A CRITICAL EVALUATION OF THE APPLICATION OF NATURAL HAZARD AND CLIMATE MODELS

A dissertation submitted by

Brent Boehlert

In partial fulfillment of the requirements for the degree of

Doctor of Philosophy

In

Civil and Environmental Engineering

Tufts University

February 2015

Committee:

Professor Richard M. Vogel, Tufts University

Professor Kenneth M. Strzepek, Massachusetts Institute of Technology

Professor Susan Solomon, Massachusetts Institute of Technology

Professor Laurie G. Baise, Tufts University

Dr. Nicholas C. Matalas, formerly U.S. Geological Survey

ABSTRACT

Planning for natural hazards and climate change requires that we develop the best possible understanding of future hydroclimatic conditions. Modeling tools have become essential in meeting this goal, but selecting the most appropriate models and distilling actionable findings from their outputs is still a major challenge. In this research, we investigate this issue from two perspectives: (1) the evaluation of hypothesized natural hazard models, with a focus on predicting flooding events (Chapters 1 and 2); and (2) patterns of agreement and uncertainty in water availability projections derived from a wide array of climate model ensembles (Chapter 3). The first chapter evaluates the appropriateness of traditional metrics of 'goodness-of-fit' as measures of the performance of a hypothesized natural hazard model (i.e., the applicability of a selected probability density function). We find that goodness-of-fit can be quite misleading, causing us to reject the correct model and generate potentially large errors in design event (e.g., 1000-year flood) estimation in the process. We propose an alternative metric that gives a more balanced assessment of goodness-of-fit. In the second chapter, we introduce a property called transformational concordance. This property can be used to evaluate whether a hypothesized model and its distributional behavior are consistent with observations. Through our analysis of concordance, we reveal systematic bias in GEV parameter estimation, which is cause for significant concern given the wide application of the model for flood and other natural hazard modeling. The third chapter focuses on improving our understanding of the timing, location, and magnitude of climate change impacts on water needs and availability. Using a wide range of recently available climate model ensembles, we explore the spatial and temporal patterns of inter-model agreement and uncertainty in projected river runoff,

irrigation water requirements, and basin storage yield. Cost estimates of adapting global water supply systems are developed for each ensemble, and implications for water management are discussed.

ACKNOWLEDGEMENTS

I would like to thank several people who provided invaluable assistance, insights, and support, during the process of conceptualizing and drafting this dissertation. First, my committee: Richard Vogel, my advisor, provided continual and invaluable feedback throughout my time at Tufts; Ken Strzepek, my unofficial advisor and mentor from MIT, for sparking my interest in this field, advancing my career, providing continual insights, and convincing me to pursue the Ph.D.; Susan Solomon, from MIT, for introducing me to a new way of thinking about climate change; Laurie Baise, from Tufts Engineering, for providing key insights on the statistical characteristics of natural hazards; and Nick Matalas, for providing the original seed of the concordance idea, and for key comments on drafts.

I would also like to thank Jim Neumann and Bob Unsworth, both Principals at Industrial Economics, for their patience, support, and understanding these past three years as I worked under them full time while pursuing the Ph.D. My appreciation goes out to the students in the Tufts Hydrostatistics group under Professor Vogel, most notably Will Farmer and Laura Read, who provided invaluable feedback on research ideas and draft presentations. I also wanted to thank Bob Hirsh of the U.S. Geological Survey, Diane Ivy of MIT, and clients at the World Bank who provided access to datasets that were critical in the completion of this dissertation.

Finally, thanks to my parents, George and Susan Boehlert, for their kindness, generosity, and love throughout the process, and to my brother Brooks for his friendship and his service to our country over these last 11 years.

TABLE OF CONTENTS

A CRITICAL EVALUATION OF THE APPLICATION OF NATURAL HAZARD AND CLIMATE MODELS

INTRODUCTION

This dissertation is composed of three chapters that focus on the evaluation and application of natural hazard and climate change models. Modeling tools have become essential in developing the best possible understanding of future hydroclimatic conditions, but selecting the most appropriate models and distilling actionable findings from their outputs is still a major challenge. In this dissertation, we investigate this issue from two perspectives: (1) the evaluation of hypothesized natural hazard models, with a focus on predicting flooding events (Chapters 1 and 2); and (2) patterns of agreement and uncertainty in water availability projections derived from a wide array of climate model ensembles (Chapter 3). The first two chapters illustrate the challenges of selecting and fitting statistical models using both observed flooding data from the U.S. Geological Survey (USGS) and Monte Carlo analysis to investigate questions that require numerical experimentation. The last chapter, on the other hand, investigates patterns of change in future global water resource availability using a suite of water demand and availability models that are driven by climate changes from a set of 220 General Circulation Models (GCMs). As a global study that relies on simulation rather than statistical models, the research presented in Chapter 3 is considerably more data intensive than research presented in Chapters 1 and 2.

Chapter 1 is titled *Goodness-of-Fit Can Be Misleading* and is co-authored with Professor Richard Vogel of Tufts University. This chapter evaluates the appropriateness of traditional metrics of 'goodness-of-fit' as measures of the performance of a hypothesized natural hazard model (i.e., the applicability of a selected probability density function). When modeling the relationship between the frequency and magnitude of a natural hazard, there is a need to select and fit a frequency distribution

of the hazard of interest which mimics the unknown parent distribution. Hypothesis tests and goodness-of-fit metrics are often used to assess how well a hypothesized probability density function (pdf) fits observations. The goal of such evaluations is to ensure that the chosen probability model reproduces various important properties associated with the correct pdf. In our investigation of goodness-of-fit, we focus on the Probability Plot Correlation Coefficient (PPCC), which is based on a quantile-quantile (Q-Q) probability plot and is perhaps the most widely-used goodness-of-fit metric for natural hazard model evaluation. Among 200 flood series from rivers in the U.S. with very long records, we note that it was generally only those rivers that experienced extraordinary floods that had consistently very low PPCC values, leading one to question the goodness-of-fit of commonly used pdfs to those samples with the most critical flood experience. Using Monte Carlo experiments, we find that when (a) a particular sample happens to contain a high outlier or (b) we introduce additional information about the true underlying model, the goodness-of-fit declines rather than increases. As a result, we find that PPCC can be quite misleading, causing us to reject the correct pdf and generate potentially large errors in design event estimation. Further experiments lead us to observe systematic errors in quantile prediction when the maximum sample value diverges significantly from the n-year return period event, and to observe that witnessing such a divergence may provide a warning of possible design errors. We attempt an adaptive strategy involving replacement of the largest observation(s), and although that approach has promise for its ability to reduce root mean square error (RMSE), we argue that generating better estimates of skewness using regional methods is likely to be the best approach forward for reducing both bias and RMSE associated with design events. To conclude this chapter, we introduce PPCC goodness-of-fit

metrics based on probability-probability (P-P) plots, which are shown to give a much more balanced assessment of the goodness-of-fit of a hypothesized pdf.

Chapter 2 is titled *Transformational Concordance of the Generalized Extreme Value Distribution*, and is co-authored by Professor Richard Vogel and Dr. Nicolas Matalas, who is retired from the U.S. Geological Survey. In this chapter, we first note that for a scientific hypothesis to be consistent with observations, it must exhibit concordance with observed data across space, time, as well as across functional transformations. The chapter focuses on the problem of selecting a probability distribution for modeling natural hazards that exhibits *transformational concordance* with observed data. A data series is transformationally concordant if its properties under each transformation (e.g., real space to log space) are consistent with our theoretical understanding of how those properties change across transformations. For example, in real space, if a sequence of observations is assumed lognormal and if logarithms exhibit zero skewness, the lognormal distribution in real space and the normal distribution in log space would be transformationally concordant. Exploring one of the most widely-used models for flood frequency analysis, we use 200 long flood records in the U.S. to consider the transformational concordance of the Generalized Extreme Value (GEV) model. If observations are GEV, then their adjusted logarithms should follow a Gumbel pdf. The GEV pdf exhibited the best goodness-of-fit among several alternative pdfs for the 200 flood series using traditional metrics. However, the flood series were found to be transformationally discordant under the GEV hypothesis, and the GEV model was found to significantly underestimate the frequency of extremely large design events. Both the discordance and design event bias appear to be attributable to systematic bias in GEV parameter estimation that is largely addressed by application of regional skew

estimates. To conclude Chapter 2, we identify the potential for a GEV hypothesis test based on transformational concordance.

Chapter 3 is titled *Water under a Changing and Uncertain Climate: Lessons from Climate Model Ensembles* and is co-authored by Professors Susan Solomon and Kenneth Strzepek of the Massachusetts Institute of Technology. The third chapter is motivated by the fact that climate change and rapidly rising global water demand are expected to place unprecedented pressures on already strained water resource systems. Successfully planning for these future changes requires a sound scientific understanding of the timing, location, and magnitude of climate change impacts on water needs and availability – not only average trends, but also interannual and decadal variability and associated uncertainties. In recent years, two types of large ensemble runs of climate projections have become available, those from groups of more than 20 different climate models, and those from repeated runs of several individual models. These provide the basis for novel probabilistic evaluation of both climate change and the resulting effects on water resources. Using a range of available climate model ensembles, this Chapter explores the spatial and temporal patterns of inter-model agreement and uncertainty in projected river runoff, irrigation water requirements, and basin storage yield. Cost estimates of adapting global water supply systems are developed for each ensemble. We observe strong spatial patterns of multiple-ensemble agreement and disagreement in both precipitation and runoff trends. Regions with robust cross-ensemble drying trends include southern Europe, northern Africa, western Australia, southern Africa, eastern Brazil, and northern Mexico; and wetting trends occur in the northeastern US, Canada, northern regions of the globe, and parts of southeast Asia. Relative to changes in precipitation, we find that patterns of changes in basin yield are both magnified and

systematically drier due to the dependence of river runoff on land surface dynamics and temperature. Due to the temporally integrating effects of basin yield and monetary discounting, the costs of maintaining historical yields show still stronger patterns of agreement across GCM ensembles, particularly when focusing on agreement within broad geographic regions of the globe. We recommend future research that evaluates patterns of GCM ensemble agreement and disagreement under a broader assessment that integrates projected changes in irrigation water requirements into an analysis of basin water supply, under a future that incorporates rising food demands, population increases, and environmental flow requirements.

GOODNESS-OF-FIT CAN BE MISLEADING

Brent Boehlert[1,2] and Richard M. Vogel[1]

Author affiliations: 1. Tufts University, Department of Civil and Environmental Engineering, Medford, MA
2. Industrial Economics, Inc., Cambridge, MA

ABSTRACT

When modeling the relationship between the frequency and magnitude of a natural hazard, there is a need to select and fit a frequency distribution of the hazard of interest which mimics the unknown parent distribution. Hypothesis tests and goodness-of-fit metrics are often used to assess how well a hypothesized probability density function (pdf) fits observations. The goal of such evaluations is to ensure that the chosen probability model reproduces various important properties associated with the correct pdf. The probability plot correlation coefficient test (PPCC) which is based on a quantile-quantile (Q-Q) probability plot is a widely used tool for evaluating the goodness-of-fit of a hypothesized pdf to a sample of observations. Among 200 flood series from rivers in the U.S. with very long records, we note that it was generally only those rivers that experienced extraordinary floods that had consistently very low PPCC values, leading one to question the goodness-of-fit of commonly used pdfs to those samples with the most critical flood experience. We further document, using Monte Carlo experiments, that the PPCC metric is very sensitive to observations that appear to be high outliers, and that when we introduce additional information about the true underlying model, the goodness-of-fit declines rather than increases. As a result, the metric can be quite misleading, causing us to reject the correct pdf in situations when it would have been

very important to accept it. Further experiments lead us to observe systematic errors in quantile prediction when the maximum sample value diverges significantly from the n-year return period event, and to observe that witnessing such a divergence may provide a warning of possible design errors. We attempt two adaptive strategies, and find that although an approach involving replacement of the largest observation(s) may have promise, particularly in reducing root mean square error (RMSE), generating better estimates of skewness using regional methods is likely to be the best approach forward for reducing both bias and RMSE associated with design events. We introduce PPCC goodness-of-fit metrics based on probability-probability (P-P) plots, which are shown to give a much more balanced assessment of the goodness-of-fit of a hypothesized pdf.

1. INTRODUCTION

In modeling the relationship between frequency and magnitude of natural hazards, we seek to ensure that the chosen probability density function (pdf) and associated model parameters reflect, as closely as possible, the unknown parent distribution. There are many pdfs that could represent a given physical process, and an infinite number of parameter combinations that map that pdf to the specific context being modeled. For example, in the context of flood frequency analysis, annual maximum flow series have been shown to be reasonably well modeled by the Log-Pearson Type III (LP3), Generalized Extreme Value (GEV), or three parameter Lognormal (LN3) probability models over broad geographical regions (for a review, see Tables 3 and 4 in Vogel and Wilson 1996, Kidson and Richards 2005, El Adlouni et al. 2008, and Table 2 in Gubavera 2011). By selecting a pdf and set of parameters that most accurately represent the watershed under consideration, we ensure that the outputs of our frequency analyses will also be as accurate as possible. This means that estimation of the magnitude of design events (e.g., 1,000-year flood) will be as close as possible to the "true" values that would occur in the physical system itself.

Hypothesis tests and goodness-of-fit metrics are often used to assess how well a postulated pdf fits observations. Perhaps the most widely used goodness-of-fit statistic for distributional selection is known as the probability plot correlation coefficient (PPCC) which has now been developed for a very broad range of hypothesized pdfs (see Heo et al. 2008 for a recent review). Although the PPCC goodness-of-fit metric is now widely used in the context of frequency analysis of hydrologic and other extreme events, natural hazards time series tend to exhibit outliers and other characteristics that pose unique challenges for the interpretation of such metrics (IACWD 1982, Gen and Koehler

1990, Laio et al. 2010). For the purposes of this study, outliers are observations that appears to deviate considerably from the other observations within a given sample (Grubbs 1969). Methodologies have been developed for addressing outliers in natural hazard frequency analysis, most recently by Cohn et al. (2013), who focus on mitigating the influence of low outliers in pdf parameter estimation. Their method improves the fit of the frequency distribution to observed high flows, making the flood frequency analysis more robust.

Using actual flood series, we document systematic bias associated with the PPCC goodness-of-fit metric, as well as estimates of design events resulting from samples that appear to contain high outliers or very low maximum values. These flood series are used to expose systematic bias associated with our interpretation of goodness-of-fit as well as our ability to estimate design events such as the 1,000-year flood. We rely on Monte Carlo experiments to document our findings in a more definitive manner using the three probability models noted above. Next, we document the value of potential adaptive approaches for identification and treatment of outliers to reduce bias associated with both goodness-of-fit measures and design flood estimates. Lastly, we recommend an alternative goodness-of-fit PPCC metric that provides a more balanced assessment of the ability of a hypothesized pdf to mimic flood series, regardless of whether they exhibit outliers or not.

2. GOODNESS-OF-FIT VERSUS HYPOTHESIS TESTS

Comparing the goodness-of-fit of a range of probability models to observations is a standard practice in natural hazard frequency analysis. To illustrate the bias in

goodness-of-fit measures when outliers are present, we focus on the PPCC metric, which is the correlation between the ranked observations and the fitted quantiles associated with their ranked plotting positions on a quantile-quantile (Q-Q) plot. PPCC tests have been developed for many probability models, including the normal, lognormal and Gumbel distributions (Vogel 1986), and the Pearson type 3 (Vogel and McMartin 1991) and others (see Heo et al. 2008). For two parameter pdfs with fixed shape, the value of a PPCC statistic is only affected by sample size, choice of plotting position, and significance levels. However, for 2-parameter pdfs with varying shape (e.g., Gamma) and for all 3-parameter pdfs, the PPCC goodness-of-fit metric depends additionally upon the shape parameter of the distribution.

A formal hypothesis test should not depend upon estimated parameters of the distribution. When a PPCC hypothesis test depends upon the parameter estimates of a pdf, in addition to sample size and significance level, such tests tend to have low power compared to similar tests for distributions with fixed shape (see Vogel and McMartin 1991). Instead, in such situations when the PPCC statistic depends on estimates of model parameters, the statistic becomes solely a goodness-of-fit metric which cannot be formally used to either accept or reject a scientific hypothesis. Here we focus on goodness-of-fit evaluations that tend to be relevant only for such complex pdfs with varying shape parameters, such as the GEV, LP3, LN3 and a variety of other pdfs commonly used to model natural hazards. Figure 1 provides an example of GEV Q-Q plots for four approximately 100-year series of annual maximum observed flows. Figure 1 illustrates the ordered observations along the horizontal axis, and the theoretical quantiles of the GEV distribution along the vertical. We estimate GEV model parameters and generate random samples using the method of L-moments (see Hosking

1990 and Hosking and Wallis 1997), as configured within the R software package lmomco (Asquith 2011). Also shown is the correlation between the two axes, known as the PPCC value. Importantly, each of these flood series contains an event that is modeled as having a return period greater than 1,000-years; this event is apparent as an off-diagonal data point on the right side of each graph.

