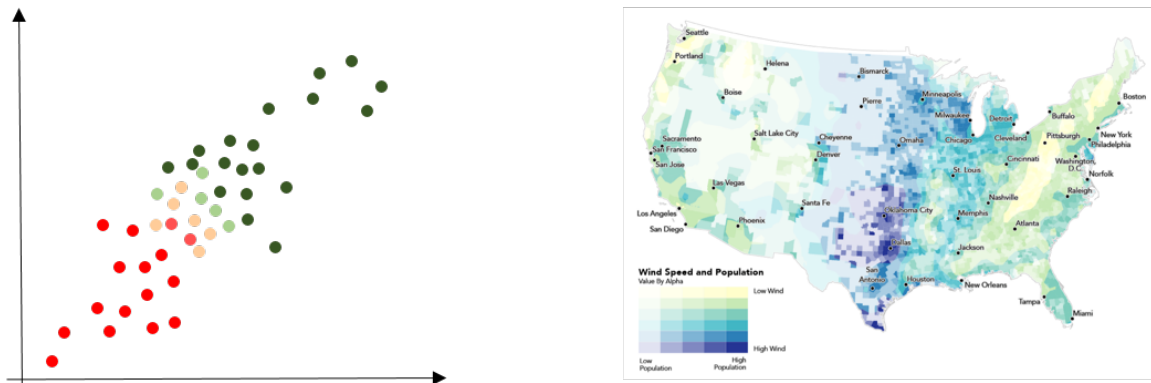


Figure 4.2: Visualization issues identified by the respondents. The left one shows the issue that people will easily get distracted by the data points with larger values or salient colors. The right one shows the issue that people can hardly distinguish each data interval when too many color hues are used.

Quantification of Uncertainty: Another gap located between the visualization domain and other domains is the use and understanding of uncertainty quantification methods. The goal of these questions is to better understand what, and how, quantification methods are used in climate research.

First, the participants were asked to select the quantification approaches they used from the given list (See as Appendix I). They were also encouraged to specify methods outside this list. Traditional methods, such as mean, standard deviation, and

confidence interval, were chosen by every respondent. However, metrics in information theory, such as Shannon entropy, were not chosen by any respondent. One of the respondents specified the probability of detection method which is not included in the list. In general, we conclude that, compared to the variety of visualization methods, the quantification of uncertainty is more limited to traditional methods.

The respondents were also asked to describe their concerns regarding quantification errors. For example, one of them said that the mean and median might be misleading for some highly tailed data. In the questionnaire, we have listed some common pitfalls (See Appendix I) and asked the participants if they were aware of those pitfalls in their previous uncertainty research. From their responses, most of them were aware of the pitfalls. Another noteworthy issue in quantifying uncertainty is to find the boundary of uncertainty. Besides describing the uncertainty within a single value, it is more reasonable to know what the maximum and minimum uncertainty is. However, two of the respondents did not realize its importance.

Visualization and Quantification of Uncertainty: We also wanted to elicit more personal opinions towards uncertainty analysis, from which we may envision future directions in uncertainty visualization and quantification. First, an interesting result is that over half of the respondents think some of the results cannot be verified by their domain knowledge. From our analysis, we may attribute this issue to the use of unfamiliar methods or limited knowledge of the data. Secondly, many respondents prefer to limit the uncertainty analysis within their own research fields instead of showing to the stakeholders. Thirdly, when they are asked about what gaps will obstruct their understanding of uncertainty, most of them select the visualization gap and knowledge gap. The visualization gap happens because some visualization methods are rarely used in their own domains and require effort to learn. The knowledge gap happens because of disagreements in domain conventions. Lastly, when they are

asked that if they realize the uncertainty propagation during the quantification and visualization stage, four out of six respondents chose No. From all these discussions, we can see the gap of uncertainty analysis and visualization between visualization designers and domain experts in other domains.

VISUAL ANALYTIC TOOLS

Based on our findings from the regular interviews with domain experts, I defined the uncertainty in our data into two levels: model-level uncertainty and ensemble-level uncertainty. For each type of uncertainty, I developed an interactive web-based tool which aims to assist domain experts in understanding both types of uncertainty as well as building the awareness of the differences among uncertainty quantification approaches. Shown in Figure 5.2 and Figure 5.4, the framework is composed of two tools. The first tool (See Figure 5.2) enables the visualization and exploration of model-level uncertainty for each model. The second tool (See Figure 5.4) enables the ensemble comparison and uncertainty analysis.

5.1 System Architecture

The system is developed based on the Model-View-Controller (MVC) architecture. A general model-view-controller architecture is composed of three parts: the model that manages the data, logic and rules of the systems, a view that shows the output of the model, and the controller that accepts the requests from users and sends them to the model for processing the requests. The following diagram (See as 5.1) shows the MVC architecture of our tools.

On the model side, we have built two servers using J2EE and Node to manage the climate model data and process the requests from controllers. As the main server, Node server supports the communication between clients and servers, such as sending HTML and JavaScript files and receiving parameter settings from clients. Other work, such as multithreaded rendering of geospatial data, are separately assigned to J2EE.

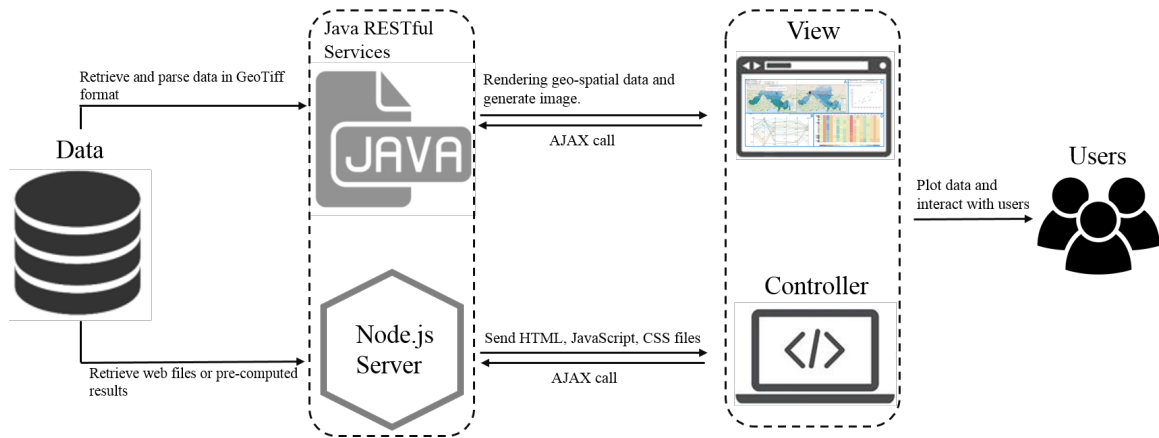


Figure 5.1: The system is designed based on the Model-View-Controller (MVC) architecture. The model part is composed of two servers, Java and Node.js, which retrieve and process the data in response to the requests sent from the controller. The view shows the output of the model through multiple visualization methods. The control panel is implemented through JQuery and sends requests to the model using AJAX.

On the view side, the output of the model is visualized using D3.js, and includes spatial maps, scatter plots, matrix views, and parallel coordinate plots. As the most important part of our work, the view part will be further explained in the following sections.

On the controller side, we have implemented a parameter control panel using JQuery which contains multiple select boxes, time sliders, buttons, etc. For each visualization tool, the controller also supports multiple interactions and animations, such as zooming in or zooming out on the map view. Many REST (Representational State Transfer) services, such as retrieving water supply data or computing mean values, were developed on the model side to respond and fulfill the request sent from the controller side. Major programming languages used for implementing the system include Java and Javascript. On the Java side, we use Jersey to implement the REST web services. On the JavaScript part, we use Node.js as the server and implement the view part and controller part with D3.js and JQuery. Other open source libraries, such as OpenStreetMap and Leaflet.js, were also used.

5.2 Data Description and Uncertainty Definition

The climate data we used in this tool contains three parts: precipitation, minimum temperature, and maximum temperature. For each type of data, it was estimated from outputs of different combinations of Global (GCMs) and Regional (RCMs) Climate Models. These were provided by the Coordinated Regional Climate Downscaling Experiment (CORDEX), a project sponsored by the World Climate Research Program that uses a set of advanced RCMs to dynamically downscale the latest set of GCM climate scenarios and predictions produced within the 5th Coupled Model Inter-comparison Project (CMIP5) (Giorgi and Asrar (2009)). The GCM-RCM combinations of CORDEX were run in a historical period from 1950 to 2005 and for future climate projections from 2006 to 2100 under the newly developed Representative Concentration Pathways (RCPs) (Vuuren *et al.* (2011)). Specifically, each pixel contains a three-dimensional distribution in which the Z axis represents simulated values, X axis represents time, and Y axis represents models.

Due to the complexity of the data, we define the uncertainty in the data into two levels:

- **Model-level uncertainty** refers to the disagreements in mutable simulation results from multiple runs of a single model. Usually, this type of uncertainty is expressed as a 2D PDF in which the x axis represents the runs and y axis represents the mutable results. It can be quantified as a single value by measures of data divergence, such as standard deviation or interquartile range, where traditional visualization methods (e.g. line charts or scatter plots) can be easily applied.
- **Ensemble-level uncertainty** refers to the disagreements of multiple models in which each model contains its own model-level uncertainty. Therefore, this type

of uncertainty is often expressed as a three-dimensional distribution in which the Z axis represents simulated values, X axis represents time, and Y axis represents models. To quantify such uncertainty, a common method in visualization and climate research is to convert the three-dimensional PDF into a 2D PDF by summarizing the model-level uncertainty of each model into a single value (e.g. mean or entropy). Therefore, visualization designers can apply the visualization methods from model-level uncertainty on ensemble-level uncertainty.

5.3 Exploring Model-Level Uncertainty

The goal of our model-level uncertainty visual analytic tools is to help domain experts explore the model-level uncertainty inside each model. This tool is composed of four parts (See Figure 5.2). Part A is only used for visualization purposes. Part B, C, and D are control panels for changing the visualization scheme.

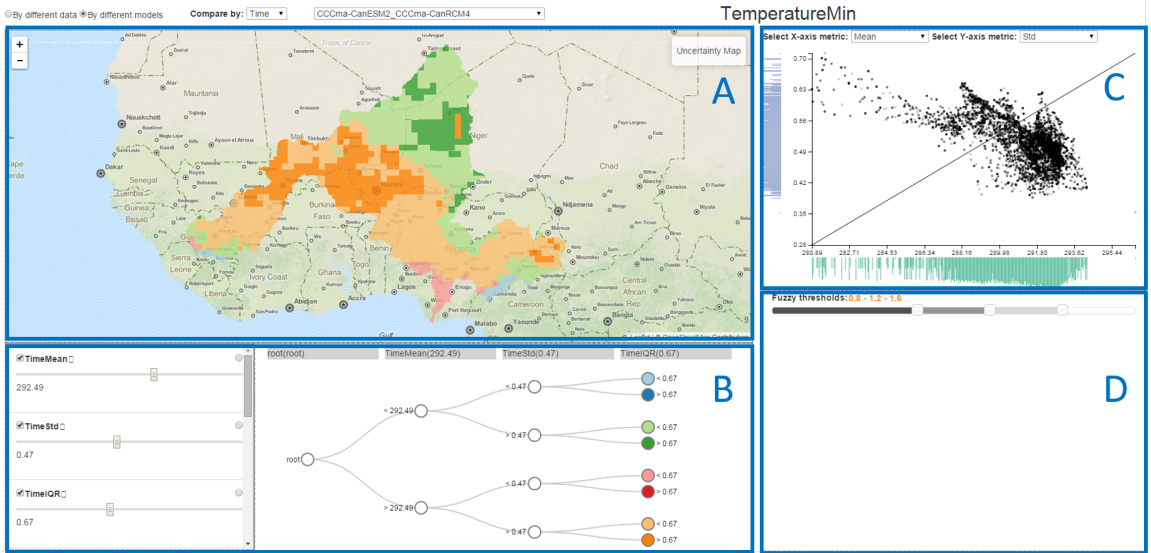


Figure 5.2: The tool for exploring the model-level uncertainty of a single climate model. (A) shows the spatial distribution of the model-level uncertainty. The color mapping scheme is decided by the tree in part(B). Each layer in the tree, except the root node, represents one uncertainty metric selected from the list on the left. In the scatter plot of part(C), each dot represents a spatial point on the map. The x-axis and y-axis can be any uncertainty metric. The slider in part(D) colors the data distribution within three discrete color scales.

Part C is a scatter plot of all spatial points in which X axis and Y axis can be selected as any uncertainty quantification method. Dots with higher opacity represent more overlapped points. Through this view, users can explore the relationship between two quantification approaches. As shown in Figure 5.3, it shows four combinations between mean and another quantification approach. We can see that as the mean value becomes larger, the entropy, standard deviation, kurtosis and skewness stretch out further. The histogram beside the axis shows the data distribution computed by that approach. For example, in Figure 5.3, the histogram below the X-axis represents the data distribution of mean values. Part D is a range slider which can map the data distribution within three gray scales.

In order to compare more quantification approaches, users can use the tree-like structure in Part B. The left list gives a pool of quantification approaches where users can choose as many methods as they want. The right tree divides the spatial points into several groups. Each tier in the tree, except the root tier, represents one quantification method, and each node represents a set of spatial points of which the quantification results are in a specific range. For example in Figure 5.2, the upper node in the second tier from left to right represents the points with their mean values larger than 292.49. Likewise, the uppermost node in the third tier represents the points with mean larger than 292.49 and standard deviation smaller than 0.47. Till the colored leaf nodes, each node represents the spatial points of which the quantification results match all the conditions from each tier. For example, the darker orange leaf node represents the spatial points with a model-level IQR values larger than the threshold 0.67, SD values larger than 0.47 and mean values larger than 292.49. Through the combination of conditional filters at each tier, we find that some regions with the same values in one method could be very different in another similar method, which illustrates the differences of these methods. For instance, even though

standard deviation and IQR are both used for quantifying the data divergence, the contrast between darker green regions and lighter green regions demonstrate that they are very different in IQR values. In order to better support the visual comparison, we use a hierarchical color scheme: 1) different color tones (warm or cold) denote the differences among mean values; 2) different color hues in the same color tone denote the differences among standard deviation values; and 3) different color saturations in the same color hues denote differences among IQR values.

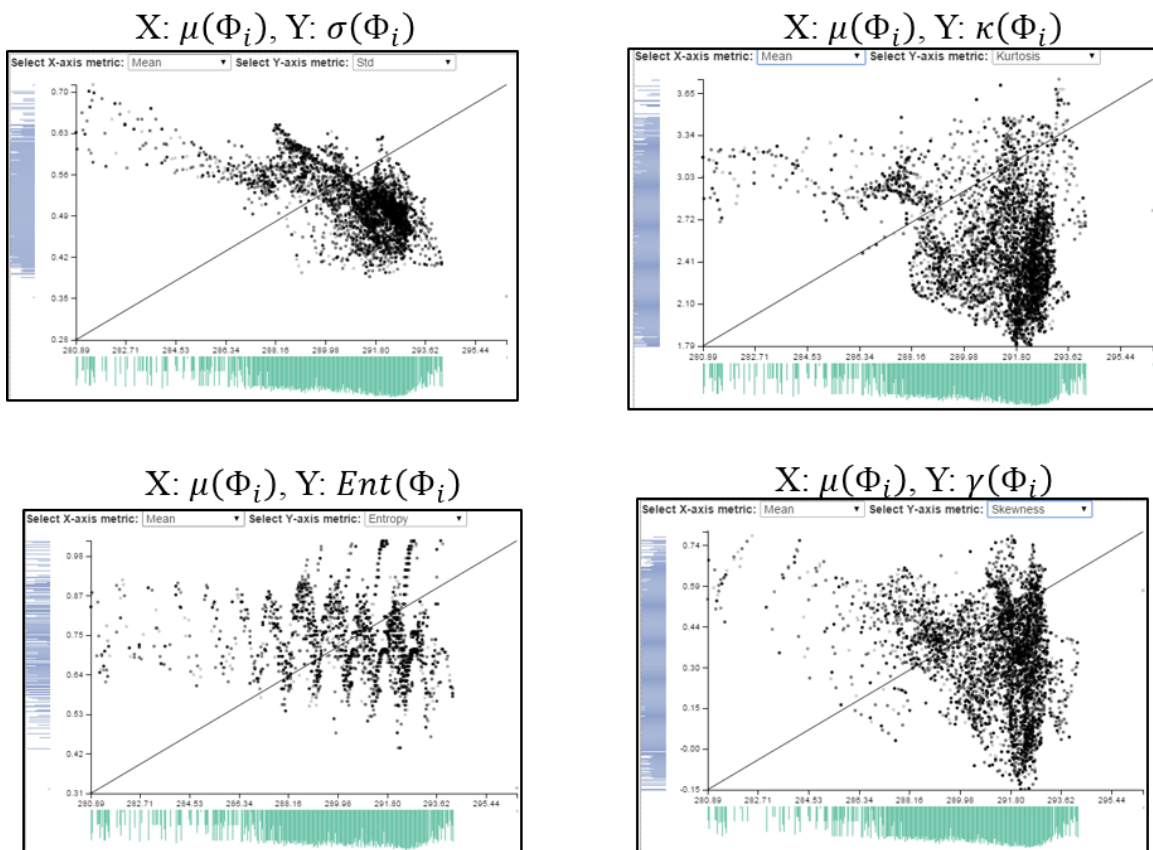


Figure 5.3: For each scatter plot in clockwise order, X axis represents the mean value over all time slices and Y axis represents standard deviation, kurtosis, entropy, and skewness respectively. We can see that as the mean value becomes larger, the entropy, standard deviation, kurtosis and skewness stretch out further.

In conclusion, there are three levels of functionalities in this tool. The fundamental functionality, shown as Part D, is to support mapping each data distribution

within three discrete colors. The middle-level functionality, shown as part C, is to compare the relationship between two quantification approaches as well as their data distributions. The high-level functionality, shown as Part B, is to compare more quantification approaches and visualize their differences.

5.4 Exploring Ensemble-Level Uncertainty

The second tool, shown as Figure 5.4, is used for comparing ensemble-level uncertainty using different quantification approaches. In general, Part A shows the overall spatial distribution of ensemble-level uncertainty and other parts are used for looking into the uncertainty in a specific group of points which can be selected and extracted through Part A and Part B.

To clarify the confusion between ensemble-level and model-level uncertainty quantification in this tool, we give following definitions and notations. A model-level distribution represents the time series in one single model. Based on this definition, for any model M_i , where $i = 1, 2, 3, \dots, 18$, the feature vector of model M_i at a geographical point (x, y) is composed as Equation 5.1.

$$\vec{M}_i(x, y) = [\mu_i(\cdot), \sigma_i(\cdot), IQR_i(\cdot), \dots, Kurt_i(\cdot)]^\top \quad (5.1)$$

where $\mu(\cdot), \sigma(\cdot), IQR(\cdot), Kurt(\cdot)$ represents the features we extract from the model-level distribution using some quantification approaches, such as mean, standard deviation, and IQR. Similarly, the ensemble-level uncertainty is computed based on the model-level features of each model, which can be represented as Equation 5.2:

$$Ens \left\langle \vec{M}_1(x, y), \vec{M}_2(x, y), \dots, \vec{M}_i(x, y), \dots, \vec{M}_n(x, y) \right\rangle \quad (5.2)$$

where n represents the number of models, r represents the row index of the feature vector $\vec{M}_i(x, y)$, and $Ens \langle \cdot \rangle$ represents quantification methods for ensemble-level un-

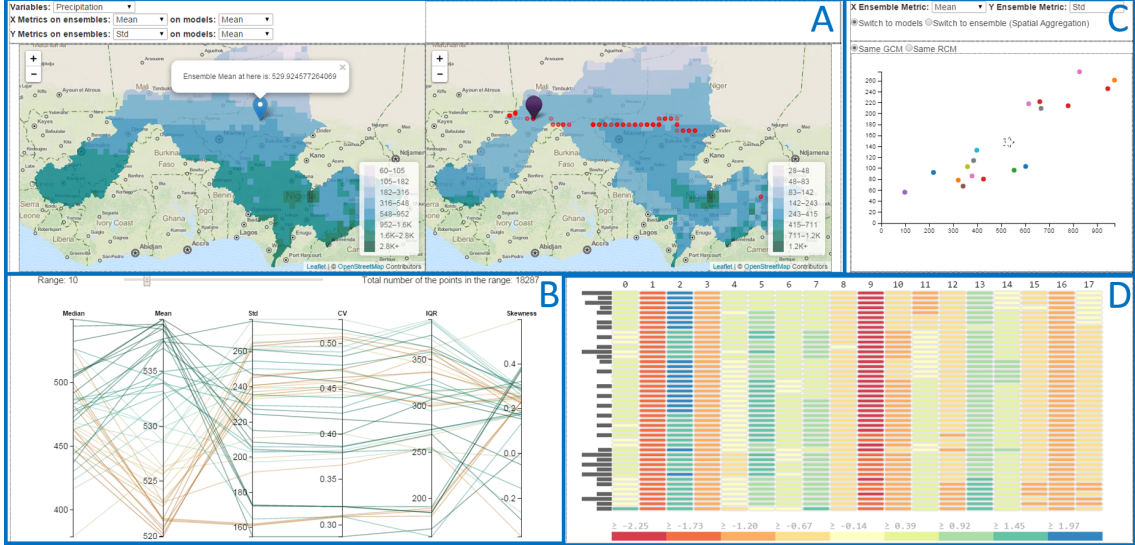


Figure 5.4: The tool for exploring ensemble-level uncertainty among ensembles of climate models. Part(A) visualizes the ensemble-level uncertainty quantified by various uncertainty quantification metrics and also supports multiple interactions directly on the map, such as dragging, zoom in and out, and clicking. By clicking on the left map, the scatter plot in part(C) shows the distribution of climate models on that clicked point. Each dot represents a single model and its color is grouped by model settings. In part(B), each vertical axis in the parallel coordinate represents an uncertainty metric and each line represents a group of points having that metric value. Part(D) visualizes how each model deviates from the average performance within a matrix grid where each column represents a climate model and each row represents a group of points which equals to one line in the parallel coordinate of part(B). The legend for the color is shown at the bottom of the matrix.

certainty which could be $Std(\cdot)$, $IQR(\cdot)$, $CV(\cdot)$, etc. Therefore, if we extract the mean from the feature vector of each model-level distribution, we can compute the ensemble-level uncertainty by computing the standard deviation of the mean values, which can be written as $Std\langle\mu_1(\cdot), \mu_2(\cdot), \dots, \mu_i(\cdot), \dots, \mu_n(\cdot)\rangle$.

To make ensemble-level uncertainty understandable, we take the mean of each model-level distribution as the only feature of each model. However, users are given the freedom to change the feature they want to use in describing the model-level distribution, such as standard deviation, through the select box in Part A. Given the two maps in Part A, users can directly compare the results by changing different features in model-level distributions or changing the uncertainty quantification approaches for

ensemble-level distributions. Also, the two maps are synchronized with each other for the convenience of comparisons across maps.

To investigate one specific area, users can click any point, which is highlighted as the blue marker in the left map. They will see the detailed distribution of each model at that point in the scatter plot of Part C. Each dot in the scatter plot represents a model and can be colored by model settings, such as GCM and RCM. Similar to part C in Figure 5.2, domain experts can also change the quantification methods on axis X and Y such as $\mu(\cdot)$, $\sigma(\cdot)$, $IQR(\cdot)$, etc. The dash star in the middle represents the mean value of all models over X and Y dimensions.

Afterwards, we extract a set of points which have similar ensemble mean $Mean \langle \cdot \rangle$ values. From these points, we want to know if they will still be similar in other quantification approaches. In the parallel coordinate of part B, we take each axis as an ensemble-level uncertainty quantification method $Ens \langle \cdot \rangle$ and each line represents one or a set of points. From Figure 5.4, we will see many bifurcate lines indicating that they have the same values in some methods but become different in other methods. For instance, in Figure 5.4, the lines gathering around the top of the Mean axis diverge at the Std axis ranging from 160 to 260. Also, the slope between two neighbor axes shows their relationship. In definition, standard deviation, IQR and CV are all used for quantifying data divergence. Since the lines between standard deviation and CV are nearly parallel in Figure 5.4, we can know that these two approaches are very similar as their definition. However, for the irregular crossings between axis IQR and axis CV, we may guess these two methods have different emphasis in quantifying data divergence and therefore behaved very differently.

Part D is another overview showing the distribution of models. Each row represents one or more spatial points in the map and each column represents a model. Here we use the model-level mean as the feature of each model and then standardize

the model-level mean value (z-score) of each model with respect to the mean values of all models. From the standardized values, we can see how far each model is deviated from their average. Encoded by the diverging color from red to blue, the change of z-score value shows the change from negatively maximum deviation to positively maximum deviation. As shown in Figure 5.4, it is easy to see that model 9 has the darkest red for all rows, i.e. all lines in the parallel coordinate, and therefore telling that model 9 is positively most deviated.

In conclusion, this tool supports more detailed data analysis and exploration in comparing differences among models and quantification approaches. It also supports various types of interactions. Users can mouse over, drag and zoom in or out the parallel coordinate and the matrix to highlight interested data points.

5.5 Global Climate Assessment Models

We have collaborated with climate scientists from the Joint Global Change Research Institute and Pacific Northwest National Laboratory, the developers of the Global Change Assessment Model (GCAM). GCAM is a global integrated assessment model combining representations of the global economic, energy, agricultural, land use and climate systems (Clarke *et al.* (2007) and Jin and Guo (2009)). It has made significant contributions to the Inter-governmental Panel on Climate Change (IPCC) climate change assessment reports.

Running GCAM will generate multiple scenarios. In each scenario, the whole world is separated into multiple regions. Each region includes multiple variables such as population, GDP, electricity, and geothermal. And each variable is represented as a time series which involves both historical data and future prediction data. Based on such complex data structures, naive data mining techniques, such as K-means, in clustering algorithms cannot be directly used for data analysis. Particularly, regarding

its spatial and temporal characteristics, its high dimensional space, which is composed by regions, time, and variables (e.g. population, GDP, and electricity), cannot be simply projected to the low dimensional space through dimension reduction skills, e.g. principle component analysis (PCA). To mitigate such issue, we try to extract a few data features to describe the original data and then conduct uncertainty analysis over those derived features.