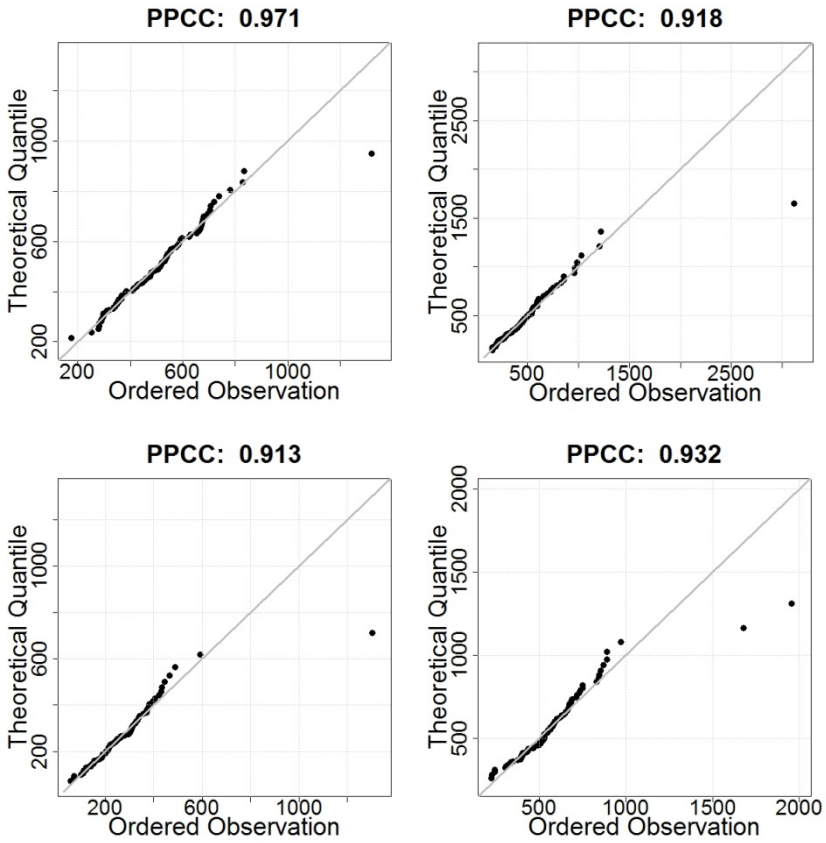

Figure 1. Examples of Q-Q plots from four observed maximum annual flow series (in cubic-feet per second) of approximate length 100. Ordered observations along the horizontal axis, and the theoretical quantiles of the GEV distribution are along the vertical. These four series each contain at least one event that is modeled as having a return period greater than 1,000-years.

The analysis in Figure 1 was repeated for the 200 annual maximum instantaneous streamflow time series recently assembled, analyzed and summarized by Hirsh and Ryberg (2012). These 200 annual maximum flood series have record lengths ranging from 85 to 126 years, with an average of 94.2 years, and are located across the

conterminous U.S. The resulting values of the PPCC statistics for five common distributions for modeling annual maximum flows, including GEV, are summarized using boxplots in Figure 2. We also highlight PPCC values (as black triangles) that correspond to rivers whose largest flood observation was greater than the estimated 1,000-year flood. For each hypothesized pdf, L-moment estimators were used to estimate the pdf model parameters and to estimate the average return period associated with the largest observation. Here we observe the almost uniform phenomenon that, for all pdfs considered, those rivers with low values of the PPCC goodness-of-fit statistic correspond to exactly those rivers in which the largest observed flood was greater than the estimated 1,000-year event. This result is rather compelling, because it indicates that the goodness-of-fit of each distribution is lowest for those samples that have experienced extraordinary floods, perhaps the most important experience on record for flood frequency analysis. The results in Figure 2 lead to only one of the following two conclusions: (1) the probability models considered in Figure 2 perform poorly for flood samples that have extraordinary flood experience, or (2) the PPCC goodness-of-fit statistic, based on a Q-Q plot, is misleading for flood samples that have extraordinary events. To ascertain which of these conclusions is correct, we perform several Monte Carlo experiments in the following sections.

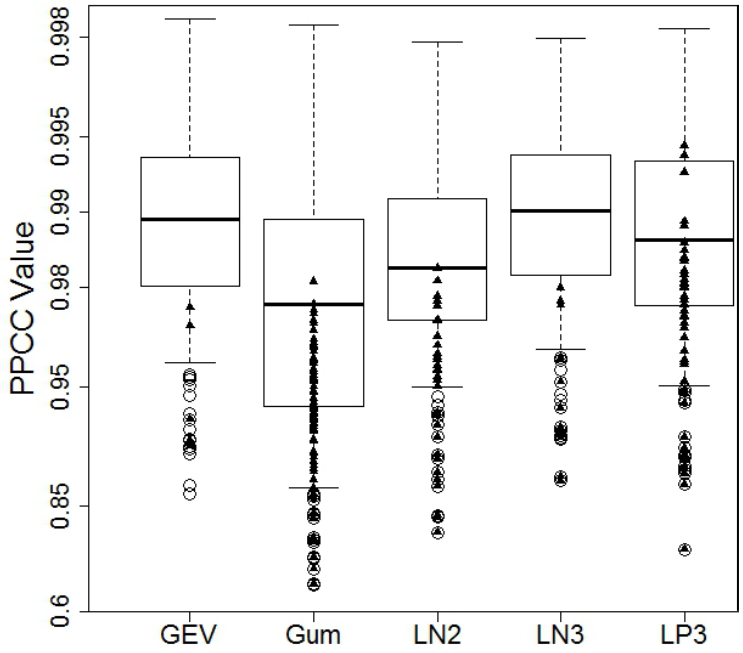

Figure 2. Boxplots of PPCC statistics using Q-Q plots for five distributions. Each boxplot includes 200 stations. Circles are outliers, and triangles are events with greater than 1,000-year return periods. GEV is Generalized Extreme Value; "Gum" is Gumbel; "LN2" is 2-parameter lognormal; "LN3" is 3-parameter lognormal; and "LP3"is Log-Pearson Type III.

3. THE PPCC GOODNESS-OF-FIT STATISTIC CAN BE MISLEADING

The previous section suggests that the PPCC goodness-of-fit statistic based on a Q-Q plot may be misleading for samples that include extraordinary flood events. To investigate this issue, we examine the relationship between PPCC estimates that rely on sample estimates of pdf model parameters versus PPCC estimates that rely on the true underlying pdf model parameter values. Initially, our intuition led us to hypothesize that goodness-of-fit as measured by the PPCC statistic would be higher when the true pdf parameter values are used instead of sample estimates of pdf parameters. To test

this hypothesis, we generate 100,000 random samples of length 100 from the three pdfs used most widely to model floods: GEV, LP3, and LN3. We also ran these experiments using sample sizes of 25 and 50, but the conclusions were unaffected, thus we only report results here for sample sizes of 100. The pdf model parameters needed to generate the random samples are fixed for each distribution, and are taken as the "true" parameter values of each model. We then estimate PPCC values for each sample by computing the correlation between the ordered observations and their theoretical values, which we evaluate based on the sample and true parameters. For the GEV, LP3, and LN3 distributions, we use the Gringorten, Blom, and Weibull plotting positions, respectively, based on recommendations by Vogel (1986) and Vogel and McMartin (1991) and others.

Figure 3 illustrates that over two-thirds of the 100,000 samples from each of the three distributions had lower PPCC values when using the true parameters than when using the estimated parameter values. Apparently, incorporation of more information about the true nature of the probability model into the PPCC calculation results in a reduction rather than an improvement in the goodness-of-fit for all three of the distributions considered. Figure 3 provides evidence in support of our primary hypothesis, that is, goodness-of-fit can be misleading. In each case illustrated in Figure 3, we assume that the hydrologist has chosen the correct model, thus one would expect goodness-of-fit to improve as more information is available concerning the values of the true model parameters. Instead, we observe that 'fitting' each pdf to the observations, results in a 'kind of' model tuning, where the estimated model parameters act to make the model 'adhere' more closely to the observations. Thus, the PPCC goodness-of-fit statistic is misleading in such instances, because it leads us to conclude the model fits

better than it actually does. This phenomenon was demonstrated by Vogel and McMartin (1991) when they showed that when constructing an LP3 probability plot, sample estimates of the skewness of the logarithms act to create more linear probability plots than one would expect, leading one to accept the LP3 alternative more often than one should. In other words, Vogel and McMartin (1991) found that LP3 probability plots tended to look more linear than they should, leading the analyst to accept the LP3 model more often than they should, due in part to its extraordinary flexibility and ability to 'adhere to' or 'mimic' the behavior of the observations.

Goodness-of-fit is about assessing the ability of a model to mimic the observations, so that better fit leads to better mimicry of the behavior of the observations. What our experiments indicate is that this 'better mimicry' of the observations is not necessarily consistent with our goal, which is identification of the true underlying parent pdf of a flood series.

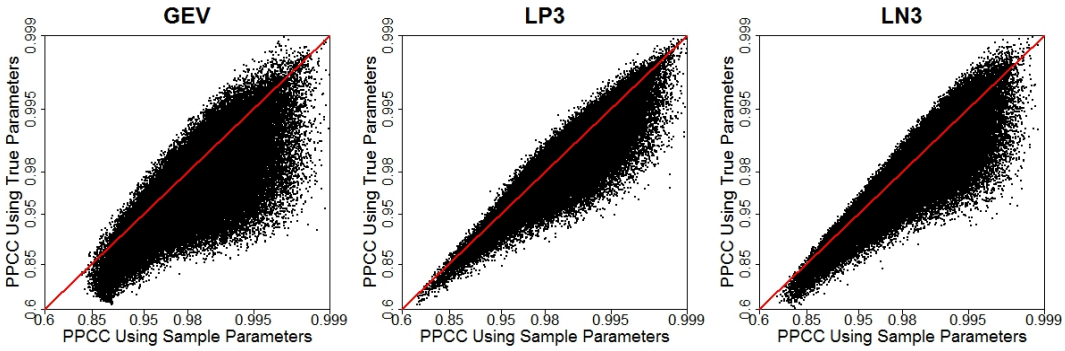

Figure 3. The Goodness-of-Fit index PPCC based on sample parameters versus PPCC using true parameters for the GEV, LP3 and LN3 distributions, based on 100,000 samples, each of length 100

4. THE MOST IMPORTANT NATURAL HAZARD OBSERVATIONS TEND TO CONFOUND OUR ABILITY TO SEE THE TRUE MODEL

The experiments summarized in Figure 3 led us to question the value of the PPCC statistic for assessing the 'goodness-of-fit' of alternative pdfs to flood samples. Here we explore this issue further with the goal of determining what causes the PPCC goodness-of-fit statistic based on Q-Q plots to be misleading. Using the results of the previous experiment, Figure 4 plots the average return period of the maximum observation in each of the 100,000 generated samples versus the PPCC statistic based on sample estimates of the model parameters. Samples whose largest observations have extremely high average return periods tend to have the lowest PPCC values. Yet these samples are the most critical samples in the sense that they exhibit extreme flood experience that should be crucial for future flood hazard planning. Extremely large floods lead to very low values of PPCC due to the nature of the PPCC computation. The PPCC based on a Q-Q plot is defined as the correlation between the ordered observations and an estimate of the ordered observations based on the fitted pdf. The correlation coefficient is a measure of linearity, which is known to be heavily influenced by observed values that are very far away from their fitted counterparts. When the maximum observation happens to have a very small or very large average return period, its magnitude tends to be very far away from any of the fitted quantiles based on the hypothesized distribution, and exerts large influence on the correlation between theoretical and observed quantiles.

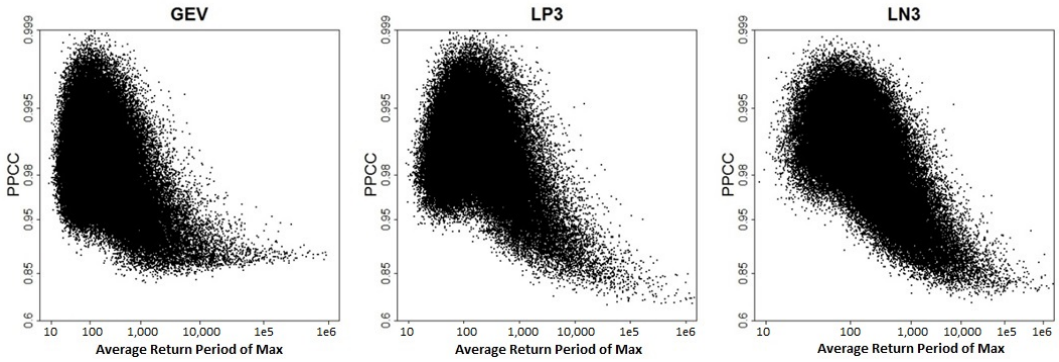

Figure 4. Average return period associated with maximum sample value (years) versus PPCC value based on sample parameters for the GEV, LP3, and LN3 distributions; based on 100,000 samples of length 100

The reduction in values of the PPCC statistic is even more striking when the fitted pdf is based on the true model parameters as is shown in Figure 5. In Figure 5, the PPCC values peak in samples that contain a maximum value with an average return period which is roughly equal to the length of the samples used to fit the distribution. Again, this is because of the configuration of the Q-Q plot, which allocates n plotting positions evenly in probability space. In this example, the 100^{th} plotting position corresponds approximately to the 1 in 100 event, so if such an event exists in the sample, the upper right end of the Q-Q plot is fixed near the unity line. Because this maximum value has the greatest leverage in the Q-Q plot, if it is greater or less than the 1 in n event, the PPCC value will decline.

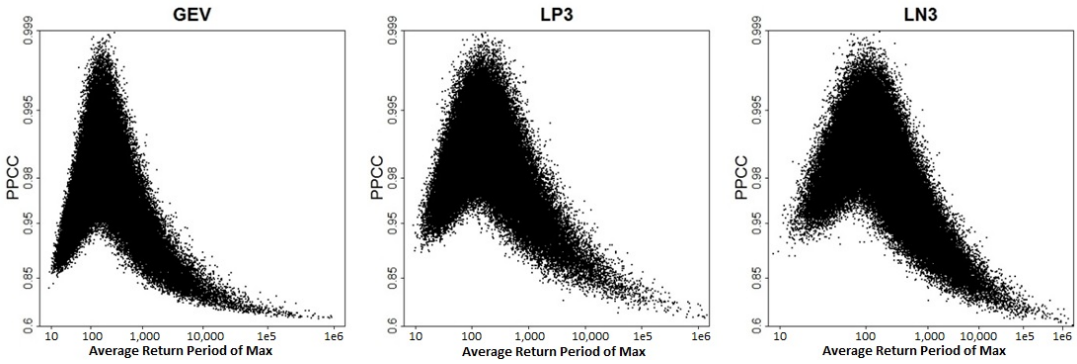

Figure 5. Average return period of maximum sample value (years) versus PPCC value based on true parameters for the GEV, LP3, and LN3 distributions; 100,000 samples of length 100

Figure 6 illustrates the difference between PPCC values generated using sample and true theoretical quantiles. As we saw earlier in Figure 3, the PPCC corresponding to the fitted pdf is generally greater than the PPCC corresponding to the true pdf. The differences are largest as the exceedence probability of the maximum sample value moves in either direction away from the T=100 year =n year event. In the case of the GEV model, for example, the difference in PPCC values can be greater than 0.15 in samples that contain observations with very small maximum return periods, and nearly 0.25 within samples containing the extremely large maximum return periods. Instances where the difference between PPCC values using sample and true parameters are greatest correspond directly with samples that contain events we care about most.

The experiment displayed in Figure 4 indicates that samples containing the largest natural hazards tend to perform most poorly from a goodness-of-fit perspective, and in Figure 6, that by introducing information about the true properties of these samples, we see the largest reductions in goodness-of-fit. We next inquire into the implications of this finding if models are chosen based on goodness-of-fit.