The goal of this line chart tool, shown as Figure 5.5, is to reflect the disparities among different regions for any given output variable and also to point out the most outlying regions (maximum and minimum). Before drawing the line chart, users need to pick one interested output variable. In the line chart, the x-axis represents the time and y-axis represents the variable values. For all given scenarios, we compute the mean value from all regions, which is shown as the black line in the middle of the line chart. We also compute its standard deviation and draw the three sigma range ($\mu \pm 3 * \sigma$) as the blue area in the line chart. We can see this range as the expected disparities among different regions if they follow the Normal distribution. On the other hand, we draw the maximum and minimum as the red and blue line in the line chart respectively. Users can mouse over the red or blue line to further check what are these regions. But in order to properly show more detailed information without further mousing over the points, we draw the regional polygon, rather than a dot, in the line chart to show the regional information and only replace the regional polygon if one or more new regions become the maximum or minimum. For example, as shown in Figure 5.5, we can see China region in year 1990 at the beginning of the red line but it changes to India region in year 2020. That means China region has the maximum population at the beginning and then India region turns into having maximum population in 2020.

To demonstrate the use of the line chart tool, we take Figure 5.5 as an example.

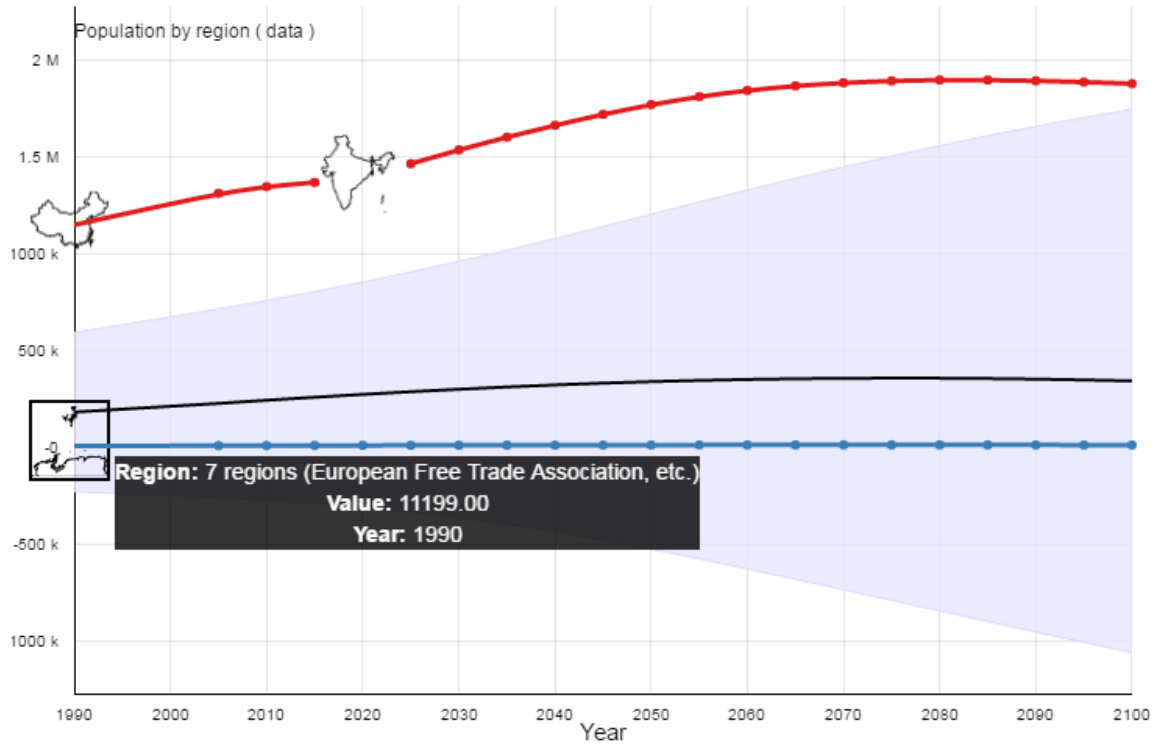


Figure 5.5: In the line chart, the x-axis represents time and y-axis represents the variable values. The blue area shows the expected disparities among regions. The black line shows the average value of all regions. The red and blue line shows the maximum and minimum value in the data respectively. The regional polygon in the red and blue line visualize the geographic information in each dot.

This figure shows the change of populations from 1990 to 2100. From the red and blue line, you can see that in 1990 China has the largest population while European Free Trade Association and other 7 regions have the smallest population. And when it comes to 2020, India replaces China as it has the largest population but European Free Trade Association and other 7 regions, such as the South Pole, still have the smallest population. From the black line, we find the average population is very close to the blue line, which means the majority of all regions is still within small populations. Also, under the assumption of Normal distribution, the growing blue area shows that different regions are expected to have larger differences in population.

One more thing worthwhile to note is the relative position of the red/blue line

and the area chart. As we introduced above, the area chart represents three sigma range $\mu \pm 3 * \sigma$ which can be explained as the expected 99.73 percentile of the data values if they follow the Normal distribution, while on the other side the red/blue line represents the actual maximum/minimum value in the data. Based on their differences, we may infer some interesting distribution patterns. First of all, if both red and blue line are in the blue area, we may guess that the actual data distribution is narrower than Normal distribution. Secondly, if both red and blue line are outside the blue area, we may infer that the actual data distribution is much sparser and spread than Normal distribution. Last and more often in our actual cases, one of the red and blue line is inside the area while another line is outside the area, which can be seen in Figure 5.5. For this case, we may infer that the center in the actual data distribution is shifted to one side.

5.6 Data Processing and Implementation

In our project, there are two major data sources. The first type of source data is provided by the Coordinated Regional Climate Downscaling Experiment (CORDEX) which is sponsored by the World Climate Research Program. The second data with larger regional scale and more outputs is provided by the Joint Global Change Research Institute and Pacific Northwest National Laboratory. During the collaboration with two groups of domain experts, they proposed different requirements for the tool, e.g. one requiring a web application and another one requiring a desktop application, and therefore we used different techniques to process the data and implement the tool.

For the first data source, it is estimated from outputs of different combinations of Global and Regional Climate Models (GCMs and RCMs) and is extracted within the range of the Niger River basin by the domain experts. The data has three variates:

precipitation, minimum temperature, and maximum temperature. The GCM-RCM combinations of CORDEX were run in a historical period from 1951 to 2000. Therefore, we have 18 combinations of GCMs and RCMs, each of which contains 50 time slices map data and is generated under a chaotic process. The resulting maps of each model have a resolution of 30 sec and units of m^3/year . Particular technique details such as the data discretization in computing normalized Shannon entropy are covered as described in Section 3. To enable the fast and interactive analysis environment in a light web-based framework, all the data is precomputed and stored on the server side. The total size of all computed data using seven quantification approaches is over 30 Gigabytes but will be on-call as request. Parallel processing, such as parallelly drawing the map, is also supported on the server side. In the front-end side, the libraries we used include JQuery, Bootstrap, D3.js and Leaflet. In the back end, the server is installed on Apache Tomcat and developed by Java. Particularly, Java library GDAL is used for parsing the data in GeoTiff format.

For the second data source, compared to the first data source, its size is very small and hence all computation work can be executed on the fly. Regarding its complex data structure which has been discussed in section 5.5, we first aggregate the data in different levels. For example, to compute the mean value of all regions, we need to aggregate the data for all regions and scenarios. Due to the requirements of running the tool locally, we implemented this tool using Electron which is a high-level framework based on JavaScript. In order to avoid the reproduction of some basic data mining techniques such as PCA and hierarchical clustering, we linked an external call to the python packages. Since the tool is enabled by the Chromium V8 engine in nature, all the functionalities are implemented through JavaScript frameworks and the performance is up to the performance of running machines.

Chapter 6

CASE STUDIES

In this chapter, I will demonstrate how I integrate the taxonomy studies and developed systems into practical uncertainty analysis through two case studies. In the first case study, I aim to help users explore the agreement levels on water scarcity in the Niger River Basin area among different models as well as assess the future uncertainty. In the second case study, I will present an initial visualization prototype that combines data from population and climate simulation as inputs to a patch-based mosquito spread model for analyzing the potential disease spread vectors and their relationship to climate variability.

6.1 Prediction of Water Scarcity in Niger River Basin

In this case study, I have developed a geovisual analytics tool for exploring simulation results under combinations of climate models, climate policies, and future population growth. Moreover, our tool is capable of ensemble-visualization and allows users to explore agreement levels among different climate models to assess future uncertainty.

6.1.1 Data Description:

Water Supply Data: The water supply was estimated from outputs of different combinations of Global (GCMs) and Regional (RCMs) Climate Models. These were provided by the Coordinated Regional Climate Downscaling Experiment (CORDEX), a project sponsored by the World Climate Research Program that uses a set of advanced RCMs to dynamically downscale the latest set of GCM climate scenarios and

predictions produced within the 5th Coupled Model Intercomparison Project (CMIP5) (Giorgi and Asrar (2009)). The GCM-RCM combinations of CORDEX were run in a historical period from 1950 to 2005 and for future climate projections from 2006 to 2100 under the newly developed Representative Concentration Pathways (RCPs) (Vuuren *et al.* (2011)). Specifically, for each pixel, water supply was computed as the sum of the local runoff plus river corridor discharge. The resulting maps of water supply have a resolution of 30 sec and units of m³/year (as shown in Fig. 6.1 (b).)

Water Demand Data: Water demand data can be measured in multiple ways. For example, it can be measured in terms of the liters of water per person needed based on daily usage such as drinking and bathing, or based on usage by different sectors such as agricultural and industrial demand. In this work, we use the Falkenmark index (Falkenmark (1989)), which is an average regional indicator (with pre-defined thresholds) that measures water demand by the total cubic meters of water availability per person per year in a region. We have collected and generated historical population data as well as population projections. Historical population density data (in *person/km*²) is collected from the Gridded Population of the World (GPW) v3 from the Socioeconomic Data and Applications Center (SEDAC), Columbia University (resampled to 30 arc-second (approx. 1km) resolution). We projected the spatial distribution of future population through the year 2100 using two different models. The first is an exponential growth model assuming that population in the basin will grow at a given percentage each year. The second model for population projection is based on the Shared Social Path (SSP) population projections proposed by Moss *et al.* (2010). The SSP provides the projected total population for each of the basin countries at 5 year intervals until 2100.

Water Scarcity Data: Based on the per capita water usage in cubic meters, the water conditions in a pixel can be categorized per Falkenmark (Falkenmark (1989))

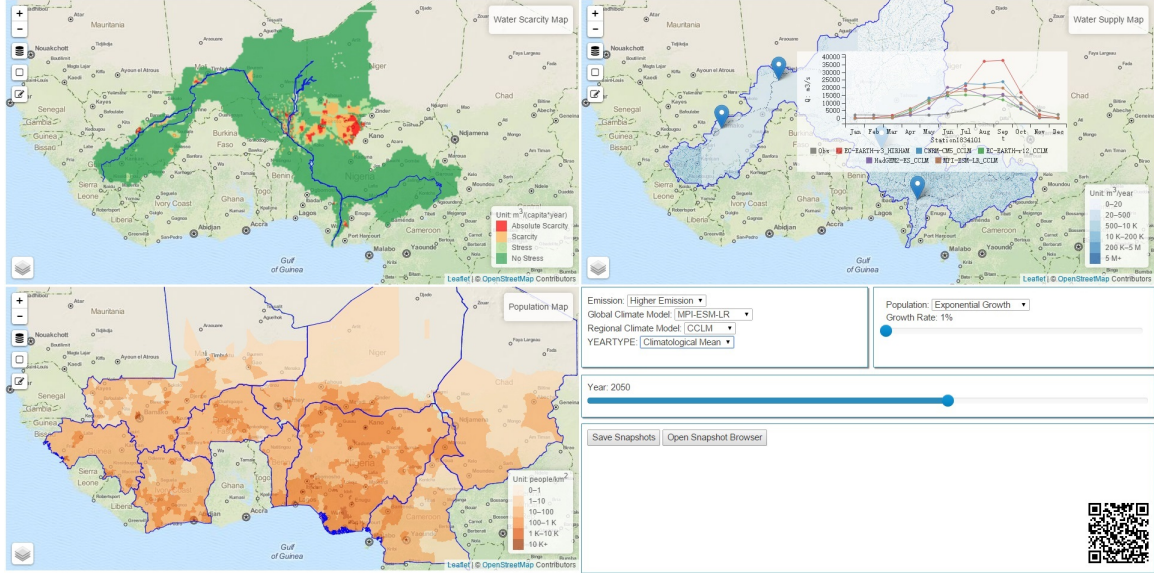


Figure 6.1: Simulation view for the RCP85 emission policy with MPI-ESM-LR GCM and CCLM RCM in climatological mean condition in 2050. In clockwise order, starting from the upper left view, they are (a) Water scarcity map in the 4 scarcity levels using Falkenmark indicator, (b) Water supply map with gauge station view and river networks, (c) Water demand map with political boundaries using a 1% exponential growth model, and (d) Control panel for composing water scarcity scenarios.

as: no stress (greater than 1700), stress (between 1000 and 1700), scarcity (between 500 and 1000), and absolute scarcity (less than 500). To apply the Falkenmark indicator, we (i) calculated the water availability per capita per year as the ratio between water supply and population layers (in $m^3/person\ year$), and (ii) classified the supply demand ratio according to the thresholds of 500, 1000 and 1700 (Fig. 6.1(c)), which is also how we apply the color scheme in the visualization part.

6.1.2 Visual Analytic View

The goal of our visualization platform is two-fold: 1) to simulate historical and future scenarios, and; 2) to compare and analyze the uncertainty associated with these scenarios. Two visualization views are designed accordingly.

The first view is the simulation view (Fig. 6.1), which consists of three map pan-

els for geographical information (about water supply, demand, and scarcity, resp.) and one control panel that allows users to enter parameters and generate scenarios. Quantities of water supply and demand were color-coded in an intuitive manner in the map panels. The blue color was used (shown in Fig. 6.1(b)) to present water supply, and tones from deep blue to light blue represent water supply values from large to small. Similarly, orange colors were used in Fig. 6.1(c) to represent different levels of water demand. For water scarcity (Fig. 6.1(a)), four colors: light green, yellow, orange, and red were used to represent increasing levels of water scarcity defined by the Falkenmark index, with alarming colors such as yellow and red indicating areas of scarcity.

In addition to basic information about spatial distributions of water supply and demand, we used interactive visual elements on top of the base layers to provide users with rich information about data, modeling inputs, and spatial contexts. As shown in Fig. 6.1(b), we have implemented popup windows to visualize the volume of stream flows at stream gauge stations, which were used to derive and calibrate the amount of water supply. Time series of stream flows (m^3/s) by month were plotted as line charts for different climate models.

The control panel (Fig. 6.1(d)) provides a summary of basic model parameters. Data were organized by the modeling year and then by the types of water demand/supply models. A slide bar and dropdowns were used to allow a user to select any modeling year and combinations of water supply model and demand models. To facilitate decision making and communication in a collaborative environment, we have developed an interface to allow users to store interesting scenarios as model profiles in a database. We have also implemented other features to facilitate data exploration. For example, when exploring maps with dragging or zooming, three maps are synchronized to the same zoom level and view center, which helps in targeting

problem areas. Users are also allowed to use an “area selection” mode to highlight regions of interest.

In Fig. 6.2, following the color schemes suggested by Kaye *et al.* (2012), we used a two-dimensional color matrix that employs color and hue in order to simultaneously visualize the average intensity of water scarcity as predicted by the models as well as uncertainty associated with ensemble predictions using multiple models. The same four colors as in the simulation view were used to represent average water scarcity.

To visualize the uncertainty from ensemble prediction, we introduce the level of agreement among model predictions as a measure of uncertainty. We employ tones as a second channel and use lighter colors to indicate greater uncertainty about an ensemble-predicted water scarcity level. The agreement level is measured as the percentage of the dominant water scarcity value out of all predicted water scarcity values. For example, if 3 out of 5 water demand/supply models predicted “absolute scarcity” for a cell, and the other two predicted “scarcity” and “stress”, respectively, the agreement (or certainty) level is $3/5 = 60\%$. In this case, the color of the cell would be salmon red, corresponding to the second column and first row in the legend. By definition, the greater the agreement level, the lower the uncertainty is in an ensemble prediction. We have also explored using an entropy metric to evaluate uncertainty of ensemble prediction as the chaos/entropy in the ensemble of model results, which showed similar spatial patterns of uncertainty as in Fig. 6.2.

6.1.3 Case Studies

We focus on the comparisons of water scarcity computed from combinations of five different water supply models under the RCP45 emission scenario and one water demand model (with a 1% exponential growth) as shown in Fig. 6.2. In total, there are five combinations of water models. We categorize the agreement level into four

regions: 40%, 60%, 80%, and 100% for each level of water scarcity. The minimum agreement is 40% because there are only four scarcity levels in five water models and at least two of them must have the same scarcity level.

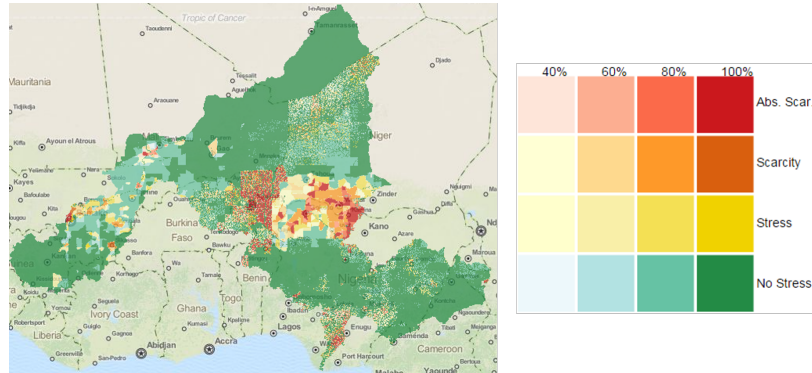


Figure 6.2: Agreement view for 5 combinations of water models in 2025. Color matrix (right) shows the visualization scheme. Each row represents different levels of water scarcity with color; each column represents different agreement levels (40% to 100%) with hue.

From Fig. 6.3, we can observe that under the assumption of relatively mild future population growth (1%), there is little water scarcity (or absolute scarcity) over the majority of in the basin. Notable exceptions are areas around the cities of Niamey and Sokoto (see also a closer view in Fig. 6.3), where there is significant water scarcity partly due to the much higher population density there. However, from the tones of the color, we can observe that the five models reach low levels of agreement on the water scarcity around the two cities for 2025 projections. This case study exemplifies the uncertainty associated with ensemble predictions that can be explicitly visualized with our tool and can be neglected when only mean values of model results are used. With our tool, we have also identified cases in which the models predict water scarcity with higher levels of agreement over a longer term. As shown in Fig. 6.3, we can observe that the water scarcity in city Sokoto (pentagram) and Katsina (X-Star) are merging with each other. While some surrounding regions (the lighter red regions)

have higher uncertainty in 2025, they are mostly turning into the absolute scarcity category with high certainty in the more distant future.

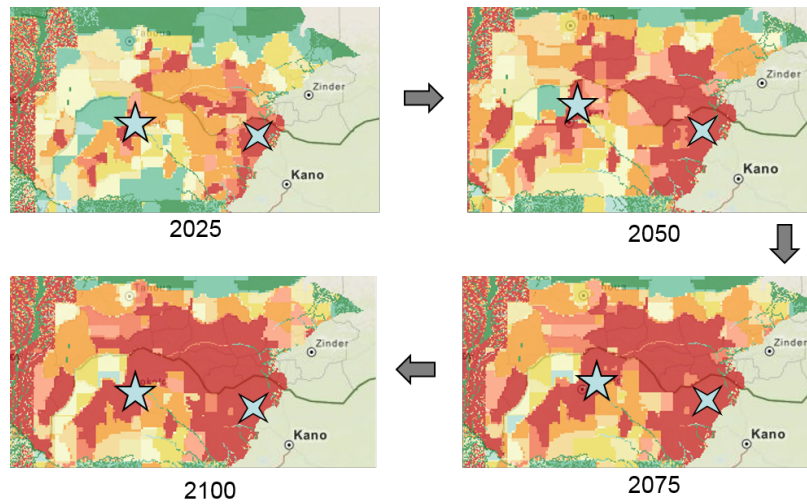


Figure 6.3: The scarcity of two cities, Sokoto and Katsina, is merging from 2025 to 2100.

6.2 Prediction of Malaria Spread in Western Africa

In this case study, I will present an initial visualization prototype that combines data from population and climate simulations as inputs to a patch-based mosquito spread model for analyzing potential disease spread vectors and their relationship to climate variability.

6.2.1 Data Modeling

To simulate and explore the relationship between climate variability and transmission of mosquito borne diseases, we first need to provide a model with a set of initial conditions. We use the historical population data from the first case study as the input for the initial amounts of potential human hosts, the historical temperature and precipitation data as the descriptors of climate variability (which future work

will replace with a downscaled ensemble of climate simulations for predictive analysis), and an epidemic model of malaria transmission. Patches in the transmission model are simplified as geographical rectangles in a region. As input to the model, each patch uses the average precipitation and temperature value of all pixels contained within a patch. Future work will explore downscaling methods for increased temporal and spatial resolutions to improve the model efficacy.

The epidemic model used in our study is based on a meta-population mathematical model for the transmission dynamics of malaria in a community consisting of multiple patches, which takes into account the effect of temperature (air and water) and precipitation variability on the hosts and vectors (Agusto *et al.* (2015)). The total host population at time t for each patch i , denoted by $N_H^{(i)}(t)$, and is split into four epidemiological states, namely mutually-exclusive susceptible $S_H^{(i)}(t)$, exposed (with no clinical symptoms of malaria) $E_H^{(i)}(t)$, infectious $I_H^{(i)}(t)$ and recovered individuals $R_H^{(i)}(t)$, where $N_H^{(i)}(t) = S_H^{(i)}(t) + E_H^{(i)}(t) + I_H^{(i)}(t) + R_H^{(i)}(t)$. Similarly, the total population for vectors at time t for each patch i , denoted by $N_V^{(i)}(t)$, is subdivided into three compartments where $L_V^{(i)}$ denotes the immature mosquitoes (eggs, larvae and pupae), adult mosquitoes $S_V^{(i)}(t)$, and infectious mosquitoes $I_V^{(i)}(t)$. Hence, $N_V^{(i)}(t) = L_V^{(i)}(t) + S_V^{(i)}(t) + I_V^{(i)}(t)$. The equations for the patch model considered in this study take the simplified form of the deterministic system of non-linear differential equations given by Agusto *et al.* (2015) within the multi-patch framework. Though there is a large number of parameters in this model, we primarily use the values suggested by (Agusto *et al.* (2015)) and only tweak the climatic variables based on historical data. In order to solve the differential equations based on the web-based system, related computation is parallelly performed by the support of Parallel Javascript library (Savitzky (2016)) and Numeric Javascript library (Loisel (2016)).

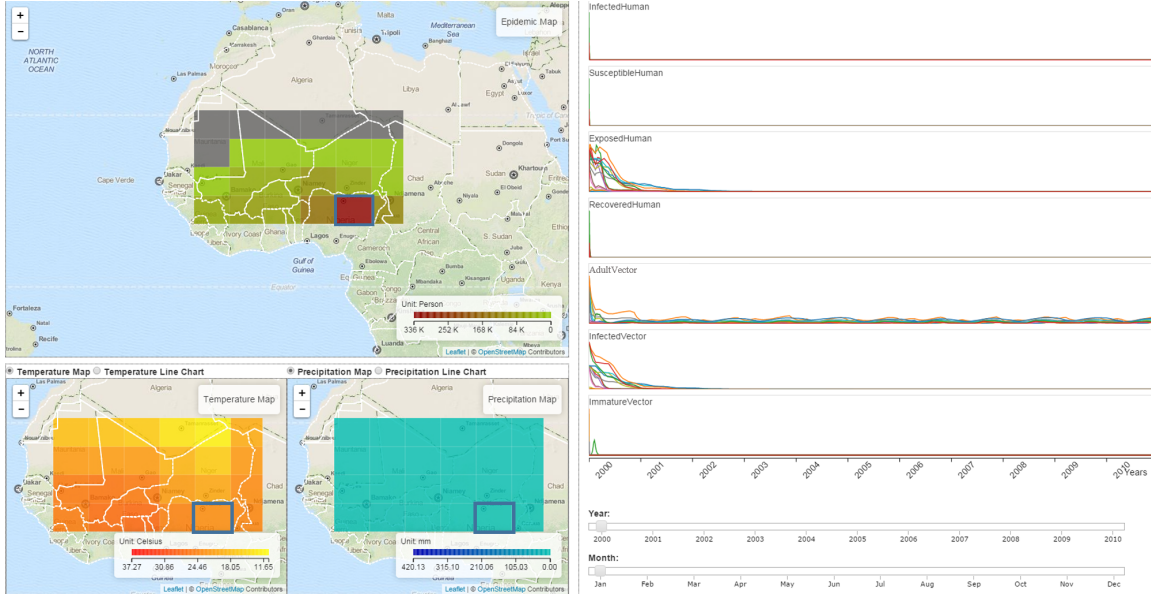


Figure 6.4: Overview of the default system. The top left map shows the spatial distribution of infected humans. The patch highlighted by the blue rectangle is identified as the center of disease and population diffusion. The lower left two maps show the distribution of temperature and precipitation which can also be switched to line chart views showing their temporal changes. The right line charts show the changing amount of seven states in host and vectors by the time. The lower right sliders are used for changing the time by months or by years.