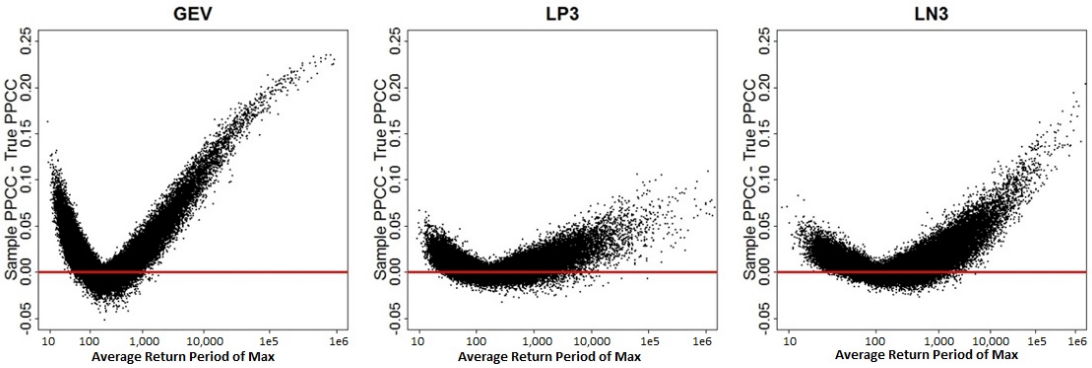

Figure 6. Average return period of maximum sample value (years) versus the difference in PPCC value based on sample and true parameters for the GEV, LP3, and LN3 distributions; 100,000 samples of length 100

5. USING GOODNESS-OF-FIT TO CHOOSE AMONG MODELS

Given the possibility that goodness-of-fit can be misleading, what are the consequences of employing the PPCC goodness-of-fit statistic for selecting among candidate probability models? That is, when using PPCC as a sole criterion for choosing a pdf, how often is an incorrect model selected, and how much larger are the quantile prediction errors stemming from that selection? As an illustration, we employ Monte Carlo analysis to evaluate how frequently the LP3 model is chosen over GEV when the underlying model is GEV. We generate 100,000 random samples of lengths 15, 25, 50, and 100 from a GEV distribution with a fixed set of parameters, and then use the PPCC to assess and compare the goodness-of-fit of the fitted GEV and LP3 models. The GEV and LP3 parameters are estimated for each sample using the method of L-moments, and then the PPCC values are calculated as the correlation between ordered GEV observations and the theoretical quantiles from the GEV and LP3 sample parameters (based on the Gringorten and Blom plotting positions, respectively). This allows us to

determine how often the LP3 PPCC value exceeds the GEV PPCC value when we know the underlying model is GEV.

In Figure 7, we plot the difference between PPCC values for GEV and LP3, where positive values indicate GEV has better goodness-of-fit than LP3, and negative values mean LP3 fits better than GEV. Regardless of sample length, the LP3 PPCC is higher than the GEV values in approximately half of samples. Thus the incorrect pdf, LP3, is selected in 57% of the cases for samples of size 15, but even for sample sizes of 100, the pdf is misspecified in 40% of samples. This suggests that the PPCC goodness-of-fit statistic based on a Q-Q plot may not be the most reliable method for choosing a probability distribution to model an individual sample.

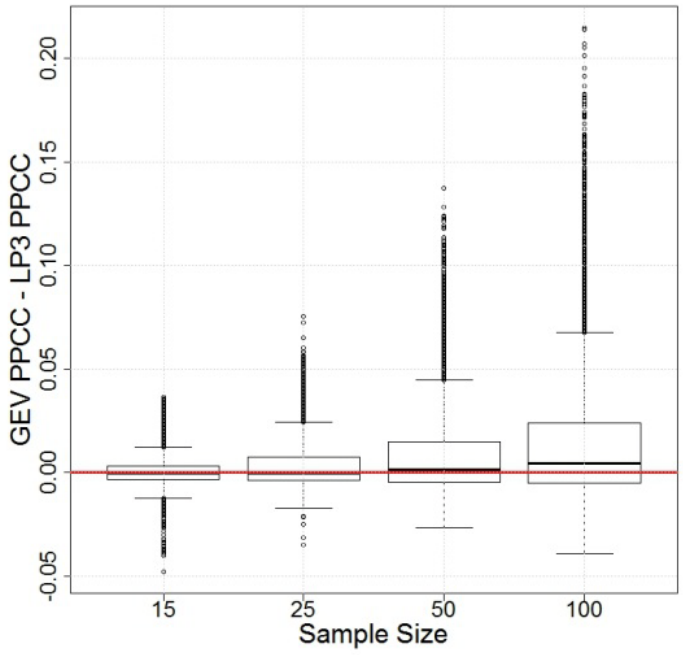

Figure 7. Boxplots of differences between PPCC of GEV and LP3 based on 100,000 synthetically generated GEV samples of length 15, 25, 50, and 100

6. RELATIONSHIP BETWEEN DESIGN ERROR AND MAXIMUM SAMPLE VALUE

The analysis above suggests that the largest observation in a sample can have a profound effect on our ability to discern and/or specify the correct pdf and associated parameter values. Here we further explore how the largest observation can in turn lead to systematic prediction errors by investigating the relationship between design event prediction errors and maximum sample value. Again we generated 100,000 samples of length n=100 years from the same GEV model used previously. Figure 8 illustrates the percentage error associated with predicting the magnitude of a 1,000-yr event, as a function of the average return period of the largest observation (top set of plots) and of the PPCC goodness-of-fit statistic (bottom set of plots) for the GEV, LP3, and LN3 pdfs. We observe that there is a systematic underestimation in quantile prediction using sample estimates of GEV model parameters when the maximum sample value has a low average return period and a systematic overestimation when the maximum sample value has a high average return period. Interestingly, there is a less clear relationship between the goodness-of-fit metric PPCC and prediction error as is shown in the bottom row of graphs in Figure 8. We conclude from Figure 8 that knowledge of goodness-of-fit cannot provide us with warning regarding systematic bias in model predictions. Nevertheless, witnessing a flood event with an average return period much larger or smaller than the record length n, does provide some warning regarding possible over or under design errors, respectively.

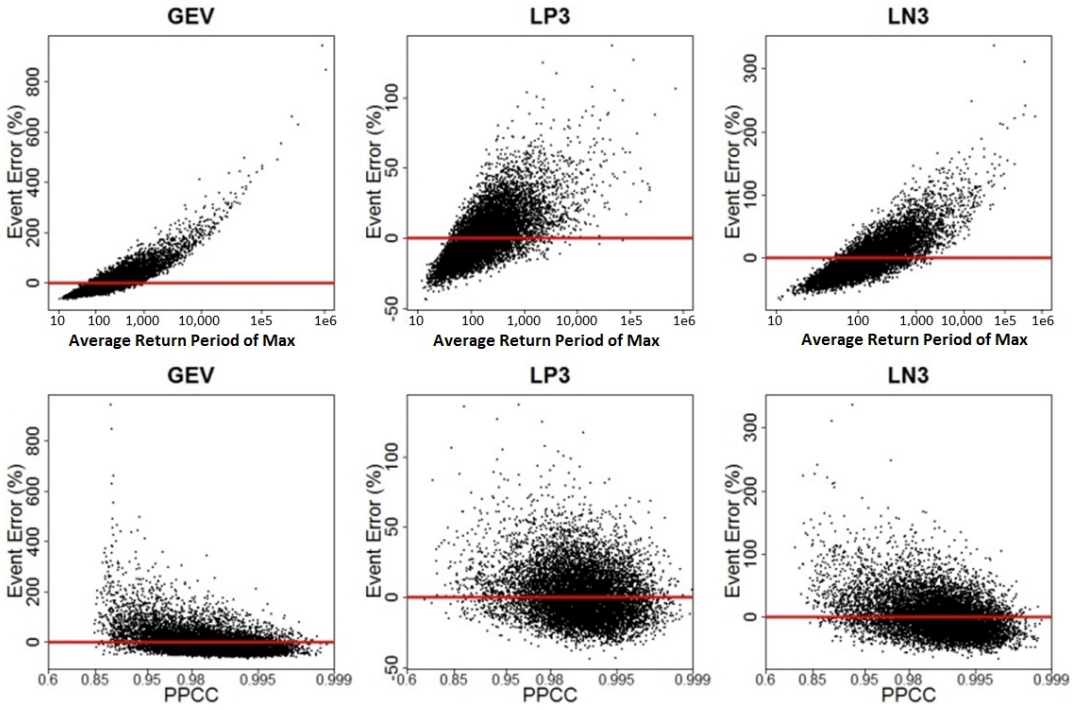

Figure 8. Average Return Period of maximum sample value (years) versus % quantile prediction error (top graphs) and PPCC versus % quantile prediction error (bottom graphs) for the GEV, LP3, and LN3 distributions; 100,000 samples of length 100

In the previous Figures, GEV samples all had fixed values of κ equal to -0.6. Figure 9 expands on the above results for the GEV distribution, by comparing the average return period of the maximum sample value and the systematic error associated with the estimated 1,000-year design event based on a range of κ and L-cv values. On the left are results based on sample estimates of GEV model parameters for κ ranging from -0.6 to 0.2, and L-cvs ranging from 0.2 to 0.6. On the right are an identical set of plots, but generated assuming the true value of κ is known. Knowing the true value of κ removes the bias introduced by the occurrence of maximum sample values that happen to have either very low or very high average return periods.

Figure 8 and Figure 9 indicate that we observe systematic errors in quantile prediction when the maximum sample value diverges significantly from the n-year return period event, and that witnessing such a divergence may signal possible design

errors. Although we also find that knowing the true value of κ removes these errors, given that in practice, we cannot know the true value of κ, is there an adaptive strategy that would reduce the bias in design event estimation introduced by the largest events? We next consider this question.

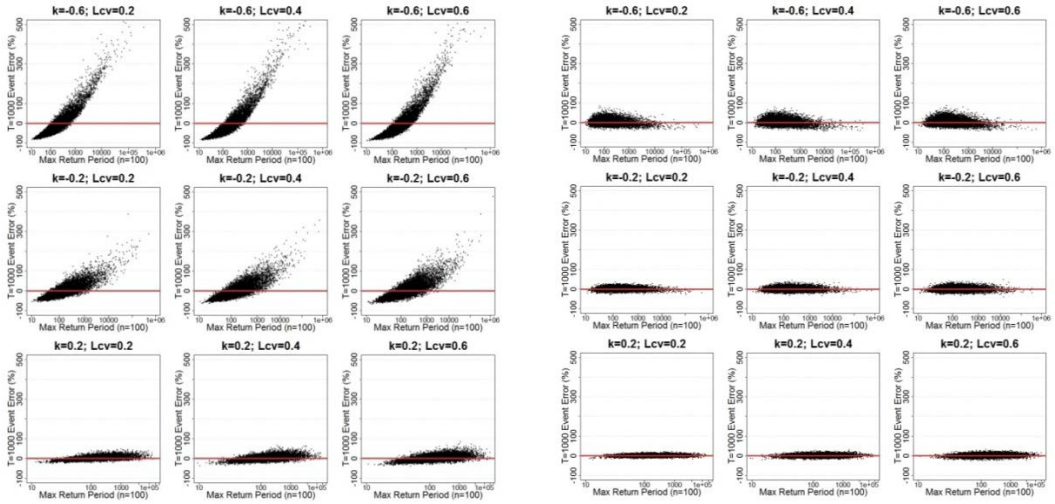

Figure 9. Average Return Period of maximum sample value versus % quantile prediction error for the GEV distribution (at-site on left and known κ on right); 10,000 samples of length 100 with mean = 1.

7. POTENTIAL ADAPTIVE STRATEGIES TO REDUCE DESIGN EVENT BIAS

To illustrate the influence of the largest observations and point toward possible adaptive strategies, we consider the effect of removing and/or replacing the largest observation on bias and root mean square error (RMSE) of design events. By removing the largest observation, we evaluate how bias and RMSE are affected had we not witnessed the largest observation and estimated pdf parameters and design events using that truncated sample. We also replace the largest observation with the expected value of the largest observation using the Gringorten plotting position in the GEV quantile function. Note that replacing the largest value by its expected value is

analogous to the use of the probability plot regression method for censored observations summarized by Helsel and Hirsch (2002, see chapter 13) and originally introduced by Travis and Land (1990).

Figure 10 displays the results of this inquiry. The first three bars and boxplots in each plot correspond to bias and RMSE associated with estimates of the 100-year flood using the at-site (AS) parameters of a GEV pdf based on: (1) the full sample, (2) the sample with the largest observation removed, and (3) the sample with the largest observation replaced by its expected value. These three cases are denoted AS-AS, Omit-AS, and Rep-AS, respectively in Figure 10. The second set of three bars and boxplots in Figure 10 are the same cases, except using the known value of κ in all estimates which we term AS-KK, Omit-KK, and Rep-KK, respectively. Figure 10 illustrates clearly that both bias and RMSE associated with the design event increases dramatically due to sampling that results from estimating all three GEV model parameters. Dropping the largest observation leads to reductions in RMSE, but increases bias considerably—this is not an attractive alternative. Replacing the maximum observation with its expectation, further improves RMSE, but still increases bias somewhat relative to the estimation with the full sample. A refined version of this alternative, following an approach similar to that recommended by Cohn et al. (2013) to detect and remove low outliers, may be appealing when RMSE is the primary concern.

On the other hand, knowing the true value of κ greatly reduces both bias and RMSE associated with design event estimates. Our conclusion based on the figures above and below is that although some adaptive strategy involving replacement of the largest observation(s) may have promise, generating better estimates of κ using regional methods is likely to be the best approach forward for reducing both bias and RMSE

associated with design events. This further underscores recommendations in the existing body of research to 'substitute space for time' by using hydrologic records at different locations to compensate for short records at a single site (IAWCD 1982, Cunnane 1988, Stedinger et al. 1992, Griffis et al. 2004, Griffis and Stedinger 2009).

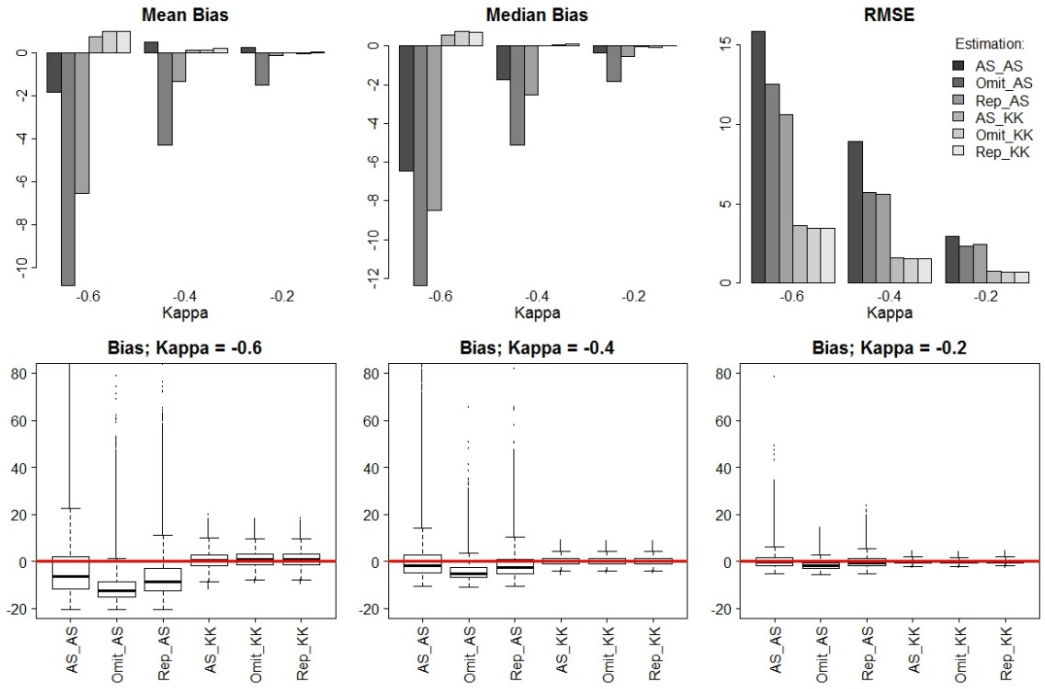

Figure 10. The effect of removing the largest observation (Omit) and replacing the largest observation with its expected value (Rep) on bias and RMSE of 100-year design event magnitude when estimation uses all at-site (AS) parameter estimate, and known κ (KK)

8. CONCLUSIONS AND RECOMMENDATIONS

We document systematic bias associated with the widely used PPCC goodness-of-fit metric. Monte Carlo experiments indicate that the value of the PPCC statistic, based on Q-Q plots: (a) declines when we introduce additional information about the true underlying statistical model and (b) is lowest for those samples that contain exactly the

extreme events that we are trying to model. Owing to these issues, we find that goodness-of-fit may not be the most reliable metric for choosing a probability distribution to model an individual sample. These findings highlight the fact that goodness-of-fit is assesses the ability of a model to mimic the observations, so that better fit leads to better mimicry of the behavior of the observations. This 'better mimicry' of the observations is not necessarily consistent with our goal, however, which is identification of the true underlying parent pdf of a flood series so that we can accurately predict quantiles, such as the 1000-year flood.

Further experiments lead us to observe systematic errors in quantile prediction when the maximum sample value diverges significantly from the n-year return period event, and to observe that witnessing such a divergence may provide a warning of possible design errors. We attempt two adaptive strategies, and find that although an approach involving replacement of the largest observation(s) may have promise, particularly in reducing RMSE, generating better estimates of κ using regional methods is likely to be the best approach forward for reducing both bias and RMSE associated with design events.

Overall, we seek to select the probability model that best represents the true relationship between the frequency and magnitude of our observations. The most effective way to do this is through hypothesis tests, which allow us to reject, or invalidate, a model with a chosen degree of confidence. Such hypothesis test have been developed for two parameter distributions (e.g., Stedinger et al. 1992), but few are available for the three parameter distributions used most widely in natural hazards analysis and in many other fields. Even those hypothesis tests which are available for three parameter pdf's such as for the LP3 hypothesis (Vogel and McMartin 1991) and

for the GEV hypothesis (Chowhurdy and Stedinger 1991), such tests lack power due to the need to estimate the third shape parameter of the pdf, in practice. Goodness-of-fit, evaluations are not the only alternative to hypothesis tests. For example, Laio et al. (2010) use a test that evaluates whether the maximum value in a sample is consistent with the hypothesis that a given distribution is the parent distribution to validate the statistical model in question. This approach can be particularly useful in contexts where the largest values are the most relevant observations, such as in flood frequency analysis. In a recent study, Renard et al. (2013) introduce a data-based comparison framework for pdf evaluation that emphasizes predictive ability and stability over goodness-of-fit. If multiple sites are being evaluated within the same geographic region, then Multiple Comparison Procedures (MCPs) can be used to test the hypothesis of a shared distribution across multiple samples. For a review of MCP's applied to the selection of a regional pdf of natural hazards, see section 6 of Thompson et al. (2011) as well as section 6 of Vogel et al. (2009). Lastly, if goodness-of-fit must be used, then a more reliable alternative may be to use P-P plots to calculate PPCC statistics rather than Q-Q plots (Gen and Koehler 1990).

In P-P plots, the percentiles associated with each of the ranked observations computed from the fitted pdf are plotted against the their unbiased plotting positions, The PPCC statistic based on a P-P plot is the correlation between these two axes, both of which have values between 0 and 1. Using PPCC statistics based on P-P plots treats each observation with effectively equal weight, and as a result, is influenced far less by outliers than a PPCC based on a Q-Q plot. Figure 11 compares our original boxplots of PPCC statistics using Q-Q plots shown previously in Figure 2 above with the equivalent set of PPCC statistics based on P-P plots. Recall here that each value of PPCC is

computed from one of the relatively long flood records at 200 rivers considered by Hirsch and Ryberg (2012). We note in Figure 11 that for all probability distributions, the 1,000-year events are much more evenly spaced across the range of PPCC values based on P-P plots, suggesting a more balanced assessment of the goodness-of-fit of a hypothesized pdf.

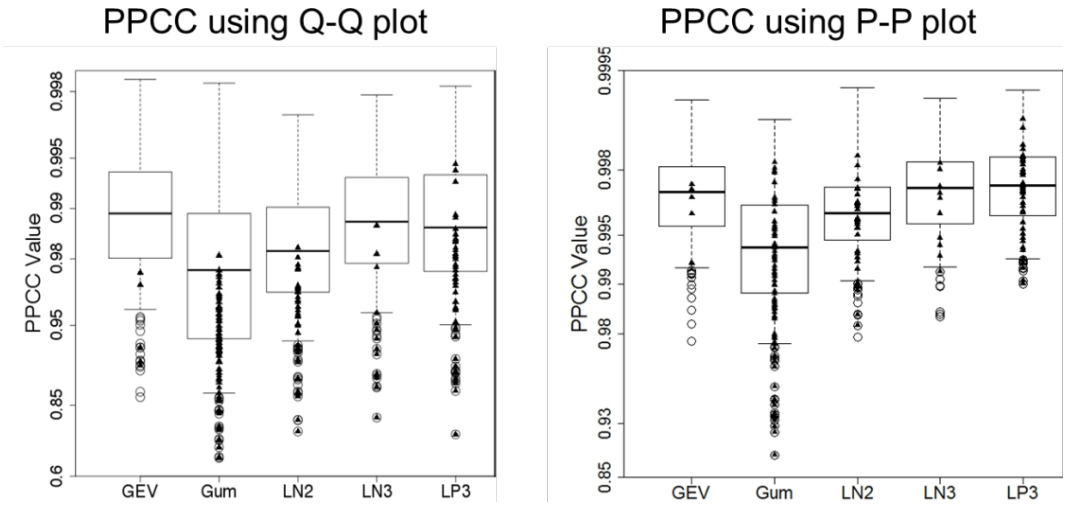

Figure 11. PPCC using Q-Q versus P-P probability plots for five distributions. Each boxplot includes 200 stations. Circles are outliers, and triangles are events with greater than 1,000-year return periods. GEV is Generalized Extreme Value; "Gum" is Gumbel; "LN2" is 2-parameter lognormal; "LN3" is 3-parameter lognormal; and "LP3"is Log-Pearson Type III.

REFERENCES

Asquith, W. H. 2011. Univariate Distributional Analysis with L-moment Statistics using R. Dissertation to the Department of Civil and Environmental Engineering, Texas Tech University.

Chowdhury, J.U., J.R. Stedinger, L.H. Lu. 1991. Goodness-of-fit tests for regional generalized extreme value flood distributions. Water Resources Research, 27 (7):1765–1776

Cohn, T. A., J. F. England, C. E. Berenbrock, R. R. Mason, J. R. Stedinger, and J. R. Lamontagne. 2013. A generalized Grubbs-Beck test statistic for detecting multiple

potentially influential low outliers in flood series, Water Resour. Res., 49, 5047–5058, doi:10.1002/wrcr.20392.

Cunnane, C., 1988. Methods and merits of regional flood frequency analysis. *J. Hydrol.*, 100: 269 290.

El Adlouni, S., B. Bobée, T.B.M.J Ouarda. 2008. On the tails of extreme event distributions in hydrology. *Journal of Hydrology.* 355: 16-33.

Gen, F. and K. Koehler. 1990. Goodness-of-Fit Tests Based on P-P Probability Plots. *Technometrics.* 32(3): 289-303.

Griffis, V. W., J. R. Stedinger, and T. A. Cohn. 2004. Log Pearson type 3 quantile estimators with regional skew information and low outlier adjustments, *Water Resour. Res.*, 40, W07503, doi:10.1029/2003WR002697.

Griffis, VW and JR Stedinger. 2009. Log-Pearson Type 3 Distribution and Its Application in Flood Frequency Analysis. III: Sample Skew and Weighted Skew Estimators, *Journal of Hydrologic Engineering*, 14(2).

Grubbs, F. 1969. Procedures for detecting outlying observations in samples, *Technometrics*, 11, 1–21.

Gubavera, T.S. 2011. Types of Probability Distributions in the Evaluation of Extreme Floods. *Water Resources*. 38(7): 962–971.

Helsel, D.R. and R. M. Hirsch, 2002. Statistical Methods in Water Resources Techniques of Water Resources Investigations, Book 4, chapter A3. U.S. Geological Survey. 522 pages.

Hirsch, R.M. and Ryberg, K.R., 2012, Has the magnitude of floods across the USA changed with global CO2 levels?, *Hydrological Sciences Journal,* 57(1).

Hosking, J. 1990. L-moments: Analysis and Estimation of Distributions using Linear Combinations of Order Statistics. *J. R. Statist. Soc. B*. 52(1): 105-124.

Hosking, J. and J. Wallis. 1997. Regional Frequency Analysis: An Approach Based on L-moments. Cambridge University Press.

Heo, J.H, Y.W. Kho, H. Shin, S. Kim, and T. Kim. 2008. Regression equations of probability plot correlation coefficient test statistics from several probability distributions. *Journal of Hydrology*. 335(1-4):1-15.

IACWD (Interagency Committee on Water Data). 1982. Guidelines for Determining Flood Flow Frequency. *Bulletin 17B* of the Hydrology Subcommittee, Washington, D.C.

Kidson, R. and K.S. Richards. 2005. Flood frequency analysis: assumptions and alternatives. *Progress in Physical Geography*. 29(3): 392-410.

Kim, S. and Heo, J. 2010. Comparison of the Probability Plot Correlation Coefficient Test Statistics for the General Extreme Value Distribution. World Environmental and Water Resources Congress 2010: pp. 2456-2466. doi: 10.1061/41114(371)253

Laio, F., P. Allamano, and P. Claps. 2010. Exploiting the information content of hydrological "outliers" for goodness-of-fit testing. *Hydrol. Earth Syst. Sci.* 14: 1909-1917. doi:10.5194/hess-14-1909-2010

Renard, B., K. Kochanek, M. Lang, F. Garavaglia, E. Paquet, L. Neppel, K. Najib, J. Carreau, P. Arnaud, Y. Aubert, F. Borchi, J.-M. Soubeyroux, S. Jourdain, J.-M. Veysseire, E. Sauquet, T. Cipriani and A. Auffray, 2012. Data-based comparison of frequency analysis methods: a general framework. *Water Resources Rese*arch. 49: 825–843, doi:10.1002/wrcr.20087.

Stedinger, J., Vogel, R., and Foufoula-Georgiou, E. 1992. Handbook of Hydrology, chap. 8: Frequency analysis of extreme events, McGraw-Hill, New York.

Thompson, E.M., Hewlett, J.B., Baise, L.G., and Vogel, R.M. 2011. The Gumbel hypothesis test for left censored observations using regional earthquake records as an example, *Nat. Hazards Earth Syst. Sci.*, 11, 115-126, doi:10.5194/nhess-11-115-2011.

Travis, C.C., and M. L. Land. 1990. Estimating the mean of data sets with nondetectable values. *Environ. Sci. Technol*. 24, 961-962.

Vogel, R. 1986. The probability plot correlation coefficient test for the normal, lognormal, and Gumbel distributional hypotheses. *Water Resour. Res.* 22: 587–590.

Vogel, R. and McMartin, D. 1991. Probability plot goodness-of-fit and skewness estimation procedures for the Pearson type 3 distribution. *Water Resour. Res.* 27: 3149–3158.

Vogel, R.M. and I. Wilson. 1996. The Probability Distribution of Annual Maximum, Minimum and Average Streamflow in the United States, Journal of Hydrologic Engineering, ASCE, Vol. 1, No. 2, pp. 69-76.

Vogel, R.M., J.R.M. Hosking, C.S. Elphick, D.L. Roberts, and J.M. Reed, 2009. Goodness-of-fit of Probability Distributions for Sightings as Species Approach Extinction, *Bulletin of Mathematical Biology*, DPO 10.1007/s11538-008-9377-3.

TRANSFORMATIONAL CONCORDANCE OF THE GENERALIZED EXTREME VALUE HYPOTHESIS

Brent Boehlert[1,2], Richard M. Vogel[1], and Nicholas C. Matalas[3]

Author affiliations: 1. Tufts University, Department of Civil and Environmental Engineering, Medford, MA
2. Industrial Economics, Inc., Cambridge, MA
3. 709 Glyndon St. S.E., Vienna, VA 22180

ABSTRACT

One way to evaluate whether a scientific hypothesis is consistent with observations is to determine whether the data are concordant with that hypothesis from several perspectives. For a hypothesis to be consistent, it must exhibit concordance with observed data across space, time, as well as across functional transformations. We focus on the problem of selecting a probability distribution for modeling natural hazards that exhibits *transformational concordance* with observed data. A data series is transformationally concordant if its properties under each transformation (e.g., real space to log space) are consistent with our theoretical understanding of how those properties change across transformations. For example, in real space, if a sequence of observations is assumed lognormal and if logarithms exhibit zero skewness, the lognormal distribution in real space and the normal distribution in log space would be transformationally concordant. Using 200 long flood records in the U.S., we explore the transformational concordance of the Generalized Extreme Value (GEV) distribution. If observations are GEV, then their adjusted logarithms should follow a Gumbel pdf. The GEV pdf exhibited the best goodness-of-fit among several alternative pdfs for the 200 flood series using traditional metrics. However, the flood series were found to be transformationally discordant under the GEV hypothesis, and the GEV model was found to significantly underestimate the frequency of extremely large design events. Both the

discordance and design event bias appear to be attributable to systematic bias in GEV parameter estimation that is largely addressed by application of regional skew estimates. We identify the potential for a GEV hypothesis test based on transformational concordance.

1. INTRODUCTION

Determining whether a scientific hypothesis is consistent with observations is a fundamental step in developing any mathematical model. One way to evaluate consistency of a particular hypothesis is to determine whether the data are concordant with that hypothesis from several perspectives. We call this notion *concordance*. For a hypothesis to be consistent, it must exhibit concordance with observed data across space, time, as well as across functional transformations, which we term *transformational concordance*. In this initial study, we focus on the problem of selecting a probability distribution for modeling natural hazards that is transformationally concordant with observed data. Appendix A introduces other forms of concordance, which may be useful in future investigations.

In all fields of natural hazards, the problem of estimating the magnitude of a design event and its associated frequency of occurrence is of basic concern. Estimation of these metrics requires hypothesizing a particular probability density function (pdf) and then evaluating the consistency of that pdf with observed data. Because of their flexibility, three-parameter pdfs are typically employed to model natural hazards, and several approaches have been advanced for evaluating alternative hypotheses for these pdfs. Hypothesis tests have been developed for some three-parameter pdfs including the log Pearson type III (LP3) (Vogel and McMartin 1991) and GEV (Chowhury and Stedinger 1991) hypotheses, although the need to estimate the third shape parameter of the pdf from data series limits the power of such tests. For example, Vogel and McMartin (1991) document that even though the third parameter of the LP3 distribution (log skew) can improve the apparent 'goodness-of-fit' of resulting probability plots over alternative pdfs, that 'goodness-of-fit' is misleading, because the

resulting test often failed to detect departures from the LP3 model for the sample sizes typically experienced in hydrologic applications. Similar experiences led Boehlert (2015) to conclude that 'goodness-of-fit' can mislead us to select probability models and parameter estimates that perform poorly in producing the results we care about most in frequency analysis, such as the magnitude and average return period of design flood events. The maximum value test (see Laio et al. 2010), evaluates whether the maximum value in a sample is consistent with the hypothesis that a given distribution is the parent distribution. In regional applications, many hypothesis tests are applied repeatedly resulting in a regional hypothesis test; this idea was first introduced to the climate and hydrology literature by Livezy and Chen (1983) using the concept of field significance. The concept of field significance is more generally termed Multiple Comparison Procedures (MCP) by statisticians. Thompson et al. (2011) and Vogel et al. (2009) provide an introduction, review and application of the use of MCPs and field significance for use in natural hazards applications, and Douglas et al. (2000) show the importance of accounting for spatial correlation of flow series when performing repeated distributional hypothesis tests. Yet another approach for evaluating three-parameter pdfs was introduced by Renard et al. (2013), who advance a data-based comparison framework for pdf evaluation that emphasizes predictive ability and stability over goodness-of-fit.

We introduce and evaluate the concept of transformational concordance, which holds promise for gaining insight into the true statistical properties underlying a given dataset. A data series is transformationally concordant if its properties under each transformation (e.g., real space to log space) are consistent with our theoretical understanding of how those properties change across transformations. For example, in real space, if a sequence of observations is presumed to be distributed as lognormal and

if the logarithms of the observations yield zero skewness, the skewness would not contradict the normal distribution and therefore, the lognormal distribution in real space and the normal distribution in log space would be transformationally concordant. On the other hand, if the logarithms of the observations yield significantly non-zero skewness, then the two distributions would be transformationally discordant, indicating that the original data may not be lognormally distributed.

The broader concept of concordance has been applied in a range of fields to evaluate particular scientific hypotheses. For instance, in disciplines that require aggregation of ranks from different sources, concordance describes the agreement between n ranks of k objects (e.g., different judges evaluating a set of competitors); one measure of such agreement is the *concordance correlation coefficient* (Lin 1989). In biology, ecology, and other physical sciences, *spatial concordance* measures the degree to which the spatial pattern of a particular driver predicts the spatial pattern of some outcome (see Garcia et al. 2005). To measure these relationships, ecologists have applied the *Kendall coefficient of concordance* to evaluate species associations in community ecology (see Legendre 2005). In the more relevant context of natural hazard frequency analysis, researchers have used the concept of concordance to describe the degree of correlation between the flow sequences of two or more converging rivers and other correlations between hydrological variables (see Favre et al. 2004). Hosking and Wallis (1997) describe a "discordancy measure" that identifies sites that are grossly discordant with the group as a whole, where discordancy is measured based on L-moments of the sites' data.

In this initial study, we focus on the transformational concordance of the Generalized Extreme Value (GEV) distributional hypothesis. GEV is a three-parameter

distribution used widely for analysis of natural hazards, and for flood frequency analysis (see Renard et al. 2013, Villarini et al. 2011/2012, Gubavera 2011, Salinas et al. 2013). Note that the U.S. is one of the only nations that does not use the GEV model and instead mandates the use of the LP3 distribution for use in flood frequency analysis (see *Bulletin 17B* [IACWD 1982] and Table 3 in Vogel and Wilson 1996). As described further below, if a data series is distributed as GEV, then its logarithms (adjusted by its upper or lower bound) follow a Gumbel pdf (i.e., Extreme Value Type I). As a result, if a dataset is transformationally discordant between GEV in real space and Gumbel in log space, that dataset is either not GEV distributed or the estimated GEV parameters used to calculate the lower/upper bounds are incorrect.