6.2.2 Visual Analytic View

Our prototype system consists of two visualization elements: line charts and map views. Line charts are used for representing the change of each model variable over the time. Map views are used for showing the disease transmission over space or temperature or precipitation distributions, as shown in Figure 6.4. In the geographic views, each patch is a semi-transparent rectangle on the map. In the largest map, patches in red denote higher volumes of infectious humans and patches in green denote lower volumes of infectious human. Gray patches represent regions where population data is unavailable, and future work will explore methods for automatically estimating population from satellite imagery. For the other two maps, the left map shows the temperature distribution with orange representing higher temperatures and yellow

lower temperatures. The right map shows the precipitation with darker blue being larger amounts of precipitation. To see the temporal trend of the temperature or precipitation, users may click the radio button above the two maps and switch between the map view and line chart view.

The right part of our interface, Figure 6.4, consists of seven line charts each showing the host and vector states over time (these are the states previously defined in Section 6.2.1). The x-axis represents time and the y-axis represents the amount of the host or vector in that state. Each line in the line charts represents a patch and is distinguished by the color. From top to bottom, these states are infectious host, susceptible host, exposed host, recovered host, adult vector, infected vector and immature vector. Despite the visual clutter in the line charts, users can mouse over the patch on the map to highlight the corresponding line.

6.2.3 Case Studies

Our current prototype focuses on the West Africa region near the Niger basin and shows regular patterns of temperature and precipitation. What is of interest is exploring how changes in climate will impact the resultant amount of malaria cases. In the epidemic model, the influence from the temperature are directly projected onto the natural mortality rate of immature mosquitoes, egg deposition rate and maturation rate, which will further affect the amount of infected vectors and the infectious host. Therefore, it is reasonable to expect that the change of adult and immature vectors, even for the infected host, may follow the change of temperature or precipitation. Also, due to the mobility of the population among patches, which is also considered in the epidemic model, we can make a bold presumption that the infectious host or infected vectors will also diffuse or transmit among patches. To validate the above assumptions, the model visualization is explored by modeling experts.

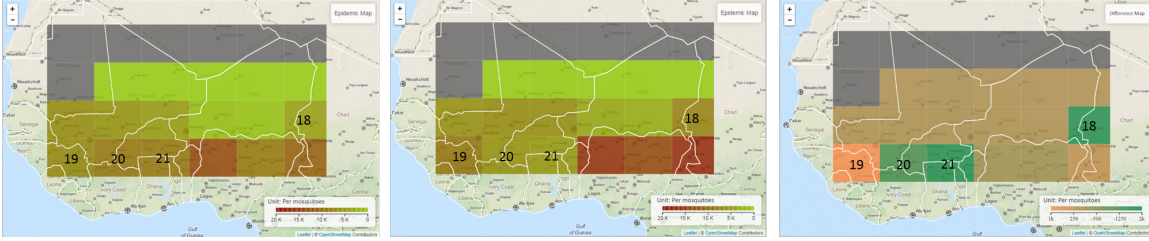


Figure 6.5: Difference maps for the infected vector in March 2000 between simulations with historical temperature (left) and 2°C higher temperature (right). The color in the right map from orange to green represents the positive growth rate 1000 to negative growth rate 2000.

Parameter Settings: As a simulation system, parameter settings and data processing plays an important role in generating the simulated results. In this model, humans and mosquitoes are considered as the disease host and vectors respectively. All parameters, except the initial total of hosts, vectors, initial infectious hosts, infected vectors, temperature and precipitation, are fixed using the values suggested by Augusto *et al.* (2015). The amount of vectors is proportional to the magnitude of the total hosts and can be modified as more mosquito collection reports are provided. The amount of initial infectious hosts and infected mosquitoes are set as ten percent of the total amount of host and vector separately. The West Africa region is uniformly split into 24 patches and they are numbered from 1 to 24 for convenience from left to right and from top to bottom. During the computation, temperature and precipitation are loaded from the collected historical data directly. The step size during the integration of the differential equations is taken as 10^{-6} .

Simulation Results And Analysis: Figure 6.4 shows the default visualization at the beginning of the simulation. The epidemic map on the top-left shows the distribution of initial infectious humans. As an explorative trial, Figure 6.5 shows a simple first pass difference map where the user simply compares what happens to disease transmission if the average temperature increases by 2°C . We can see more infected vectors emerge in patch 19 and less infected vectors in patch 18, 20 and 21.

CONCLUSION AND FUTURE WORK

The goal of this work is bidirectional: 1) to help visualization scientists correctly recognize the role of domain experts in uncertainty analysis, and; 2) to help domain experts understand the uncertainty visualization and quantification techniques in the visualization field. To achieve this goal, I have conducted several surveys and interviews as well as developed three web-based visualization frameworks. The surveys include a long-term interview and two short-term surveys. The long-term interview occurred over one and half years with weekly face-to-face and email meetings. The short-term surveys include a thorough literature review and an online questionnaire. In the literature review, I have summarized and compared three types of uncertainty quantification approaches: descriptive statistics, inferential statistics and information theory. Related visualization approaches are also summarized and analyzed in the taxonomy. Through the questionnaire, I have collected some domain gaps in uncertainty analysis, and also characterized several uncertainty quantification and visualization problems. The implementation of those web-based visualization frameworks is a practice of the knowledge from the previous surveys. During the implementation, each step involves the choice of visualization and quantification methods, how to let the domain experts engage in uncertainty analysis, and how to take advantage of the domain knowledge to explain the uncertainty analysis results. In the first case study, I presented a visual analytics framework for analyzing water scarcity in the Niger River Basin. Basin-wide water demand and supply as well as derived water scarcity are visualized using interactive maps, which can provide the user with auxiliary information, such as the climatological input, population distribution, and geographical

context within which water scarcity occurs. More importantly, this tool can assist users in exploring different future scenarios by allowing multiple water supply and demand models to be used for predicting future patterns of water scarcity. In the second case study, I have developed a prototype system which provides a method for linking multi-source data with disease transmission models. The goal of this tool is to develop a way to explore the impact of future climate variability. Thus, future work will explore the creation of disease risk maps using an ensemble of climate simulations as input to the model. In conclusion, each tool presented in the work successfully allows users to explore the model data from multiple perspectives and effectively visualizes uncertainty information in intuitive ways, which demonstrates the usefulness of the taxonomy and effectiveness of the uncertainty analysis.

There are many extensions and work that are worth further exploration. Given the time sensitivity of the literature review, the taxonomy on this topic should be updated regularly. More detailed analysis over the three types of quantification methods should be added in future. Also, if possible, the online questionnaire should have more participants in order to find more insights into the domain gaps between the climate research and visualization field. The personal bias in the questionnaire from the author also need to be reduced. The structure and form of the questionnaire should be improved by taking the advice from experts in related fields. For the implemented tools, more new features could be added. For example, because the tools are all developed as web-based frameworks, it is possible to extend them for different devices such as tablet, cell phones, etc. Other features, such as the similarity analysis over ensembles of features, linking and coupling the implemented visualization tools into an integrated system, and more organized visual components, could be added.

REFERENCES

- Adamowski, J. F., “Peak daily water demand forecast modeling using artificial neural networks”, *Journal of Water Resources Planning and Management* **134**, 2, 119–128 (2008).
- Agusto, F., A. B. Gumel and P. Parham, “Qualitative assessment of the role of temperature variations on malaria transmission dynamics”, *Journal of Biological Systems* **23**, 04, 1550030 (2015).
- Bensema, K., L. Gosink, H. Obermaier and K. Joy, “Modality-driven classification and visualization of ensemble variance”, *IEEE Transactions on Visualization and Computer Graphics* **PP**, 99, 1–1 (2015).
- Biswas, A., S. Dutta, H.-W. Shen and J. Woodring, “An information-aware framework for exploring multivariate data sets”, *IEEE Transactions on Visualization and Computer Graphics* **19**, 12, 2683–2692 (2013).
- Bonneau, G.-P., H.-C. Hege, C. R. Johnson, M. M. Oliveira, K. Potter, P. Rheingans and T. Schultz, “Overview and state-of-the-art of uncertainty visualization”, in “Scientific Visualization”, pp. 3–27 (Springer, 2014).
- Buttenfield, B. and R. Weibel, “Visualizing the quality of cartographic data”, in “Third International Geographic Information Systems Symposium (GIS/LIS 88), San Antonio, Texas”, (1988).
- Chen, H., S. Zhang, W. Chen, H. Mei, J. Zhang, A. Mercer, R. Liang and H. Qu, “Uncertainty-aware multidimensional ensemble data visualization and exploration”, *IEEE Transactions on Visualization and Computer Graphics* **21**, 9, 1072–1086 (2015).
- Chen, M. and H. Jaenicke, “An information-theoretic framework for visualization”, *IEEE Transactions on Visualization and Computer Graphics* **16**, 6, 1206–1215 (2010).
- Claessen, J. H. and J. J. Van Wijk, “Flexible linked axes for multivariate data visualization”, *IEEE Transactions on Visualization and Computer Graphics* **17**, 12, 2310–2316 (2011).
- Clarke, L., J. Lurz, M. Wise, J. Edmonds, S. Kim, S. Smith and H. Pitcher, “Model documentation for the minicam climate change science program stabilization scenarios: Ccsp product 2.1 a”, Pacific Northwest National Laboratory, PNNL-16735 (2007).
- Correa, C. D. and P. Lindstrom, “The mutual information diagram for uncertainty visualization”, *International Journal for Uncertainty Quantification* **3**, 3 (2013).

- Dasgupta, A., J. Poco, Y. Wei, R. Cook, E. Bertini and C. T. Silva, “Bridging theory with practice: An exploratory study of visualization use and design for climate model comparison”, *IEEE Transactions on Visualization and Computer Graphics* **21**, 9, 996–1014 (2015).
- Demir, I., C. Dick and R. Westermann, “Multi-charts for comparative 3d ensemble visualization”, *IEEE Transactions on Visualization and Computer Graphics* **20**, 12, 2694–2703 (2014).
- Efron, B. and R. J. Tibshirani, *An introduction to the bootstrap* (CRC press, 1994).
- Endert, A., M. S. Hossain, N. Ramakrishnan, C. North, P. Fiaux and C. Andrews, “The human is the loop: new directions for visual analytics”, *Journal of intelligent information systems* **43**, 3, 411–435 (2014).
- Falkenmark, M., “The massive water scarcity threatening africa-why isn’t it being addressed”, *Ambio* **18**, 2, pp. 112–118 (1989).
- Feng, D., L. Kwock, Y. Lee, R. M. Taylor *et al.*, “Matching visual saliency to confidence in plots of uncertain data”, *IEEE Transactions on Visualization and Computer Graphics* **16**, 6, 980–989 (2010).
- Giorgi, C. J., F. and G. R. Asrar, “Addressing climate information needs at the regional level: the cordex framework”, *WMO Bulletin* **58**, 3, 175–183 (2009).
- Gleckler, P. J., K. E. Taylor and C. Doutriaux, “Performance metrics for climate models”, *Journal of Geophysical Research: Atmospheres* **113**, D6 (2008).
- Goodwin, S., J. Dykes, S. Jones, I. Dillingham, G. Dove, A. Duffy, A. Kachkaev, A. Slingsby and J. Wood, “Creative user-centered visualization design for energy analysts and modelers”, *IEEE Transactions on Visualization and Computer Graphics* **19**, 12, 2516–2525 (2013).
- Gosink, L., K. Bensema, T. Pulsipher, H. Obermaier, M. Henry, H. Childs and K. I. Joy, “Characterizing and visualizing predictive uncertainty in numerical ensembles through bayesian model averaging”, *IEEE Transactions on Visualization and Computer Graphics* **19**, 12, 2703–2712 (2013).
- Hall, A. and S. Manabe, “Can local linear stochastic theory explain sea surface temperature and salinity variability?”, *Climate Dynamics* **13**, 3, 167–180 (1997).
- Havre, S., E. Hetzler, P. Whitney and L. Nowell, “Themriver: Visualizing thematic changes in large document collections”, *IEEE transactions on visualization and computer graphics* **8**, 1, 9–20 (2002).
- Jaynes, E. T., “Information theory and statistical mechanics”, *Physical review* **106**, 4, 620 (1957).
- Ji, G. and H.-W. Shen, “Dynamic view selection for time-varying volumes”, *IEEE Transactions on Visualization and Computer Graphics* **12**, 5, 1109–1116 (2006).