We explore the concept of transformational concordance in flood frequency analysis using a dataset of annual maximum streamflow from 200 stream gages operated by the U.S. Geological Survey (USGS) within the contiguous U.S. Each of these series is at least 85 years in length through water year 2008, and is within "unmanaged" drainage basins that have few or no reservoirs, water withdrawal intakes, or other water infrastructure development (see Figure 12). Hirsh and Ryberg (2012) provide additional details on this dataset as well as further explorations concerning their stationary behavior. Throughout our analyses, we employ the method of L-moments (Hosking 1990, Hosking and Wallis 1997) for fitting pdfs to these and other flood series.

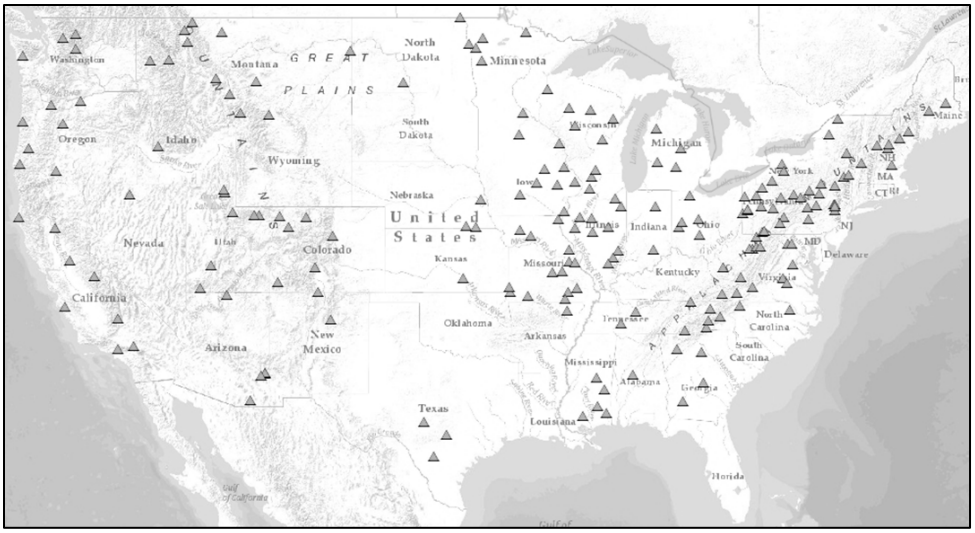

Figure 12. Locations of 200 USGS station with annual maximum flow data, introduced by Hirsh and Ryberg (2012)

In the remainder of the paper, we provide a brief introduction to the theory of transformational concordance, evaluate the transformational concordance of the Hirsch and Ryberg (2012) annual maximum flow dataset, and then examine the source of apparent discordance we find in that evaluation. We further explore the transformational concordance concept to explain apparent systematic bias in the frequency of extreme floods when fitting the GEV distributions.

2. TRANSFORMATIONAL CONCORDANCE: THE CASE OF THE GEV DISTRIBUTION

There is general consensus among hydrologists that sequences of annual floods are well described by one or another of the three asymptotic distributions of the largest value. These distributions may be compactly represented in the form of a generalized extreme value (GEV) distribution formulated by Mises (1936) and subsequently utilized by Jenkinson (1955) to investigate meteorologic extremes. Since then, the GEV distribution has found application to numerous other natural hazards, including

earthquakes (Pisarenko et al. 2014, Thompson et al. 2012), floods (Renard et al. 2013, Villarini et al. 2011/2012), droughts (Sousa et al. 2011), and drought and flood stages of lakes (Paynter and Nachabe 2010). Its cumulative distribution function may be expressed as:

$$F(Y < y) = \exp\left[-(1-\kappa y)^{1/\kappa}\right] \tag{1}$$

where

$$y = (x-\varepsilon)/\alpha \tag{2}$$

where ε and $\alpha > 0$ are parameters of location and scale and κ is the shape parameter.

As $\kappa \to 0$, Eq. (1) tends to the Type 1 extreme value distribution or Gumbel pdf,

$$F(Y < y) = \exp\left[-\exp(-y)\right] \tag{3}$$

The distribution is unbounded below and above, $-\infty \le y \le \infty$, and has a fixed value of skew, $\gamma = 1.139$. If $\kappa < 0$, Eq. (1) corresponds to the Type II extreme value distribution,

$$F(Y < y) = \exp\left[-\left(y^{-\kappa}\right)\right] \tag{4}$$

The distribution is bounded below but not above, $1/\kappa \le y \le \infty$. If $\kappa > 0$, Eq. (1) corresponds to the Type III extreme value distribution also referred to as the Weibull pdf,

$$F(Y < y) = \exp\left[-(-y)^{\kappa}\right] \tag{5}$$

The distribution is bounded above but not below, $-\infty < y \le 1/\kappa$.

If X is distributed as Type II, then

$$Z = \ln(X - m) \tag{6}$$

is distributed as Type I, and likewise, if X is distributed as Type III, then

$$Z = -\ln(m - X) \tag{7}$$

is distributed as Type I. See e.g., Johnson and Kotz (1995). The lower bound m is defined as:

$$m = \varepsilon + \frac{\alpha}{\kappa} \tag{8}$$

And ε, α, and κ are the three parameters of the GEV distribution. In real space, let $\{x_i : i = 1,...,n\}$ represent a sequence of annual peak streamflows spanning a period of *n* years. The sequence may be fitted to a generalized extreme value (GEV) distribution using various approaches including the method of moments and L-moments (Stedinger et al. 1992) or maximum likelihood estimators introduced by Martins and Stedinger (2000). If in log space, there is no contradiction in the shape of the distribution and the shape implied by the sequence $\{z_i = \ln(x_i - m) : i = 1,...,n\}$, then the generalized distributions in real space and in log space and the associated GEV hypothesis are transformationally concordant.

In real space, if the skewness estimated from the sequence $\{x_i : i = 1,...,n\}$ is positive, then the sequence may be fitted with the GEV distribution, whether $\kappa < 0$ (Type II EV distribution) or $\kappa > 0$ (Type III EV distribution). And if in log space, the skewness estimated from $\{z_i = \ln(x_i - m) : i = 1,...,n\}$ is positive, then the GEV distributions in real and in log space are concordant if skewness in log space is in the neighborhood of 1.139, the skewness of the Type I EV distribution, equivalently the GEV distribution with $\kappa = 0$. If, however, the skewness in log space is outside the

neighborhood of 1.139, then the GEV distributions in real and in log space are discordant.

In real space, if skewness is negative, the observations may be fitted with the Type III EV distribution, equivalently the GEV distribution for which $\kappa > 0$. Note that the Type III EV distribution may be positively or negatively skewed. In either case, the distribution is bound above and unbounded below. Unless the skewness in log space is in the neighborhood of 1.139, the GEV distributions in real space and in log space are discordant.

Whether the skewness in real space is positive or negative, the sequence in log space cannot be fitted with a GEV distribution if the skewness in log space is negative. In this case, the GEV distributions in real space and in log space are discordant.

3. GOODNESS-OF FIT OF GEV DISTRIBUTION FOR U.S. ANNUAL MAXIMUM FLOOD DATA

We begin our evaluation of the GEV hypothesis by using traditional goodness-of-fit metrics to assess whether the GEV model mimics U.S. flood observations. Figure 13 illustrates L-moment diagrams which enable us to compare the behavior of several theoretical pdfs (shown using curves) with sample L-moments computed from the 200 flood series. See Hosking (1990), Vogel and Fennessey (1993), Stedinger et al. (1992) and Hosking and Wallis (1997) for a review of the application of L-moment diagrams for assessing the goodness-of-fit of alternative pdfs to flood samples. Based on a comparison of the estimated L-skewness and L-kurtosis of the 200 stations the theoretical relationship for a GEV, three parameter lognormal (LN3) and a three

parameter Pearson (PE3) distribution in Figure 13, the GEV model appears to provide a good fit to the data. As expected from other assessments of suitable pdf's for modeling observed flood series (Vogel and Wilson 1996, Kidson and Richards 2005, El Adlouni et al. 2008, Gubavera 2011), we conclude from Figure 13 that the GEV, LN3 and PE3 pdfs are all consistent with flood experience on these 200 U.S. rivers.

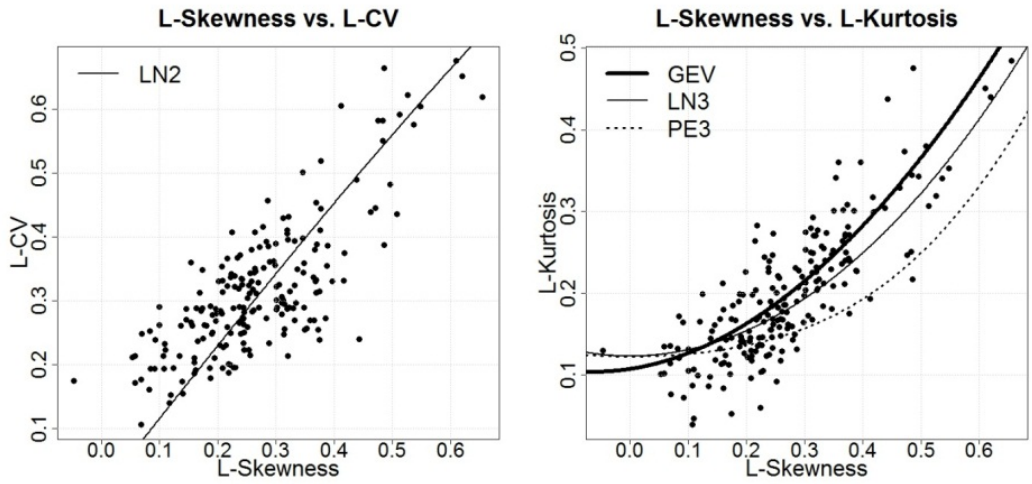

Figure 13: L-moments of annual maximum streamflow dataset

In addition to L-moment diagrams as measure of goodness-of-fit, we considered the linearity of probability plots based on various hypothesized pdfs. We evaluated the Probability Plot Correlation Coefficient (PPCC) using quantile-quantile (Q-Q) probability plots for the 200 stations. The PPCC statistic is the correlation between the ordered observations versus estimates of the expectation of those ordered observations based on a hypothesized fitted pdf. Such probability plots and associated PPCC goodness-of-fit comparisons are now widely used in the field of hydrologic frequency analysis (Stedinger et al. 1992, Vogel 1986, Vogel and McMartin 1991 and Heo et al. 2008). A determination of the expected value of the ordered observations based on a fitted distribution used to compute the PPCC is based on a plotting position chose to

reproduce the expected values of the order statistics for the hypothesized pdf, along with the quantile function and set of model parameters. Figure 3 illustrates PPCC goodness-of-fit statistics for the following distributional alternatives: GEV, Gumbel (Gum), 2-parameter lognormal (LN2), LN3, and Log Pearson Type III (LP3), all based on at-site parameter estimates using the method of L-moments. The mean PPCC value for the GEV pdf was higher than for the other four distributions, at approximately 0.99 as shown in Figure 14, thus we conclude that among the five pdf's considered in Figure 14, the GEV model exhibits the best overall goodness-of-fit.

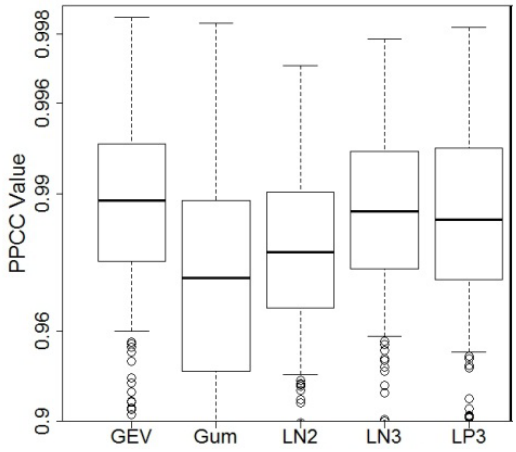

Figure 14: Boxplots of PPCC values of the 200 annual maximum flow series fit to various distributions

Although both L-moment diagrams and PPCC goodness-of-fit evaluations suggest that the GEV pdf fits the datasets well, we found that the GEV does a poor job of estimating the frequency of extremely large events. In the 200 flood series, each ranging in length from 85 to 126 years (average length 94.2 years), there is a total of approximately 19,000 site-years of observations. If all these observations are assumed to be independent in both space and time, among the 19,000 site years of data, we would expect to observe about 19 (0.1%) flood observations which exceed our estimate of the 1,000 year flood using a GEV model. The number 19 is also a random number

with a distribution that follows a Bernoulli pdf, enabling us to compute the likely range associated with this random variable. Equation (2) of Vogel et al. (1993a) documents how to construct the likely range associated with this random variable assuming spatial and temporal independence of the flood samples. Here the 95% likely range of events which would be expected to exceed the 1,000-year flood is from 12 to 26 events. If one accounted for the weak temporal and spatial correlation of these flood series using methods described by Matalas and Langbein (1962), the likely range would likely increase slightly, because spatial and temporal correlation are analogous to decreasing the effective sample size associated with this experiment.

Beard (1977) and Vogel et al. (1993a/b) performed similar analyses using three different datasets, where they counted the number of times actual streamflow observations exceeded estimates of the 1,000-year flood , although Beard did not consider the GEV distribution. Vogel et al. (1993a) focused on the southwestern U.S. and found that the number of observations which exceeded GEV estimates of the 1,000-year events using no expected probability adjustment fall within, but at the bottom of, their 95% likely intervals of 11 to 28 events. In an analysis of Australian flood frequency, Vogel et al. (2013b) find that 17 GEV estimates of the 100-year event occur when their 90% likely interval is 13 to 26. In our analysis, on the other hand, only six observed floods exceeded the GEV estimates of the 1,000-year floods based on a GEV model, suggesting that the magnitude of extreme design events are systematically overestimated when fitting a GEV model using at-site L-moment estimators of the model parameters. Note that these estimates do not incorporate an expected probability adjustment (Stedinger, 1983) which has currently only been derived for the normal distribution. According to the findings of Stedinger (1983), without an expected

probability adjustment, one would expect the number of exceedances of a sample estimate of the 1,000-year flood to be greater than our expectation. As such, incorporating an expected probability adjustment in our context would actually decrease the already-underestimated frequency of 1,000-year events. In the following sections, we further explore the inconsistency between the fact that among various plausible pdfs considered, the GEV model appears to exhibit the highest goodness-of-fit, yet cannot reproduce the frequency of large flood events.

4. TRANSFORMATIONAL CONCORDANCE OF U.S. ANNUAL MAXIMUM FLOOD DATA

Our previous goodness-of-fit evaluations based on L-moment diagram and probability plots led us to conclude that the GEV model is consistent with U.S. flood observations. Now we evaluate whether the flood series exhibit transformational concordance under the GEV hypothesis. If concordant, then the L-moment diagram of transformed data given in equations (6) and (7) should have L-skew and L-kurtosis values that fall within the neighborhood of the theoretical Gumbel values of 0.17 and 0.15, respectively. Here the neighborhood is an elliptical region centered at this point, the size of which is specified based on desired level of confidence and sample size (see Chapter 3 of Hosking and Wallis 1997 and Liou et al. 2008). Figure 15 illustrates these elliptical regions for 10,000 synthetically-generated Gumbel series each of length n = 15, 25, 50 and 100. The elliptical region encloses 95% of the points, and thus constitutes the 95% confidence region for the L-moment ratio estimates derived from Gumbel data.

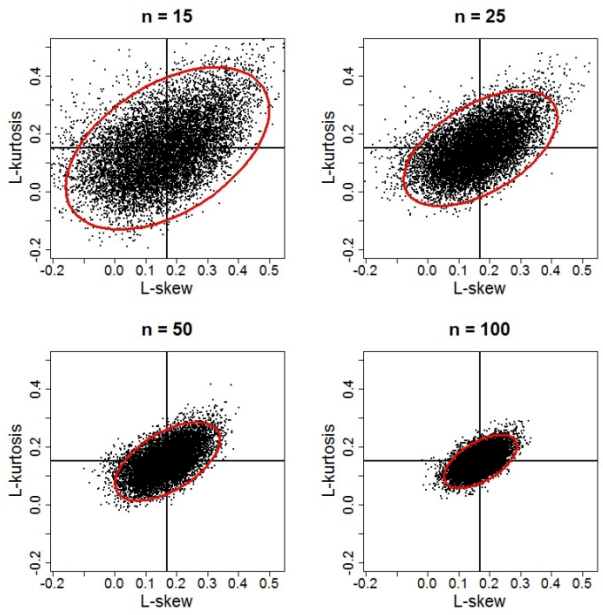

Figure 15: 95% confidence ellipses which define the expected neighborhood for estimates of L-skew and L-kurtosis for Gumbel samples of length n= 15, 25, 50, and 100

To investigate the transformational concordance of the USGS annual maximum flow observations, we first used the method of L-moments to estimate the GEV parameters for (a) the full datasets, and (b) a truncated set of the data containing only the first 10 and 25 observations of each series. Recall that if X is GEV then the transformation Z given in equation (6) and (7) should follow a Gumbel pdf. We transform the observed time series X to Z using at-site L-moment estimates of the GEV parameters to estimate the lower/upper bound m (equation 8). If concordant, the transformed series should fall within the neighborhood of the Gumbel distribution illustrated earlier in Figure 4. Figure 16 illustrates L-moment diagrams for the values of Z, which do not look anything like the expected behavior we saw in Figure 15. We conclude from Figure 16 that either (a) the observed flood series are not concordant with the GEV hypothesis, or (b) sampling variability associated with parameter estimates is occluding our ability evaluate the GEV hypothesis.

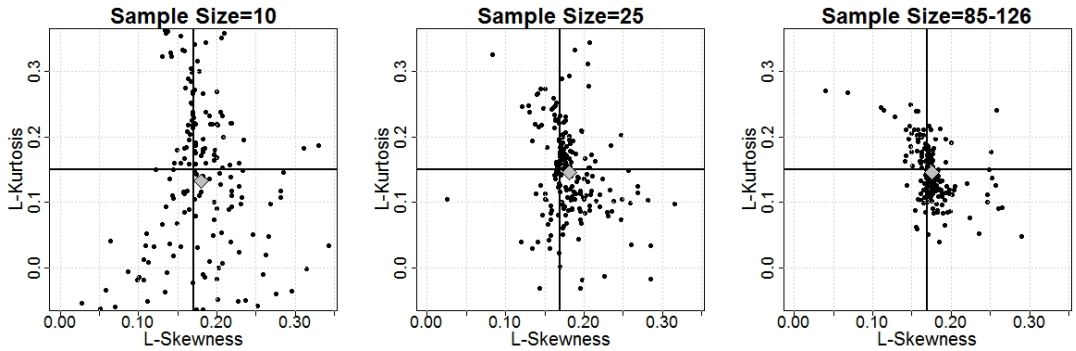

Figure 16: L-skewness versus L-kurtosis of values of Z corresponding to annual maximum flood data; sample sizes of 10, 25, and 85-126

5. SOURCE OF DISCORDANCE IN GEV DATA: PARAMETER ESTIMATION

In this section, we explore the possible reasons for the apparent GEV discordance illustrated in the previous section. We hypothesized that the discordance was caused by at-site estimates of κ, as the literature has long-recommended using regional estimates of shape parameters to effectively lengthen sample records using data from other sites (see *Bulletin 17B* [IACWD 1982], Stedinger and Lu 1995, Reis et al. 2005, and Griffis and Stedinger 2009). Figure 17 compares L-moment diagrams for the 200 stations using at-site parameter estimates based on the method of L-moments (A; at left), at site ε and α estimates but regional estimates of κ (B; at center), and synthetic data generated from the 200 sets of at-site GEV model parameters that by definition have no parameter uncertainty and therefore known lower bounds (C; at right). For the purposes of this study, each regional κ estimate is simply the average of at-site κ values that fall within a common 4-digit USGS hydrologic unit code (HUC; there are 204 4-digit HUCs in the contiguous U.S.). When we used the regional instead of at-site κ value in calculation of the lower bound, the resulting data pattern (B) is much more elliptical and thus more consistent with the theoretical results observed in Figure 15. Using synthetic

data (C) when the values of κ are assumed known apriori, produce the ellipses as expected. Figure 18 illustrates the variability in sample estimates of the shape parameter κ. Not surprisingly, at-site L-moment estimates of the κ values for the 10- and 25-year truncated records vary widely, whereas the at-site and regional values that rely on the complete series are much more tightly bound between approximately -0.5 and 0.2.

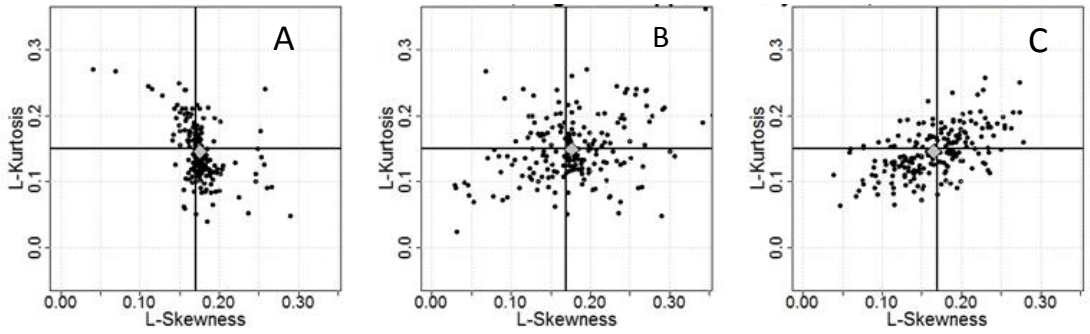

Figure 17. Comparison of L-moments of the values of the transformed GEV variate Z based on at-site (A), regional (B), and true values of κ (C). The At-site and regional results are based on U.S. flood series and the synthetic results are based on synthetic GEV samples with known parameters.

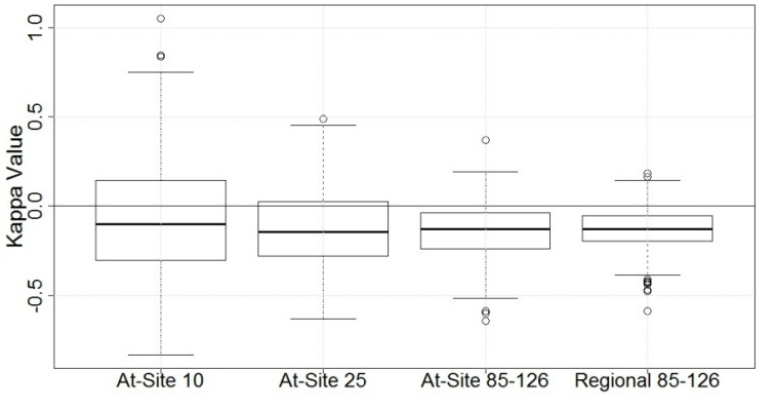

Figure 18: Boxplots of at-site and regional κ values for truncated and full sample series

Figure 17 indicates that the source of the apparent discordancy between the theoretical GEV model and U.S. flood series lies in our estimate of κ, which dictates how to transform the observations X into the transformed GEV series Z. That figure also

demonstrated that when the uncertainty in the κ value is resolved (either using regional estimates or if one actually knew the true value), the apparent GEV discordancy disappears. Yet sample lengths 85-126 should be sufficient to avoid significant bias in at-site shape parameter estimates. In recent formulations of the weighted estimate of the shape parameter (blend of at-site and regional), the at-site value dominates the weighting when sample lengths exceed 60-75 (Stedinger and Lu 1995 focus on GEV; Griffis and Stedinger 2009 focus on LP3). This suggests that at-site parameter estimates from these large samples should avoid the discordance observed in Figure 17.

To test whether the GEV discordancy resolves itself with large sample sizes, we generated 1,000 samples of length 10,000, each of which has mean of 1 and L-cv of -0.4 and κ values varying from -0.6 to 0.4. In Figure 19, the title of each subplot shows the percentage of the 1,000 runs with estimated lower bounds m that were larger than the smallest observation, or upper bounds m, that were smaller than the largest observations—these cases were omitted. In each plot the gray points are transformed using known values of parameters, whereas black points are transformed using at-site parameter estimates based on method of L-moments. Figure 19 suggests that even when relying on exceptionally large sample sizes for at-site parameter estimation, we still observe transformational discordance. For $\kappa < 0.2$, we observe strong correlation between L-skewness and L-kurtosis, and too little variance in L-skewness, particularly as $\kappa \rightarrow 0$. Although the source of this discordance is a matter for further research, given that these large sample sizes effectively eliminate sampling uncertainty, the result in Figure 19 may be caused by some fundamental aspect of the parameter estimation procedure. In the next section we explore how the variability in the GEV shape parameter can also produce systematic bias in design flood estimates.

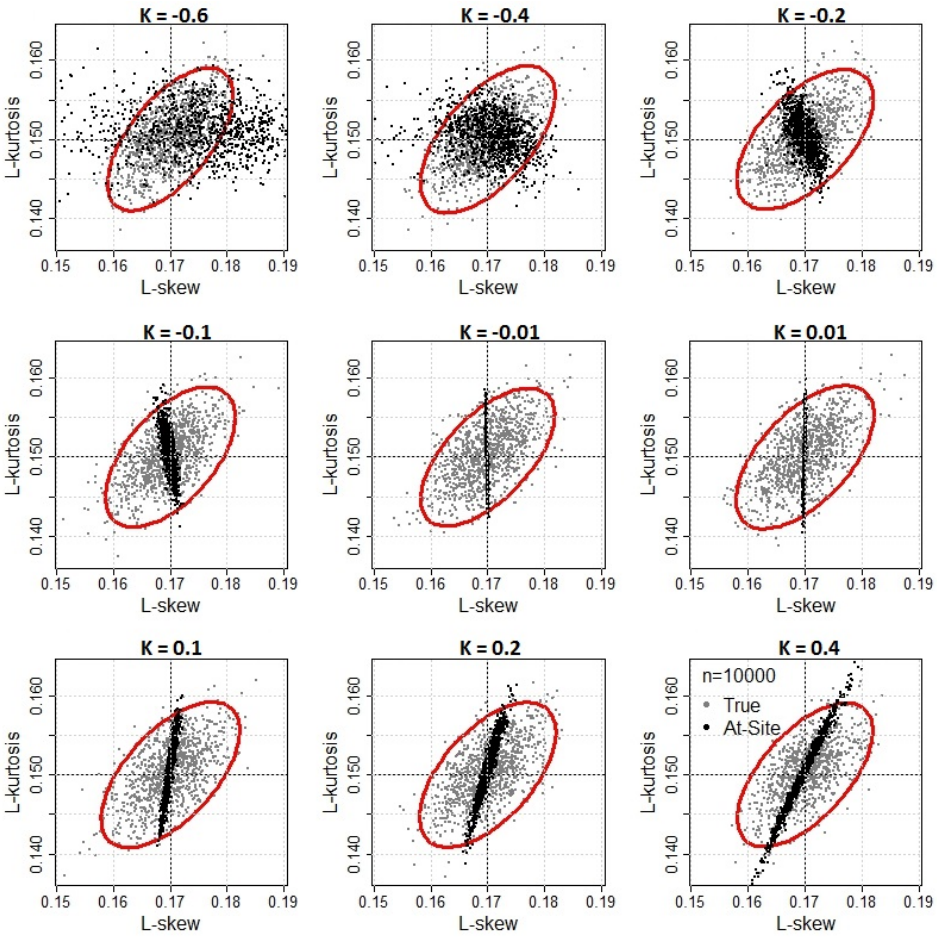

Figure 19. L-moment diagrams for transformed GEV data using known and at-site parameters to estimate lower bound; 1,000 runs of n =10,000. Gray points are transformed to log space using known parameters, whereas black points are transformed using at-site parameters.

6. GEV SHAPE PARAMETER ESTIMATION ERRORS LEAD TO SYSTEMATIC BIAS IN DESIGN FLOODS

In the previous sections we document how the sampling variability associated with at-site estimates of the GEV shape parameter κ can prevent us from understanding the true underlying probabilistic structure of the observations of interest. These investigations led us to conclude that U.S. flood observations appear discordant with the GEV model in spite of the fact that standard goodness-of-fit metrics led us to conclude

that the GEV model is preferred over other 3-parameter alternatives. We now examine the bias associated with design flood estimates from the sampling uncertainty in at-site estimates of κ. Recall that across 19,000 site-years of flood data from the USGS dataset, only six observations were found to exceed the 1,000-year event predicted using GEV at-site L-moment estimates of model parameters. If the flood series were temporally and spatially independent, we would have expected 19 such exceedances (95% likely interval: 12 to 26). Thus a GEV model fitted to relatively long hydrologic records (the average record length is 94.2 years) using at-site parameter estimates will generally underestimate the frequency of very extreme design events.

To further evaluate the potential implications of this finding, we generated 10,000 GEV samples of length 100 each for 100 sets of κ values which ranged from -0.6 to 0.4 at an interval of 0.01 (excluding 0). Thus for each value of κ considered, we generated (10,000 x 100 = 1,000,000 site years of floods). As in our previous analyses, the L-cv and mean for each data series is set at 0.4 and 1, respectively. For each sample, we estimated the at-site parameters and the magnitude of the 1,000-year return period event based on L-moment estimators of (a) at-site parameters, and (b) at-site α and ε, but known κ. We then counted the number of observations which exceeded an event with an estimated 1,000-year return period, which we would expect would be approximately 1,000 events (i.e., 0.1% of the 1 million site-years of observations). Figure 20 displays these results. When using the known value of κ for design event estimation, the number of observations which exceeded the estimated 1,000-year events matches our expectation nicely (black points). On the other hand, using the at-site L-moment estimator of κ to estimate the 1,000 year design event from synthetic GEV samples tended to produce either too many extreme flood events (for $\kappa > 0.15$) or

too few (for $\kappa < 0.15$). Given that the interquartile range of the L-moment estimates of at-site κ values for the 200 stations is approximately -0.05 to -0.25, the mean of this range corresponds to approximately 300 flood events, or 30% of the flood observations expected to exceed the 1,000-year flood. This is very similar to the 6 versus 19 actual flood events (31%) which exceeded estimates of the 1,000 year flood in our previous analysis using the Hirsch and Ryberg (2012) dataset. We conclude from this experiment that at-site estimation of the GEV shape parameter, even using reasonably long records, may cause systematic bias in flood frequency estimates.

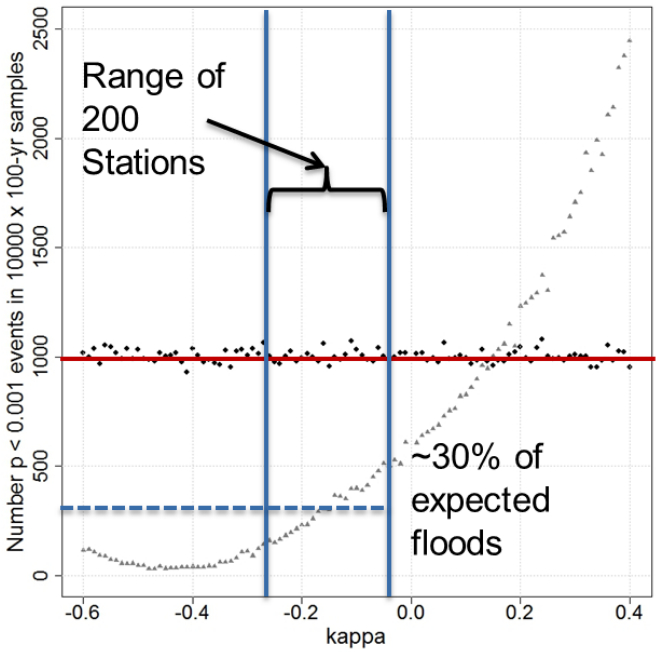

Figure 20. Number of GEV observations which exceeded the estimated 1,000-year return period event as a function of κ for the GEV distribution; 10,000 samples of n=100 yield a total of 1,000,000 years of observation from which we would expect 1,000 events to exceed our estimate of the 1,000-year return periods design event.. Black points are calculated using a known κ value, and gray points are based on at-site L-moment estimators of κ.

7. TRANSFORMATIONAL CONCORDANCE: A POSSIBLE AVENUE FOR HYPOTHESIS TESTING?

Finally, we explore the potential for using the concept of transformational concordance to develop a hypothesis test. Figure 15 illustrated 95% confidence ellipses for L-moment diagrams corresponding to the Gumbel hypothesis for various sample sizes. If X is GEV then the transformation Z should follow a Gumbel pdf. If all or a significant portion of the estimated L-moments of the values of Z fall outside of that ellipse, we would have strong evidence that the real-space data X is not GEV thus providing a possible test for the GEV hypothesis. To illustrate this concept, we synthetically generate samples of GEV and LP3 data of various sample sizes, and then we obtain L-moment sample estimates of the GEV model parameters which we use to estimate the lower and upper bounds m (equation 8) followed by computation of the transformed value of Z given in equations (6) and (7). Our expectation is that for modest sample sizes of LP3 synthetic data the L-moment diagrams for the transformed value Z will lead to estimated L-moments which fall squarely outside of the 95% confidence ellipse derived from Gumbel synthetic data. We find that the estimated L-moments of the estimated values of Z computed from LP3 data do indeed fall outside of the Gumbel ellipse, but only when using exceptionally long sample lengths as is shown in Figure 21. This suggests that there is potential for developing a GEV hypothesis test, but further development is needed to make such a test applicable to the much smaller sample sizes typically available. Such a hypothesis test could draw on goodness-of-fit tests for acceptance regions in L-moment diagrams developed by Liou et al. (2008), as well as measures of spatial homogeneity commonly used in GIS applications.