- Jin, H. and D. Guo, “Understanding climate change patterns with multivariate geovisualization”, in “2009 IEEE International Conference on Data Mining Workshops”, pp. 217–222 (IEEE, 2009).
- Kandlikar, M., J. Risbey and S. Dessai, “Representing and communicating deep uncertainty in climate-change assessments”, *Comptes Rendus Geoscience* **337**, 4, 443–455 (2005).
- Kao, D., A. Luo, J. L. Dungan and A. Pang, “Visualizing spatially varying distribution data”, in “Information Visualisation, 2002. Proceedings. Sixth International Conference on”, pp. 219–225 (IEEE, 2002).
- Kaye, N., A. Hartley and D. Hemming, “Mapping the climate: guidance on appropriate techniques to map climate variables and their uncertainty”, *Geoscientific Model Development* **5**, 1, 245–256 (2012).
- Kehrer, J., P. Filzmoser and H. Hauser, “Brushing moments in interactive visual analysis”, in “Computer Graphics Forum”, vol. 29, pp. 813–822 (Wiley Online Library, 2010).
- Klir, G. and M. Wierman, *Uncertainty-based information: elements of generalized information theory*, vol. 15 (Springer Science & Business Media, 1999).
- Kniss, J. M., R. Van Uitert, A. Stephens, G.-S. Li, T. Tasdizen and C. Hansen, “Statistically quantitative volume visualization”, in “IEEE Visualization, 2005. VIS 05.”, pp. 287–294 (IEEE, 2005).
- Kwon, B. C., B. Fisher and J. S. Yi, “Visual analytic roadblocks for novice investigators”, in “2011 IEEE Conference on Visual Analytics Science and Technology (VAST)”, pp. 3–11 (IEEE, 2011).
- Liu, S., J. A. Levine, P.-T. Bremer and V. Pascucci, “Gaussian mixture model based volume visualization”, in “Large Data Analysis and Visualization (LDAV), 2012 IEEE Symposium on”, pp. 73–77 (IEEE, 2012).
- Loisel, S., “Numeric javascript”, <http://www.numericjs.com/> (February 2016).
- MacEachren, A. M., “Visualizing uncertain information”, *Cartographic Perspectives*, 13, 10–19 (1992).
- MacEachren, A. M., “Visual analytics and uncertainty: Its not about the data”, (2015).
- MacEachren, A. M., A. Robinson, S. Hopper, S. Gardner, R. Murray, M. Gahegan and E. Hetzler, “Visualizing geospatial information uncertainty: What we know and what we need to know”, *Cartography and Geographic Information Science* **32**, 3, 139–160 (2005).
- MacEachren, A. M., R. E. Roth, J. O’Brien, B. Li, D. Swingley and M. Gahegan, “Visual semiotics & uncertainty visualization: An empirical study”, *IEEE Transactions on Visualization and Computer Graphics* **18**, 12, 2496–2505 (2012).

- Maciejewski, R., S. Rudolph, R. Hafen, A. M. Abusalah, M. Yakout, M. Ouzzani, W. S. Cleveland, S. J. Grannis and D. S. Ebert, “A visual analytics approach to understanding spatiotemporal hotspots”, *IEEE Transactions on Visualization and Computer Graphics* **16**, 2, 205–220 (2010).
- Mann, P. S., “Introductory statistics, wile y”, Tech. rep., ISBN 0-471-31009-3 (1995).
- Mascaro, G., D. D. White, P. Westerhoff and N. Bliss, “Performance of the cordex-africa regional climate simulations in representing the hydrological cycle of the niger river basin”, *Journal of Geophysical Research: Atmospheres* **120**, 24, 12425–12444 (2015).
- Mirzargar, M., R. T. Whitaker and R. M. Kirby, “Curve boxplot: Generalization of boxplot for ensembles of curves”, *IEEE Transactions on Visualization and Computer Graphics* **20**, 12, 2654–2663 (2014).
- Moss, R. H., J. A. Edmonds, K. A. Hibbard, M. R. Manning, S. K. Rose, D. P. Van Vuuren, T. R. Carter, S. Emori, M. Kainuma, T. Kram *et al.*, “The next generation of scenarios for climate change research and assessment”, *Nature* **463**, 7282, 747–756 (2010).
- Najafi, M. R. and H. Moradkhani, “Ensemble combination of seasonal streamflow forecasts”, *Journal of Hydrologic Engineering* **21**, 1, 04015043 (2015).
- Neyman, J., “Outline of a theory of statistical estimation based on the classical theory of probability”, *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences* **236**, 767, 333–380, URL <http://rsta.royalsocietypublishing.org/content/236/767/333> (1937).
- Pang, A., “Visualizing uncertainty in geo-spatial data”, in “Proceedings of the Workshop on the Intersections between Geospatial Information and Information Technology”, pp. 1–14 (National Research Council Arlington, VA, 2001).
- Pang, A. T., C. M. Wittenbrink and S. K. Lodha, “Approaches to uncertainty visualization”, *The Visual Computer* **13**, 8, 370–390 (1997).
- Plewe, B., “The nature of uncertainty in historical geographic information”, *Transactions in GIS* **6**, 4, 431–456 (2002).
- Pöthkow, K. and H.-C. Hege, “Nonparametric models for uncertainty visualization”, in “Computer Graphics Forum”, vol. 32, pp. 131–140 (Wiley Online Library, 2013).
- Potter, K., S. Gerber and E. W. Anderson, “Visualization of uncertainty without a mean”, *IEEE Computer Graphics and Applications* **33**, 1, 75–79 (2013).
- Potter, K., M. Kirby, D. Xiu and C. R. Johnson, “Interactive visualization of probability and cumulative density functions”, *International journal for uncertainty quantification* **2**, 4 (2012a).

- Potter, K., J. Kniss, R. Riesenfeld and C. R. Johnson, “Visualizing summary statistics and uncertainty”, in “Computer Graphics Forum”, vol. 29, pp. 823–832 (Wiley Online Library, 2010).
- Potter, K., J. Krüger and C. Johnson, “Towards the visualization of multi-dimensional stochastic distribution data”, in “Proceedings of The International Conference on Computer Graphics and Visualization (IADIS) 2008”, vol. 53 (2008).
- Potter, K., P. Rosen and C. R. Johnson, “From quantification to visualization: A taxonomy of uncertainty visualization approaches”, in “Uncertainty Quantification in Scientific Computing”, pp. 226–249 (Springer, 2012b).
- Potter, K., A. Wilson, P.-T. Bremer, D. Williams, C. Doutriaux, V. Pascucci and C. R. Johnson, “Ensemble-vis: A framework for the statistical visualization of ensemble data”, in “2009 IEEE International Conference on Data Mining Workshops”, pp. 233–240 (IEEE, 2009).
- Quinan, P. S. and M. Meyer, “Visually comparing weather features in forecasts”, *IEEE Transactions on Visualization and Computer Graphics* **22**, 1, 389–398 (2016).
- Riehmman, P., M. Hanfler and B. Froehlich, “Interactive sankey diagrams”, in “IEEE Symposium on Information Visualization, 2005. INFOVIS 2005.”, pp. 233–240 (IEEE, 2005).
- Saad, A., G. Hamarneh and T. Möller, “Exploration and visualization of segmentation uncertainty using shape and appearance prior information”, *IEEE Transactions on Visualization and Computer Graphics* **16**, 6, 1366–1375 (2010).
- Sacha, D., H. Senaratne, B. C. Kwon, G. Ellis and D. A. Keim, “The role of uncertainty, awareness, and trust in visual analytics”, *IEEE Transactions on Visualization and Computer Graphics* **22**, 1, 240–249 (2016).
- Sanyal, J., S. Zhang, G. Bhattacharya, P. Amburn and R. J. Moorhead, “A user study to compare four uncertainty visualization methods for 1d and 2d datasets”, *IEEE Transactions on Visualization and Computer Graphics* **15**, 6, 1209–1218 (2009).
- Sanyal, J., S. Zhang, J. Dyer, A. Mercer, P. Amburn and R. J. Moorhead, “Noodles: A tool for visualization of numerical weather model ensemble uncertainty”, *IEEE Transactions on Visualization and Computer Graphics* **16**, 6, 1421–1430 (2010).
- Savitzky, A., “Parallel.js”, <https://adambom.github.io/parallel.js/> (February 2016).
- Sedlmair, M., M. Meyer and T. Munzner, “Design study methodology: Reflections from the trenches and the stacks”, *IEEE Transactions on Visualization and Computer Graphics* **18**, 12, 2431–2440 (2012).
- Smith, R. L., C. Tebaldi, D. Nychka and L. O. Mearns, “Bayesian modeling of uncertainty in ensembles of climate models”, *Journal of the American Statistical Association* **104**, 485, 97–116 (2009).

- Spiegelhalter, D., M. Pearson and I. Short, “Visualizing uncertainty about the future”, *science* **333**, 6048, 1393–1400 (2011).
- Taylor, K. E., “Summarizing multiple aspects of model performance in a single diagram”, *Journal of Geophysical Research: Atmospheres* **106**, D7, 7183–7192 (2001).
- Thomson, J., E. Hetzler, A. MacEachren, M. Gahegan and M. Pavel, “A typology for visualizing uncertainty”, in “Electronic Imaging 2005”, pp. 146–157 (International Society for Optics and Photonics, 2005).
- Van der Wel, F. J., L. C. Van der Gaag and B. G. Gorte, “Visual exploration of uncertainty in remote-sensing classification”, *Computers & Geosciences* **24**, 4, 335–343 (1998).
- Vogt, W. P., *SAGE Quantitative research methods* (Sage, 2011).
- Vuuren, D., J. Edmonds, M. Kainuma, K. Riahi, A. Thomson, K. Hibbard, G. Hurtt, T. Kram, V. Krey, J.-F. Lamarque, T. Masui, M. Meinshausen, N. Nakicenovic, S. Smith and S. Rose, “The representative concentration pathways: an overview”, *Climatic Change* **109**, 1, 5–31 (2011).
- Wang, C. and H.-W. Shen, “Information theory in scientific visualization”, *Entropy* **13**, 1, 254–273 (2011).
- Xu, L., T. Y. Lee and H. W. Shen, “An information-theoretic framework for flow visualization”, *IEEE Transactions on Visualization and Computer Graphics* **16**, 6, 1216–1224 (2010).
- Yuan, X. and E. F. Wood, “On the clustering of climate models in ensemble seasonal forecasting”, *Geophysical Research Letters* **39**, 18 (2012).
- Zehner, B., N. Watanabe and O. Kolditz, “Visualization of gridded scalar data with uncertainty in geosciences”, *Computers & Geosciences* **36**, 10, 1268–1275 (2010).

APPENDIX A
ONLINE QUESTIONNAIRE

Survey of Uncertainty Quantification and Visualization

1. Welcome to My Survey

Welcome and thank you for participating in our survey.

My name is Xing Liang. I am a PhD student in the School of Computing, Informatics, and Decision Systems Engineering at Arizona State University. I am currently conducting a survey about ensemble comparison and uncertainty analysis. This survey contains 4 parts and 30 questions in total and will take about 20~30 minutes to complete.

Your participation in this study is voluntary and you can choose not to participate or to withdraw from the questionnaire at any time. The results of this study may be used in reports, presentations, or publications but your name will not be used. If you have any questions concerning this questionnaire, please contact me at Xing.Liang@asu.edu.



Visual Analytics &
Data Exploration Research

Survey of Uncertainty Quantification and Visualization

2. Consent Form

Why am I being invited to take part in a research study?

We invite you to take part in a research study because you are a domain expert in study ensemble data. We are requesting your feedback to assess the impact of uncertainty visualization and quantification in various domain areas. To be able to participate the survey, you should be 18 or older.

Why is this research being done?

I am conducting a research study to investigate the problems of uncertainty quantification and visualization approaches in the research field of the participants for understanding their biases and reasons in choosing the approaches.

How long will the research last?

We expect that individuals will spend 20-30 minutes maximum participating in the proposed activities.

How many people will be studied?

Up to 100 people.

What happens if I say yes, I want to be in this research?

Your participation will ask you to recall your experience in conducting uncertainty analysis in your research projects and answer both choices and opened questions related to the experience. You are free to decide whether you wish to participate in this study.

What happens if I say yes, but I change my mind later?

You can leave the research at any time it will not be held against you.

What happens to the information collected for the research?

Efforts will be made to limit the use and disclosure of your personal information. The survey will be conducted through SurveyMonkey and IP addresses will not be collected when doing the survey. Your responses will be stored on a secure server with password and only accessible to the PI. But we cannot promise complete secrecy. Your data will be only used for research purposes such as numerical data analysis, hypothesis validation, etc.

Who can I talk to?

If you have questions, concerns, or complaints, talk to the research team at rmacieje@asu.edu This research has been reviewed and approved by the Social Behavioral IRB. You may talk to them at (480) 965-6788 or by email at research.integrity@asu.edu if:

- Your questions, concerns, or complaints are not being answered by the research team.
- You cannot reach the research team.
- You want to talk to someone besides the research team.
- You have questions about your rights as a research participant.
- You want to get information or provide input about this research.

YOU HAVE HAD THE OPPORTUNITY TO READ THIS CONSENT FORM, ASK QUESTIONS ABOUT THE RESEARCH PROJECT AND AM PREPARED TO PARTICIPATE IN THIS PROJECT.

Click the given next to enter the survey if you agree to be part of this project.

Survey of Uncertainty Quantification and Visualization

3. Background Question

1. What is your major research area?

- Climate
- Hydrology
- Oceanography
- Land Use/Land Cover
- Atmosphere
- Other (please specify)

2. Have you ever used any software or worked on any research project which involves the quantification and visualization of uncertainty?

- No
- Yes (please specify)

3. Can you briefly tell us your goals while conducting uncertainty analysis? (Specific to your field)

- Support Decision Making
- Support Model Comparison
- Other (please specify)

Survey of Uncertainty Quantification and Visualization

4. Questions in Visualization

4. What are the typical types of data you are visualizing?

For example, if you are an oceanographer, you may deal with vortices more often. If you are a precipitation data modeler, you may deal with precipitation data more often.