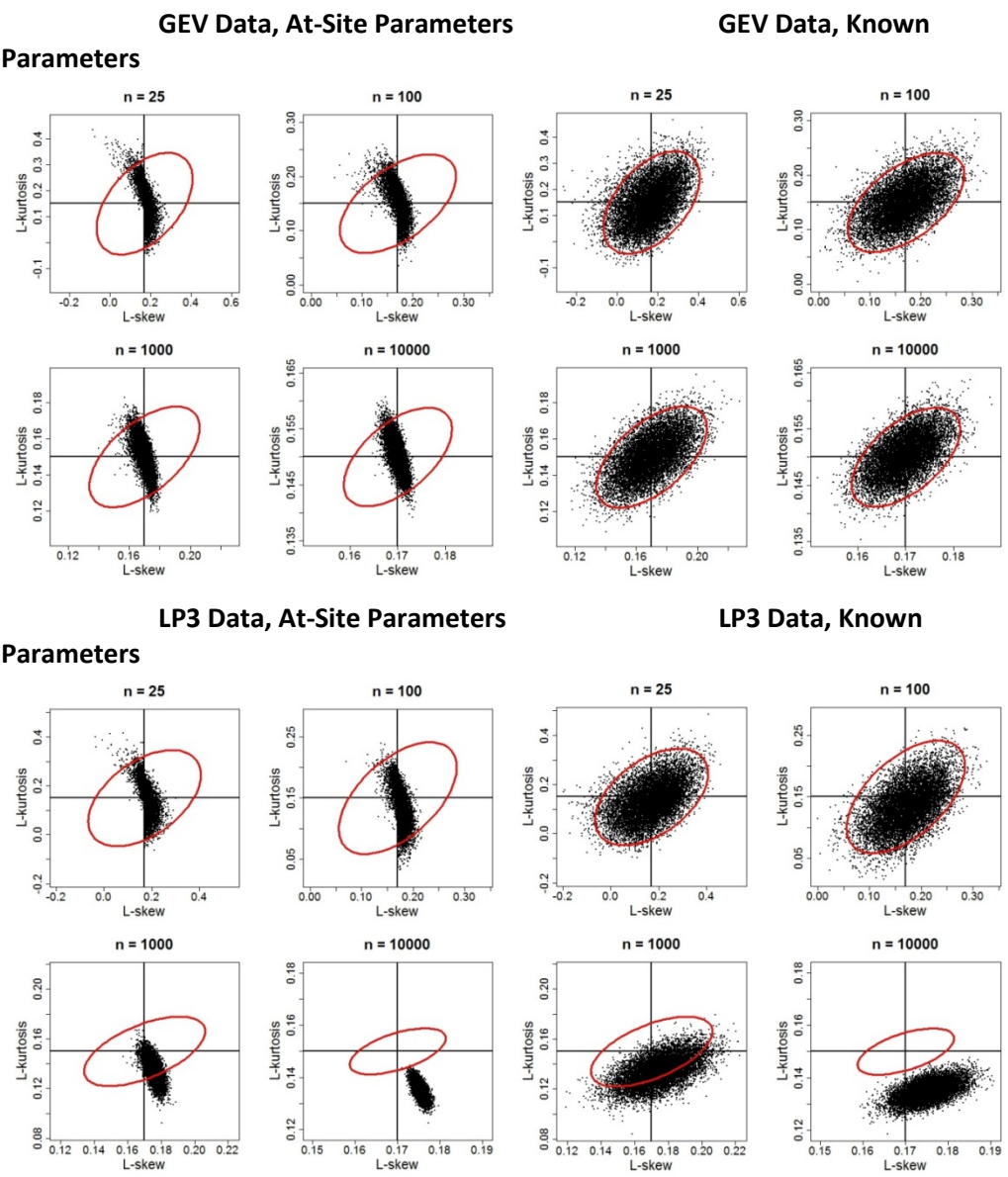

Figure 21. L-moment diagrams of the transformation Z estimated from synthetically generated GEV and LP3 for sample sizes equal to, n = 25, 100, 1,000, and 10,000. Left plots use at-site parameters for estimation of Z; right plots use known parameters.

8. CONCLUSIONS

We have introduced the concept of transformational concordance and applied it to better understand the ability of the GEV model to mimic the probabilistic behavior of both observed and synthetic flood data. In our initial evaluations of the 200 annual maximum flow records, we find that although GEV performs exceptionally well from a goodness-of-fit perspective, using at-site parameter estimates greatly underestimates the occurrence of the large events we care about most in flood frequency analysis. We then find that the 200 stations do not exhibit transformational concordance, concluding that either (a) the observed flood series are not concordant with the GEV hypothesis, or (b) sampling variability associated with parameter estimates is preventing us from evaluating the GEV hypothesis. We attribute both the discordance and errors in design event occurrence to systematic errors in estimation of the shape parameter κ of the GEV pdf.

Discordance appears to be considerably resolved by using regional κ estimates, underscoring recommendations from the body of literature to 'substitute space for time' by using hydrologic records at different locations to compensate for short records at a single site (IAWCD 1982, Cunnane 1988, Stedinger et al. 1992, Griffis et al. 2004, Griffis and Stedinger 2009). In most previous investigations, use of a regional estimate of κ for the GEV distribution or a regional estimate of skewness for the LP distribution have been advocated for the purpose of reducing the variance of estimates of the design event (Griffis and Stedinger 2009). Our findings indicate considerable bias in the design event may result when only at-site estimates are used, even when using long record lengths for parameter estimation.

9. FURTHER RESEARCH

Further research should address several issues that remain unresolved here. First, as the true value of κ for synthetically generated data series nears zero (i.e., GEV approaches Gumbel), why do at-site estimates of the parameters of those series result in transformed values Z which exhibit exactly the theoretical Gumbel L-skew in as shown in Figure 15? Second, what is the cause of the systematic parameter estimation errors that appear to explain the underestimation of design floods resulting from a fitted GEV pdf and what is the actual role of those errors in driving the underestimation? Lastly, further research should explore the usefulness and feasibility of constructing a hypothesis test based on transformational concordance. Such a test could draw on spatial homogeneity metrics commonly used in spatial statistics and cluster analysis.

REFERENCES

Asquith, W. H. 2011. Univariate Distributional Analysis with L-moment Statistics using R. Dissertation to the Department of Civil and Environmental Engineering, Texas Tech University.

Beard, L.R. 1982. Appendix 14: Flood Flow Frequency Techniques within Guidelines for Determining Flood Flow Frequency. *Bulletin 17B* of the Hydrology Subcommittee. Interagency Committee on Water Data, Washington, D.C.

Boehlert, B. 2015. Goodness-of-fit can be misleading. Tufts University Dissertation, Civil and Environmental Engineering.

Chowdhury, J.U., J.R. Stedinger, and L.H. Lu. 1991. Goodness-of-fit tests for regional generalized extreme value flood distributions. *Water Resources Research*, 27 (7): 1765–1776.

Cunnane, C., 1988. Methods and merits of regional flood frequency analysis. *J. Hydrol.*, 100: 269 290.

Douglas, E.M., R.M. Vogel, and C.N. Kroll, Trends in Flood and Low Flows in the United States, Journal of Hydrology, (240)1-2, pp. 90-105, 2000.

El Adlouni, S., B. Bobée, T.B.M.J Ouarda. 2008. On the tails of extreme event distributions in hydrology. *Journal of Hydrology.* 355: 16-33.

Favre, A-C., S. El Adlouni, L. Perreault, N. Thiemonge, and B. Bobee. 2004. Multivariate hydrological frequency analysis using copulas. *Water Resources Research.* 40(1): DOI: 10.1029/2003WR002456.

Garcia, D., J.R. Ramon Obeso, and I. Martinez. 2005. Spatial concordance between seed rain and seedling establishment in bird-dispersed trees: does scale matter? *Journal of Ecology*. 93(4): 693-704. DOI: 10.1111/j.1365-2745.2005.01004.x.

Griffis, V. W., J. R. Stedinger, and T. A. Cohn. 2004. Log Pearson type 3 quantile estimators with regional skew information and low outlier adjustments, *Water Resour. Res.*, 40, W07503, doi:10.1029/2003WR002697.

Griffis, VW and JR Stedinger. 2009. Log-Pearson Type 3 Distribution and Its Application in Flood Frequency Analysis. III: Sample Skew and Weighted Skew Estimators. *Journal of Hydrologic Engineering*. 14(2).

Gubavera, T.S. 2011. Types of Probability Distributions in the Evaluation of Extreme Floods. *Water Resources*. 38(7): 962–971.

Heo, J.H, Y.W. Kho, H. Shin, S. Kim, and T. Kim. 2008. Regression equations of probability plot correlation coefficient test statistics from several probability distributions. *Journal of Hydrology*. 335(1-4):1-15.

Hirsh, R. and K. Ryberg. 2012. Has the magnitude of floods across the USA changed with global CO_2 levels? *Hydrological Sciences Journal*. 57(1): 1-9.

Hosking, J. 1990. L-moments: Analysis and Estimation of Distributions using Linear Combinations of Order Statistics. *J. R. Statist. Soc. B*. 52(1): 105-124.

Hosking, J. and J. Wallis. 1997. Regional Frequency Analysis: An Approach Based on L-moments. Cambridge University Press.

IACWD (Interagency Committee on Water Data). 1982. Guidelines for Determining Flood Flow Frequency. *Bulletin 17B* of the Hydrology Subcommittee, Washington, D.C.

Jenkinson, A.F., 1955, The Frequency Distribution of the Annual Maximum (or Minimum) Values of Meterologic Elements, Quarterly Journal of the Royal Meteorological Society, 81, 158-171.

Johnson, N.L. and S. Kotz, 1995, Extreme Value Distributions, in Continuous Univariate Distributions, John Wiley & Sons, Inc., II, 1-112.

Laio, F., P. Allamano, and P. Claps. 2010. Exploiting the information content of hydrological "outliers" for goodness-of-fit testing. *Hydrol. Earth Syst. Sci.* 14: 1909-1917. doi:10.5194/hess-14-1909-2010

Legendre, P. 2005. Species Associations: The Kendall Coefficient of Concordance Revisited. *Journal of Agricultural, Biological, and Environmental Statistics.* 10(2): 226-245. DOI: 10.1198/108571105X46642.

Liao, J-J., Y-C. Wu, and K-S. Cheng. 2008. Establishing acceptance regions for L-moments based goodness-of-fit tests by stochastic simulation. *Journal of Hydrology*. 355: 49-62.

Lin, L.I. 1989. A concordance correlation coefficient to evaluate reproducibility. *Biometrics.* 45(1): 255-268.

Livezy, R.E., Chen, W.Y., 1983. Statistical field signifcance and its determination by Monte Carlo techniques. *Mon. Weather Rev*. 111, 46-59.

Kidson, R. and K.S. Richards. 2005. Flood frequency analysis: assumptions and alternatives. *Progress in Physical Geography*. 29(3): 392-410.

Martins, E.S. and J.R. Stedinger. 2000. Generalized maximum-likelihood generalized extreme-value quantile estimators for hydrologic data. *Water Resources Research*. 36(3): 737-744.

Matalas, N.C. and W.B. Langbein, 1962. Information content of the mean. *J. Geophys. Res.* 67 (9):3441-3448.

Mises, Richard, von, 1936, La Distribution de la Grande de n Valeurs, Review Mathematique Union Interbalcanique, 1, pp. 141-160. Reproduced in Selected Papers of Richard von Mises, II, American Mathematical Society, 1964, 271-294, Providence, Rhode Island.

Paynter, S. and M. Nachabe. 2010. Use of generalized extreme value covariates to improve estimation of trends and return frequencies for lake levels. *Journal of Hydroinformatics.* 13(1): 13-24.

Pisarenko, V.F., A. Sornette, D. Sornette, and M. V. Rodkin. 2014. Characterization of the Tail of the Distribution of Earthquake Magnitudes by Combining the GEV and GPD Descriptions of Extreme Value Theory. *Pure and Applied Geophysics*. 171(8):1599-1624.

Reis, D.S., J.R. Stedinger and E.S. Martins. 2005. Bayesian generalized least squares regression with application to log Pearson type 3 regional skew estimation, *Water Resources Research.* 41(10). W10419.

Renard, B., K. Kochanek, M. Lang, F. Garavaglia, E. Paquet, L. Neppel, K. Najib, J. Carreau, P. Arnaud, Y. Aubert, F. Borchi, J.-M. Soubeyroux, S. Jourdain, J.-M. Veysseire, E. Sauquet, T. Cipriani and A. Auffray, 2012. Data-based comparison of frequency analysis methods: a general framework. *Water Resources Rese*arch. 49: 825–843, doi:10.1002/wrcr.20087.

Salinas, J.L., A. Castellarin, S. Kohnova, and T.R. Kjeldsen. 2013. On the quest for a pan-European flood frequency distribution: effect of scale and climate. *Hydrol. Earth Syst. Sci. Discuss.* 10:6321-6358.

Sousa, P.M., R.M. Trigo, P. Aizpurua, R. Nieto, L. Gimeno, and R. Garcia-Herrera. 2011. Trends and extremes of drought indices throughout the 20th century in the Mediterranean. *Natural Hazards and Earth System Sciences.* 11: 33-51. doi:10.5194/nhess-11-33-2011.

Stedinger, J. R. 1983. Design Events with Specified Flood Risk. *Water Resources Research*. 19(2): 511-522.

Stedinger, J., Vogel, R., and Foufoula-Georgiou, E. 1992. Handbook of Hydrology, chap. 8: Frequency analysis of extreme events, McGraw-Hill, New York.

Stedinger, J.R. and L.H. Lu. 1995. Appraisal of regional and index flood quantile estimator. *Stochastic Hydrology and Hydraulics*, 9(1): 49-75.

Thompson, E.M., J.B. Hewlett, L.G. Baise, and R.M. Vogel. 2011. The Gumbel hypothesis test for left censored observations using regional earthquake records as an example, *Nat. Hazards Earth Syst. Sci.*, 11, 115-126, doi:10.5194/nhess-11-115-2011.

Villarini, G., J.A. Smith, F. Serinaldi, and A.A. Ntelekos. 2011. Analyses of seasonal and annual maximum daily discharge records for central Europe. *J. Hydrol.* 399: 299-312. Doi: 10.1016/j.jhydrol.2011.01.007.

Villarini, G., J.A. Smith, F. Serinaldi, A.A. Ntelekos, and U. Schwarz. 2012. Analyses of extreme flooding over Austria over the period f1951-2006. *International Journal of Climatology.* 32(8):1178-1192.

Vogel, R. 1986. The probability plot correlation coefficient test for the normal, lognormal, and Gumbel distributional hypotheses. *Water Resour. Res.* 22: 587–590.

Vogel, R. and McMartin, D. 1991. Probability plot goodness-of-fit and skewness estimation procedures for the Pearson type 3 distribution. *Water Resour. Res.* 27: 3149–3158.

Vogel, R.M., W.O Thomas, and T.A. McMahon. 1993a. Flood-Flow Frequency Model Selection in Southwestern United States. *Journal of Water Resources Planning and Management.* 119(3).

Vogel, R.M., T.A. McMahon and F.H.S. Chiew, Floodflow Frequency Model Selection in Australia, Journal of Hydrology, Vol. 146, pp 421-449, 1993b.

Vogel, R.M. and N.M. Fennessey. 1993. L-moment Diagrams Should Replace Product-Moment Diagrams, *Water Resources Research*, Vol. 29, No. 6, pp 1745-1752.

Vogel, R.M. and I. Wilson. 1996. The Probability Distribution of Annual Maximum, Minimum and Average Streamflow in the United States, *Journal of Hydrologic Engineering, ASCE*, 1(2):69-76.

Vogel, R.M., J.R.M. Hosking, C.S. Elphick, D.L. Roberts, and J.M. Reed. 2009. Goodness-of-fit of Probability Distributions for Sightings as Species Approach Extinction, *Bulletin of Mathematical Biology*, DPO 10.1007/s11538-008-9377-3.