5. How many dimensions does the data have?

- 1 dimension
- 2 dimensions
- 3 dimensions
- n dimensions (n>3)

For the following color schemes in question 6-8, which color scheme have you used and for what types of data (precipitation, temperature, etc)?

6. Sequential color scheme: 

- I have never used this color scheme
- I have used it for the data: (please specify)

7. Diverging color scheme: 

- I have never used this color scheme
- I have used it for the data: (please specify)

8. Qualitative color scheme: 

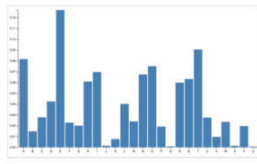
- I have never used this color scheme
- I have used it for the data: (please specify)

9. Which of the following visualizations have you ever used?

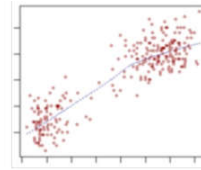
- Line Chart
- Bar Chart
- Scatter Plot
- Box Plot
- Taylor Diagram
- Original Data Space (e.g. Map)
- Spaghetti Plot
- Other (please specify)



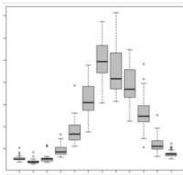
Line chart



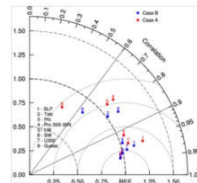
Bar Chart



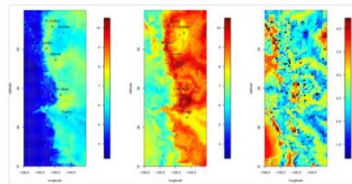
Scatter plot



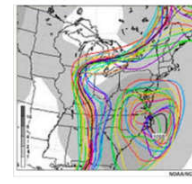
Box plot



Taylor Diagram



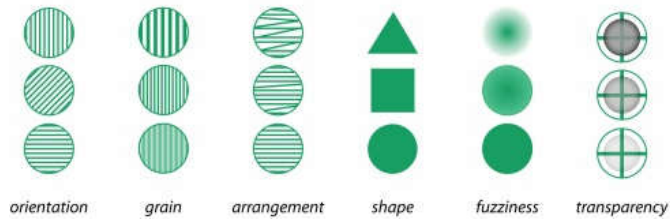
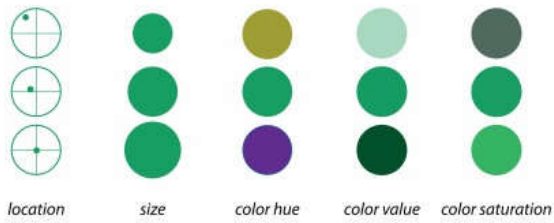
Map/Original Data Space



Spaghetti Plot

10. Given the visual variables reported in the pictures below the answers, which of the followings have you ever felt helpful in representing uncertainty in your data?

- Location
- Size
- Color Hue
- Color Value
- Color Saturation
- Orientation
- Grain
- Arrangement
- Shape
- Fuzziness
- Transparency



11. Which of the following problems have you ever met in your research?

- Clutter:** Overlap, e.g. overlaps of symbols on the map
- Clutter:** Color mixing, e.g. background color is mixing with line colors in line chart
- Distortion:** Scale inconsistency: e.g. inconspicuous Y-axis scale in two similar scatter plots
- Distortion:** Projection error, e.g. errors in mapping 3d sphere to 2d space
- Comparison Complexity:** Superposition overload, e.g. too many lines in a line chart
- Comparison Complexity:** Lack of explicit encoding, e.g. when comparing similarity among data in multiple maps, Euclidean layout distances between maps are not usually considered besides the colors
- Communication Gap: Improper Legend:** e.g. too many symbols in the legend
- Communication Gap: Missing Annotation:** e.g. annotation for the outliers in the map would communicate the intent more effectively

12. In your experience, which of the visual variables mentioned above have caused problems for you in understanding the uncertainty information in your data?

You can consider the problems given above or other problems from your research experience.

13. Can you mention an example where you had problems using the above visual variables or visualizations during uncertainty analysis?

14. In your experience, which of the visual variables mentioned above helped you quickly understand the uncertainty information in your data?

Survey of Uncertainty Quantification and Visualization

5. Questions in Uncertainty Quantification

15. Which of the following metrics have you used in quantifying uncertainty or comparing

ensemble members or simulation results?

- Mean:
 $\mu = \frac{\sum_{i=1}^N x_i}{N}$ where $\{x_1, x_2, x_3, \dots, x_n\}$ are the observed values of the samples
<https://en.wikipedia.org/wiki/Mean>
- Skewness:
 $S_N = \sqrt{\frac{1}{N} \sum_{i=1}^N (x_i - \bar{x})^3}$ where $\{x_1, x_2, x_3, \dots, x_n\}$ are the observed values of the samples and \bar{x} is their mean value while N stands for the size of the samples.
<https://en.wikipedia.org/wiki/Skewness>
- Kurtosis:
 $k = \frac{E[(X-\mu)^4]}{(E[(X-\mu)^2])^2}$ where μ is the mean, and X is the sample, E is the expectation operator
<https://en.wikipedia.org/wiki/Kurtosis>
- Standard deviation/Variance/Joint standard deviation:
 $S_N = \sqrt{\frac{1}{N} \sum_{i=1}^N (x_i - \bar{x})^2}$ where $\{x_1, x_2, x_3, \dots, x_n\}$ are the observed values of the samples and \bar{x} is their mean value while N stands for the size of the samples.
https://en.wikipedia.org/wiki/Standard_deviation
- Interquartile range:
 $IQR = Q_3 - Q_1$ where Q_3 and Q_1 are the upper quartile and lower quartile respectively. Lower quartile represents the lowest 25% of data and upper quartile represents the highest 25% of data.
https://en.wikipedia.org/wiki/Interquartile_range
- Confidence intervals:
The confidence interval contains the parameter values that, when tested, should not be rejected with the same sample. Greater levels of variance yield larger confidence intervals, and hence less precise estimates of the parameter.
https://en.wikipedia.org/wiki/Confidence_interval
- Shannon Entropy:
 $H(X) = -\sum_{i=1}^n P(x_i) \log_b P(x_i)$ where b is the base of the logarithm used and $P(x_i)$ is the probability density function.
[https://en.wikipedia.org/wiki/Entropy_\(information_theory\)](https://en.wikipedia.org/wiki/Entropy_(information_theory))
- Mutual information:
 $I(X; Y) = \sum_{y \in Y} \sum_{x \in X} p(x, y) \log \frac{p(x, y)}{p(x)p(y)}$ where $p(x, y)$ is the joint probability distribution function of X and Y , and $p(x)$ and $p(y)$ are the marginal probability distribution functions of X and Y respectively.
https://en.wikipedia.org/wiki/Mutual_information
- Jensen-Shannon divergence:
 $JSD_{\pi_1, \pi_2, \dots, \pi_n}(P_1, P_2, \dots, P_n) = H(\sum_{i=1}^n \pi_i P_i) - \sum_{i=1}^n \pi_i H(P_i)$ where $\pi_1, \pi_2, \dots, \pi_n$ are the weights that are selected for the probability distributions P_1, P_2, \dots, P_n
https://en.wikipedia.org/wiki/Jensen%E2%80%93Shannon_divergence
- Coefficient of variation:
 $C_v = \frac{\sigma}{\mu}$ where σ and μ is the standard deviation and mean.
https://en.wikipedia.org/wiki/Coefficient_of_variation

Z-Standard value:
 $z = \frac{x-\mu}{\sigma}$ where σ and μ is the standard deviation and mean and x is the raw value.

https://en.wikipedia.org/wiki/Standard_score

Other not listed above (please specify)

16. Are there two or three metrics that you have used more frequently? If so, can you explain why?

You can explain your reasons based on data types, objects and distribution characteristics or the specific analysis requirements, etc.

17. Are there any uncertainty metrics which are very helpful in your area but not commonly used in other areas? If so, can you mention one of them and explain why?

18. Are there other specific problems related to the quantification of uncertainty that you would like to highlight?

Problems can be related to data preprocessing (e.g. discretizing data for specific metrics such as entropy), justify full uncertainty boundary, the computation performance, accuracy, etc.

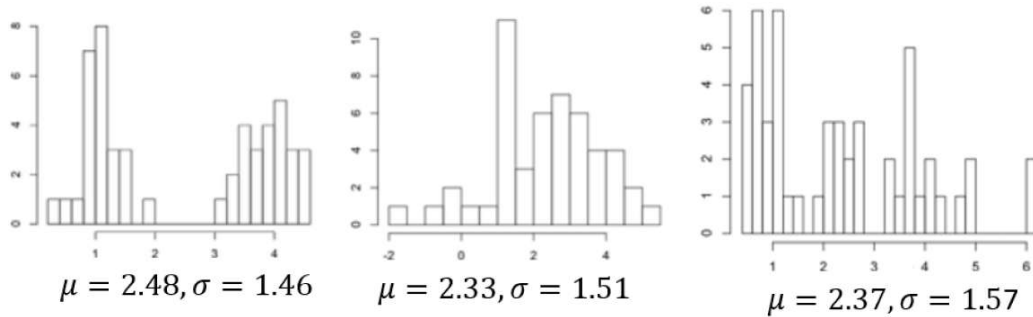
19. Are you aware of the assumption that normal distribution in using moments (e.g. mean, standard deviation, etc.)?

- Yes, I am aware of that
- No, I was not aware of that
- I don't have the experience of using moments in uncertainty analysis

20. Are you aware of that moments are limited to portray the distribution modality?

For example, shown as the figure below, three distributions have very close summary statistics but very different modality characteristics.

- Yes, I am aware of that
- No, I was not aware of that
- I don't have the experience of using moments in uncertainty analysis



21. Are you aware of that the selection of parameters in estimating probability distribution will change the results?

For example, the bandwidth and kernel function in kernel density estimation would influence the estimated distribution.

- Yes, I am aware of that
- No, I was not aware of that
- I don't have the experience of estimating probability distribution.

22. In a Bayesian framework, are you aware of the selection of prior distribution, likelihood function and posterior function would influence the results?

- Yes, I am aware of that
- No, I was not aware of that
- I don't have the experience of estimating probability distribution.

23. Are you aware of the unit problem when using entropy to quantify uncertainty?

- Yes, I am aware of that
- No, I was not aware of that
- I don't have the experience of using entropy

24. Are you aware of the importance of uncertainty boundary in quantifying uncertainty?

- Yes, I am aware of that
- No, I was not aware of that

Survey of Uncertainty Quantification and Visualization

6. Questions in Analyzing Uncertainty Quantification and Visualization

We define domain knowledge as the empirical knowledge accumulated from a considerable amount of studies in their topic and systematical knowledge learnt from the school.

25. In your experience, how frequently do you think the uncertainty quantification results cannot be verified with your domain knowledge?

- All of the results cannot be verified with my domain knowledge
- Half of the results cannot be verified with my domain knowledge
- Some of the results cannot be verified with my domain knowledge
- None of the results cannot be verified with my domain knowledge

26. In your experience, how do you think the chosen visualization metrics helped you to reach the goals?

- Effectively interpret results of the quantification for a domain expert
- Effectively present results of my research in papers or talks for peers in my domain
- Effectively convey information to stakeholders

27. In your experience, which of the followings prevented you from fully understanding the outcomes of the application of an uncertainty metric that you used in the past?

- Knowledge gaps: e.g., the uncertainty metric is hard to understand or it is misleading or insufficient to capture the information I need
- Less or inappropriate contextual information: e.g. loss of details
- Improper quantification approaches: e.g., the metric which can only be applied to continuous data is adapted on discontinuous data
- Visualization gaps: e.g. The visualization is hard to understand or be misleading because they are not frequently seen in your domain or conflicting with your common sense
- Other (please specify)

28. In your experience, what benefits can you recall that you obtained from the uncertainty analysis?

For example: I can gain a big picture of the data; I can characterize the uncertainty of a crucial parameter of a numerical model; I can explain to stakeholders why predictions or estimation of certain parameters/variables are uncertain.

29. In your experience, are you aware of the uncertainty propagation during the uncertainty quantification and visualization process?

Uncertainty can be propagated through loss of data while processing data or loss of details while visualization, etc.

- Yes
- No

30. If Yes above, did you involve them into your analysis? Can you mention an example?

APPENDIX B
IRB APPROVAL FORM



EXEMPTION GRANTED

Ross Maciejewski
Computing, Informatics and Decision Systems Engineering, School of (CIDSE)
480/965-2785
Ross.Maciejewski@asu.edu

Dear Ross Maciejewski:

On 3/23/2016 the ASU IRB reviewed the following protocol:

Type of Review:	Initial Study
Title:	Survey of Uncertainty Quantification and Visualization Approaches
Investigator:	Ross Maciejewski
IRB ID:	STUDY00004081
Funding:	None
Grant Title:	None
Grant ID:	None
Documents Reviewed:	<ul style="list-style-type: none">• IRB-Maciejewski 2016 (3).docx, Category: IRB Protocol;• Recruitment E-mail, Category: Recruitment Materials;• Survey.pdf, Category: Measures (Survey questions/Interview questions /interview guides/focus group questions);• Consent-2016 (2) (1).pdf, Category: Consent Form;

The IRB determined that the protocol is considered exempt pursuant to Federal Regulations 45CFR46 (2) Tests, surveys, interviews, or observation on 3/23/2016.

In conducting this protocol you are required to follow the requirements listed in the INVESTIGATOR MANUAL (HRP-103).

Sincerely,

IRB Administrator

cc:

Xing Liang

APPENDIX C
RESPONSES OF ONLINE QUESTIONNAIRE

Background Question

1. What is your major research area?

Answer Choices	Responses
Climate	0.00% 0
Hydrology	62.50% 5
Oceangraphy	0.00% 0
Land Use/Land Cover	0.00% 0
Atmosphphere	0.00% 0
Other (please specify) Responses	37.50% 3

2. Have you ever used any tool or worked on any research project which involves the quantification and visualization of uncertainty?

Answer Choices	Responses
No	25.00% 2
Yes(please specify) Responses	75.00% 6

3. Can you briefly tell us your goals while conducting uncertainty analysis? (Specific to your field)

Answer Choices	Responses
Support Decision Making	87.50% 7
Support Model Comparison	37.50% 3
Other (please specify) Responses	0.00% 0

Questions in Visualization

1. What are the typical types of data you are visualizing?

emissions by gas by industry by country
 4/18/2016 8:25 PM [View respondent's answers](#)

a huge range of outputs from integrated assessment models. Often costs, energy system characteristics, land use and land cover information.

4/15/2016 8:02 AM [View respondent's answers](#)

Water availability and use data
 3/27/2016 11:26 AM [View respondent's answers](#)

Precipitation, discharges
 3/25/2016 9:33 AM [View respondent's answers](#)

Water related, land use, socio-economic
 3/23/2016 4:52 PM [View respondent's answers](#)

Runoff
 3/14/2016 2:27 PM [View respondent's answers](#)

precipitation, temperature, discharge
 3/14/2016 12:33 PM [View respondent's answers](#)

2. How many dimensions does the data have?

Answer Choices	Responses
1 dimension	28.57% 2
2 dimensions	0.00% 0
3 dimensions	28.57% 2
n dimensions (n>3)	42.86% 3

3. Sequential color scheme:

Answer Choices	Responses
I have never used this color scheme	50.00% 3
I have used it for the data:(please specify)	Responses 50.00% 3

4. Diverging color scheme:

Answer Choices	Responses
I have never used this color scheme	28.57% 2
I have used it for the data: (please specify)	Responses 71.43% 5

5. Qualitative color scheme:

Answer Choices	Responses
I have never used this color scheme	57.14% 4
I have used it for the data:(please specify)	Responses 42.86% 3

6. Which of following uncertainty visualizations have you ever used?

Answer Choices	Responses
Line Chart	100.00% 7
Bar Chart	100.00% 7
Scatter Plot	100.00% 7
Box Plot	100.00% 7
Taylor Diagram	0.00% 0
Original Data Space (e.g. Map)	71.43% 5
Spaghetti Plot	42.86% 3
Other (please specify) Responses	28.57% 2

7. Given the visual variables reported in the pictures below the answers, which of the following have you ever felt helpful in representing uncertainty? (Ref: <http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=6327255>)

Answer Choices	Responses
Location	33.33% 2
Size	66.67% 4
Color Hue	66.67% 4
Color Value	50.00% 3
Color Saturation	33.33% 2
Orientation	0.00% 0
Grain	0.00% 0
Arrangement	0.00% 0
Shape	16.67% 1
Fuziness	16.67% 1
Transparency	33.33% 2

8. Which of the following problems have you ever met in your research? (Ref: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=7061479)

Answer Choices	Responses
Clutter: Overlap, e.g. overlaps of symbols on the map	83.33% 5
Clutter: Color mixing, e.g. background color is mixing with line colors in line chart	83.33% 5
Distortion: Scale inconsistency: e.g. inconspicuous Y-axis scale in two similar scatter plots	33.33% 2
Distortion: Projection error, e.g. errors in mapping 3d sphere to 2d space	33.33% 2
Comparison Complexity: Superposition overload, e.g. too many lines in a line chart	83.33% 5
Comparison Complexity: Lack of explicit encoding, e.g. when comparing similarity among data in multiple maps, Euclidean layout distances between maps are not usually considered besides the colors	16.67% 1
Communication Gap: Improper Legend: e.g. too many symbols in the legend	66.67% 4
Communication Gap: Missing Annotation: e.g. annotation for the outliers in the map would communicate the intent more effectively	16.67% 1

9. In your experience, which of the visual variables mentioned above have caused problems for you to understand the uncertainty information?

The problem we fact is typically that our space has too many dimensions.

4/15/2016 8:02 AM [View respondent's answers](#)

none

3/27/2016 11:26 AM [View respondent's answers](#)

Orientation, grain, arrangement

3/25/2016 9:33 AM [View respondent's answers](#)

Color hue and saturation, cross hatching, shape

3/23/2016 4:52 PM [View respondent's answers](#)

10. Can you mention an example where you had problems using the above visual variables or visualizations during uncertainty analysis?

scatter plot, where the uncertainty in large values dominate the visualizations.

4/18/2016 8:25 PM [View respondent's answers](#)

N/A

4/15/2016 8:02 AM [View respondent's answers](#)

Always; there are no good approaches, in my opinion, to adequately visualize uncertainty if multi-dimensional data

3/27/2016 11:26 AM [View respondent's answers](#)

Problems arise when I have to look too often at the legend, mostly when are present different symbols.

3/25/2016 9:33 AM [View respondent's answers](#)

Large range or number of intervals is hard to discern saturation, not enough hues or colors. When dimensions exceed 5 or 6 different shapes gets confusing,

3/23/2016 4:52 PM [View respondent's answers](#)

11. In your experience, which of the visual variables mentioned above helped you quickly understand the uncertainty information in your data?

box plot	4/18/2016 8:25 PM	View respondent's answers
N/A	4/15/2016 8:02 AM	View respondent's answers
None	3/27/2016 11:26 AM	View respondent's answers
Color and size	3/25/2016 9:33 AM	View respondent's answers
quickly? none.	3/23/2016 4:52 PM	View respondent's answers

Questions in Uncertainty Quantification

1. Which of the following metrics have you used in the quantification of uncertainty or comparison of ensemble members or simulation results?

<input type="checkbox"/> Metric 1	100.00%	7	
<input type="checkbox"/> Metric 2	57.14%	4	
<input type="checkbox"/> Metric 3	42.86%	3	
<input type="checkbox"/> Metric 4	100.00%	7	
<input type="checkbox"/> Metric 5	71.43%	5	
<input type="checkbox"/> Metric 6	100.00%	7	
<input type="checkbox"/> Metric 7	0.00%	0	
<input type="checkbox"/> Metric 8	0.00%	0	
<input type="checkbox"/> Metric 9	0.00%	0	
<input type="checkbox"/> Metric 10	71.43%	5	
<input type="checkbox"/> Metric 11	28.57%	2	
<input type="checkbox"/> Other not listed above (please specify)	Responses	0.00%	0

2. Are there two or three metrics that you have used more frequently? If so, can you explain why? You can explain your reasons based on data types, objects and distribution characteristics or the

specific analysis requirements, etc.

mean: for mitigation cost estimates. i actually think this is not a good measure due to a long tail in ensemble members. median: same as above, but this is a better measure for explaining general trend. 90% range:for cost data, we always need to convey how uncertain economic estimates are. we basically want to eliminate the really "weird" ones and report the full range.

4/18/2016 8:32 PM [View respondent's answers](#)

Quartiles are very frequently used in expressing ranges from integrated assessment models for simplicity.

4/15/2016 8:06 AM [View respondent's answers](#)

no

3/27/2016 11:28 AM [View respondent's answers](#)

3. Are there any uncertainty metrics which are very helpful in your area but not commonly used in other areas? If so, can you mention one of them and explain why?

N/A

4/15/2016 8:06 AM [View respondent's answers](#)

no

3/27/2016 11:28 AM [View respondent's answers](#)

Probability of detection False alarm rate

3/25/2016 9:41 AM [View respondent's answers](#)

4. Are there other specific problems related to the quantification of uncertainty that you would like to highlight? Problems can be related to data preprocessing (e.g. discretizing data for specific metrics such as entropy), justify full uncertainty boundary, the computation performance, accuracy, etc.

time-series data. explaining the data with median or mean value becomes problematic if the median member changes over time.

4/18/2016 8:32 PM [View respondent's answers](#)

It is quite difficult to quantify uncertainty in the multi-dimensional space associated with integrated assessment, particularly since many of the uncertainties have to do with human system behavior. The complexity of structural uncertainty (different models getting different answers) is also important and hard to handle.

4/15/2016 8:06 AM [View respondent's answers](#)

Noted previously

3/27/2016 11:28 AM [View respondent's answers](#)

Instrumental errors which could strongly influence the quality of dataset.

3/25/2016 9:41 AM [View respondent's answers](#)

5. Are you aware of the assumption of normal distribution in using moments, e.g. mean and standard deviation?

Answer Choices	Responses
Yes, I am aware of that	85.71% 6
No, I was not aware of that	0.00% 0
I don't have the experience of using moments in uncertainty analysis	14.29% 1

6. Are you aware of that moments are limited to portray the distribution modality? For example, shown as the figure below, three distributions have very close summary statistics but very different modality characteristics.

Answer Choices	Responses
Yes, I am aware of that	66.67% 4
No, I was not aware of that	16.67% 1
I don't have the experience of using moments in uncertainty analysis	16.67% 1

7. Are you aware of that the selection of parameters in estimating probability distribution will change the results?

Answer Choices	Responses
Yes, I am aware of that	100.00% 7
No, I was not aware of that	0.00% 0
I don't have the experience of estimating probability distribution.	0.00% 0

8. In a Bayesian framework, are you aware of the selection of prior distribution, likelihood function and posterior function would influence the results?

Answer Choices	Responses
Yes, I am aware of that	100.00% 6
No, I was not aware of that	0.00% 0
I don't have the experience of estimating probability distribution.	0.00% 0

9. Are you aware of the unit problem when using entropy to quantify uncertainty?

Answer Choices	Responses
Yes, I am aware of that	0.00% 0
No, I was not aware of that	16.67% 1
I don't have the experience of using entropy	83.33% 5

10. Are you aware of the importance of uncertainty boundary in quantifying uncertainty?

Answer Choices	Responses	
Yes, I am aware of that	71.43%	5
No, I was not aware of that	28.57%	2

Questions in Analyzing Uncertainty Quantification and Visualization

1. In your experience, how frequently do you think the uncertainty quantification results cannot be verified with your domain knowledge?

Answer Choices	Responses	
All of the results cannot be verified with my domain knowledge	20.00%	1
Half of the results cannot be verified with my domain knowledge	0.00%	0
Some of the results cannot be verified with my domain knowledge	60.00%	3
None of the results cannot be verified with my domain knowledge	20.00%	1

2. In your experience, how do you think the chosen visualization metrics helped you to reach the goals?

Answer Choices	Responses	
Effectively interpret results of the quantification for a domain expert	66.67%	4
Effectively present results of my research in papers or talks for peers in my domain	83.33%	5
Effectively convey information to stakeholders	50.00%	3

3. In your experience, which of the followings prevented you from fully understanding the outcomes of the application of an uncertainty metric that you used in the past?

Answer Choices	Responses	
Knowledge gaps: e.g., the uncertainty metric is hard to understand or it is misleading or insufficient to capture the information I need	60.00%	3
Less or inappropriate contextual information: e.g. loss of details	20.00%	1
Improper quantification approaches: e.g., the metric which can only be applied to continuous data is adapted on discontinuous data	20.00%	1
Visualization gaps: e.g. The visualization is hard to understand or be misleading because they are not frequently seen in your domain or conflicting with your common sense	80.00%	4
Other (please specify)	Responses	0.00% 0

4. In your experience, what benefits can you recall that you obtained from the uncertainty analysis? For example: I can gain a big picture of the data; I can characterize the uncertainty of a

crucial parameter of a numerical model; I can explain to stakeholders why predictions or estimation of certain parameters/variables are uncertain.

I can explain to stakeholders why predictions or estimation of certain parameters/variables are uncertain. and explaining just how large the uncertainty space is.

4/18/2016 8:34 PM [View respondent's answers](#)

This is kind of simple, but a better understanding of the underlying uncertainty or at least a sense of possible outcomes.

4/15/2016 8:10 AM [View respondent's answers](#)

unknown "unknowns" are sometimes revealed.

3/27/2016 11:31 AM [View respondent's answers](#)

When explaining why model predictions are not perfectly correct.

3/25/2016 9:50 AM [View respondent's answers](#)

5. In your experience, are you aware of the uncertainty propagation during the uncertainty quantification and visualization process? Uncertainty can be propagated through loss of data while processing data or loss of details while visualization, etc.

Answer Choices	Responses
Yes	33.33% 2
No	66.67% 4

6. If Yes above, did you involve them into your analysis? Can you mention an example?

propagation can create cones or paths of uncertainty

3/23/2016 5:00 PM [View respondent's answers](#)