APPENDIX A: CATEGORIES OF CONCORDANCE

We identify five types of concordance that can assist in evaluating whether a particular hypothesis is consistent with observed data:

1. **Transformational concordance**. A data series is transformationally concordant if its properties under each transformation (e.g., real space to log space) are consistent with our theoretical understanding of how those properties change across transformations. For example, in real space, if a sequence of observations is presumed to be distributed as lognormal and if the logarithms of the observations yield zero skewness, the skewness would not contradict the normal distribution and therefore, the lognormal distribution in real space and the normal distribution in log space would be transformationally concordant. On the other hand, if the logarithms of the observations yield significantly non-zero skewness, then the two distributions would be transformationally discordant, indicating that the original data may not be lognormally distributed.
2. **Frequency concordance:** A data series is said to be frequency concordant if the number of events predicted by some fitted pdf is consistent with the number of such events expected.
3. **Threshold concordance.** A data series is threshold concordant if the behavior of its pdf is consistent across thresholds. For example, it is well known that if the PDS series are made up of Poisson arrivals and an exponential distribution of the magnitudes above the threshold, then the AMS series which results should be consistent with a Gumbel model. Testing if all these assumptions hold, together, would constitute a test of threshold concordance.

4. **Geographic concordance.** A data series is geographically concordant if its probability distribution is consistent across geographic regions. Here, concordance may depend on scale. For example, although we may conclude that floods in the Pacific Northwest follow a GEV distribution (i.e., geographic concordance), we may not reach this conclusion looking across the entire U.S. (i.e., geographic discordance).
5. **Stationarity concordance.** A data series exhibits stationarity concordance if its pdf is temporally consistent. If the pdf is not stationarity concordant, this may point to evidence of change due to anthropogenic influences.

WATER UNDER A CHANGING AND UNCERTAIN CLIMATE: LESSONS FROM CLIMATE MODEL ENSEMBLES

Brent Boehlert[1,2], Susan Solomon[3], Kenneth M. Strzepek[3]

Author affiliations: 1. Tufts University, Department of Civil and Environmental Engineering, Medford, MA
2. Industrial Economics, Inc., Cambridge, MA
3. Massachusetts Institute of Technology, Cambridge, MA

ABSTRACT

Climate change and rapidly rising global water demand are expected to place unprecedented pressures on already strained water resource systems. Successfully planning for these future changes requires a sound scientific understanding of the timing, location, and magnitude of climate change impacts on water needs and availability – not only average trends, but also interannual and decadal variability and associated uncertainties. This study focuses on new information and its use to better understand these uncertainties. In recent years, two types of large ensemble runs of climate projections have become available, those from groups of more than 20 different climate models, and those from repeated runs of several individual models. These provide the basis for novel probabilistic evaluation of both climate change and the resulting effects on water resources. Using a range of available climate model ensembles, this research explores the spatial and temporal patterns of high confidence as well as uncertainty in projected river runoff, irrigation water requirements, and basin storage yield. Cost estimates of adapting global water supply systems are developed for each ensemble, and implications for water management are discussed.

1 INTRODUCTION

Due to rising temperatures and changing and more variable precipitation patterns, climate change will significantly affect the patterns of regional and global water availability and demand. Combined with rapidly rising water demand associated with global economic development, these changes will place unprecedented pressures on already strained water resource systems. Successfully planning for future changes that could exceed past variability and hence impact water availability in unprecedented ways requires a scientific understanding of the timing, location and magnitude of climate change, not only of average trends, but also of interannual and decadal variability and associated uncertainties. To develop local and regional adaptation responses to water resource challenges that are robust to this wide range of future conditions, it is essential to characterize the extent of these uncertainties (Lempert and Groves 2009).

In recent years, two types of large ensemble runs of climate projections have become available, those from groups of more than 20 different General Circulation Models (GCMs) that have been distributed to the community via the Coupled Model Intercomparison Project (CMIP), and those from repeated runs of several individual models. Examples of the latter include ensembles of 40 and 17 members that are available from the National Center for Atmospheric Research's Community Climate System model (CCSM3), and those of the Max Planck Institute's ECHAM climate models (Deser et al. 2012). We henceforth refer to this type of ensemble as *within-model*, as opposed to *between-model* ensembles that include runs from different modeling systems. Additional within-model ensembles are becoming available (e.g., using the fast Earth system model of the Hadley center, FAMOUS) or will become available within the next few years, and many other smaller within-model ensembles are available through

in the recent release of the CMIP5 archive. In this analysis, we compare within-model results from the CCSM, ECHAM, and several CMIP5 ensembles, as well as the between-model results from the full set of 23 different CMIP5 models.

These ensembles have provided new approaches to the probabilistic evaluation of both climate change and the resulting effects on water resources, and thus improve our understanding of the timing and location of prudent climate adaptation measures. Using a range of ensembles, this research explores the spatial and temporal patterns of uncertainty in projected river runoff, irrigation water requirements, and basin storage yield. Basin storage yield implications are translated to regional and global adaptation cost estimates for each ensemble, and the resulting implications for water management are discussed. While global-scale climate trends among ensembles are relatively robust, local scale trends are much less so. Precipitation is extremely variable at local scales, while temperatures are considerably less variable (e.g., from one model grid point to another). River runoff, irrigation requirements, and basin storage yields involve precipitation averaged over the spatial scale of the basin, and are also dependent on temperature through evaporation. Therefore, a central hypothesis of our work is that by integrating precipitation effects over space and time, projections of water resource variables will tend to have higher levels of within- and cross-ensemble agreement than precipitation.

Several recent studies evaluated patterns of precipitation and temperature uncertainty using recent within-model ensembles, and some considered the timing of signal emergence relative to the noise of climate variability. Solomon et al. (2009) identify patterns of regional agreement within precipitation patterns across the 22 CMIP3 SRES A1B scenarios, and similarly, Deser et al. (2013) show strong similarities in

the patterns of variability between the 40-member and 17-member ensembles of the NCAR and ECHAM models. From a signal-to-noise perspective, however, Mahlstein et al. (2012) evaluate the emergence of projected precipitation signals from noise at a more local level, and find few grid cells where emergence occurs. In explaining the sources of uncertainty, Hawkins and Sutton (2009) identify three types—model uncertainty, emissions uncertainty, and internal variability—and suggest that both model uncertainty and internal variability decline relative to the signal as future emissions increase in the latter part of the 21st century. Following on this theme, Hawkins and Sutton (2012) showed that many of the apparent differences in future climate change on shorter timescales (next few decades to mid-century) across the different CMIP3 models likely result from internal variability rather than modeling uncertainty.

Several previous studies have evaluated the effects of climate change on global water supply and demand. Vörösmarty et al. (2000) use the Water Balance Model (WBM) to analyze the effect of climate change on global runoff at the 0.5 x 0.5 degree grid scale, but only relied on two GCMs to generate projections. Similarly, Alcamo et al. (2007) used the WaterGAP model to compute monthly river discharge and worldwide water availability under climate change at both a grid and basin scale, but only used two GCMs under the A2 and B2 SRES scenarios. Arnell (2004) studied the effects of 24 climate scenarios on future runoff in 1,300 global basins, but only focused on a single between-model ensemble and did not extend the work beyond runoff. Milly et al.'s article (2005) considered the outputs of 12 GCMs in their analysis of how climate change will affect runoff in 163 river basins, but also did not focus on within-model ensembles. More recently, Strzepek et al. (2013) evaluated the effects of climate change under 56

different model-SRES scenario combination runs on a set of six hydrological indicators, and found that model uncertainties in river runoff and irrigation water demand tended to be higher in developing than in developed countries. Using the same set of 56 runs in a study of projected U.S. drought patterns, Strzepek et al. (2010) find that measures of drought that incorporate temperature rather than focusing on precipitation only (i.e., Palmer Drought Severity Index versus Standardized Precipitation Index) produce much greater between-model agreement due to the cross-model agreement in temperature trends. Konzmann et al. (2013) also provided a detailed investigation of the impact of climate change on irrigation water demand for a range of GCMs.

Research has been conducted on the economic effects of climate change on water resource outcomes, but typically have not used the broad range of climate models. Ward et al. (2010) investigated the potential costs of maintaining reservoir supply yield globally, but only evaluated outcomes using two climate models. In this paper, we repeat components of this analysis and demonstrate that using different model runs, even from the within-model ensemble used for one of the two scenarios, could generate a much different outcome. The Ward et al. study emerged from a broader World Bank program called the *Economics of Adaptation to Climate Change* (EACC), which estimated the costs of adapting to climate change in developing countries at $100 billion per year in 2050 (World Bank 2009). In more recent work, the costs of flooding and droughts (Strzepek et al, *in review;* Boehlert et al., *in review*) have been estimated using a set of model outputs derived from the NCAR Community Atmosphere Model (CAM) outputs.

In the following paragraphs, we document the methodologies used to investigate patterns of uncertainty across model ensembles, and then provide results of our analysis. We conclude with a discussion of results and recommended further research.

2 METHODOLOGIES

In order to translate a suite of climate model outputs into projected effects on water availability and demand, a broad set of data source and modeling approaches is required. As portrayed in Figure 22, the ensembles of General Circulation Models (GCMs) are the basis of the analysis process. Projected changes in monthly temperature and precipitation for the 21st century period were collected from 220 GCM-emissions scenario combinations available through the CMIP3 and CMIP5 archives and from the CCSM3 and ECHAM ensembles referenced above. Changes in these parameters were calculated from an historical baseline of 1961 to 1990. These 220 GCM runs, which incorporate several greenhouse gas emissions scenarios and dozens of modeling frameworks, contain five between-model ensembles and 12 within-model ensembles and thus reflect a wide variability in possible spatial and temporal distribution of precipitation and temperature outcomes. Here we focus on two "eras": 2040-2059 and 2080-2099 (referred to as the 2050 and 2090 eras, respectively). The first era is a relevant time-scale for current water infrastructure planning, and the second provides a means to evaluate later signal emergence in many regions where emergence does not occur by mid-century. As a final step in climate model processing, we normalized the climate runs to a common climate sensitivity to ensure that the precipitation and temperature projections in the models probe the impact of variability.

For the water modeling components, we next combined the projected precipitation and temperature changes with historical data from the 1961 to 1990 baseline to produce absolute temperature and precipitation projections for each basin. These absolute temperature and precipitation projections were used to estimate potential evapotranspiration (PET) using the Modified Hargreaves model (Allen et al., 1998, Droogers and Allen, 2002). PET, together with projected precipitation and temperature, was then used to project irrigation water demand using the FAO 56 model, and river runoff projections using the climate runoff model (CLIRUN)-II, a two-layer, one-dimensional rainfall-runoff model. Lastly, basin storage yield was calculated based on annual runoff, and existing basin storage using the sequent peak algorithm.

In the following paragraphs, we describe (1) characterizing the baseline conditions for the analysis, including datasets and issues of scale and resolution; (2) processing of climate model ensembles; (3) modeling runoff using CLIRUN-II; (4) modeling irrigation water demand; and (5) basin yield modeling.

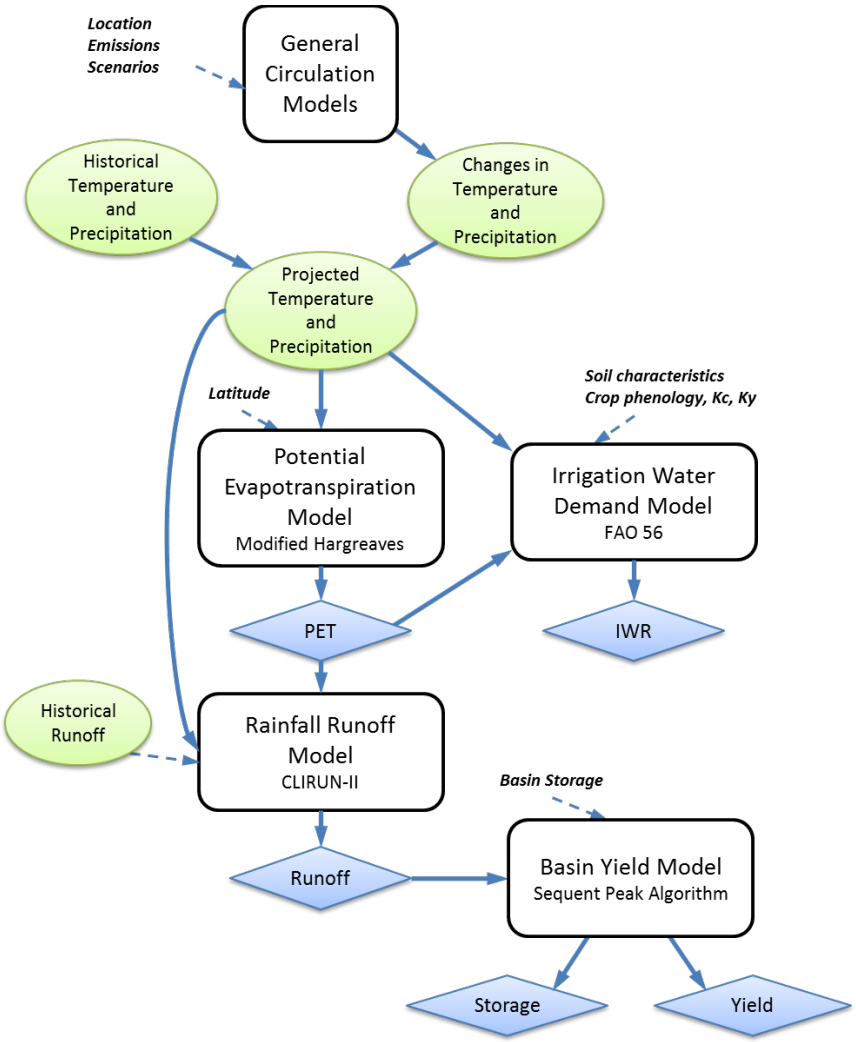

Figure 22. Diagram of modeling processes

2.1 Characterizing baseline conditions

As this study uses gridded data of varying resolutions in order to model outcomes at the basin scale, both data processing methods and baseline dataset requirements are intensive. Baseline temperature, precipitation, and runoff data have a resolution of 0.5 x 0.5 degrees, and are coupled with climate model outputs of resolutions between 1 x 1 degree and 4 x 5 degrees. These are then spatially averaged to 8,951 river basins of the world for runoff and irrigation water demand modeling, and then those 8,951 basins are

aggregated to 126 major river basins for basin yield analysis. Choices of spatial resolution and aggregation, PET calculation methods, and the baseline climate and runoff datasets are discussed below.

2.1.1 Scale and resolution

There is a trade-off between precision and accuracy when deciding on spatial and temporal resolution (Figure 23). A smaller scale analysis requires information that is more detailed, which means higher relative error, whereas a larger scale analysis allows for greater accuracy, but may not provide necessary levels of spatial or temporal precision.

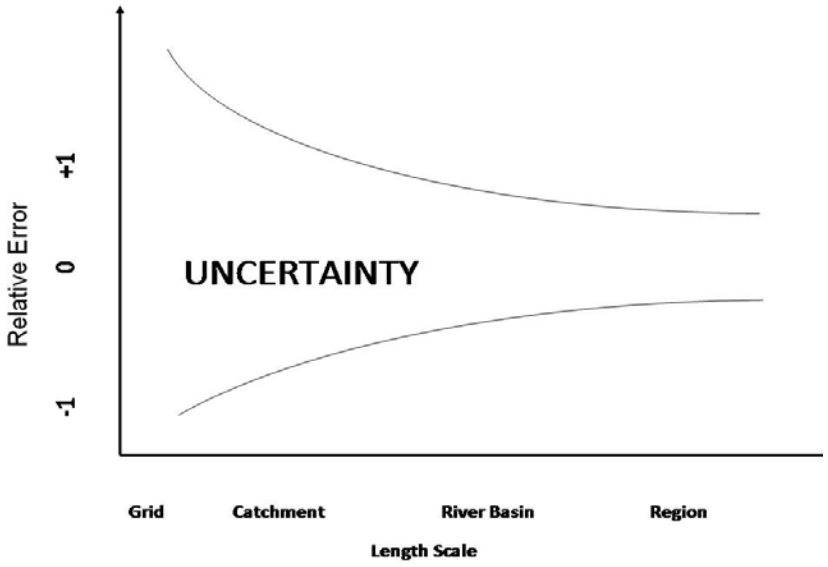

Figure 23. The cone of uncertainty in scale and resolution of modeling (source: Strzepek et al. 2011)

This study employs two river basin resolutions that strike a balance between precision and accuracy, and are appropriate for the respective analyses in which they are used. The first is 8,951 river basins of the world, which were developed by Strzepek

et al. (2011) using the Hydro1k data set from the US Geological Survey (USGS) for geographic delineation of basin boundaries (Figure 24). The Hydro1k dataset is currently the best available for global river basin delineation (Strzepek et al. 2011). The basins in the raw Hydro1k dataset range significantly in size, from the smallest catchments of less than one square kilometer to drainage areas for rivers such as the Nile or Amazon that are well over the typical grid scale of a GCM (that is, 2.5 x 2.5 degrees). The 8,951 basins were selected to be no smaller than the resolution of available baseline climate data (0.5 x 0.5 degree), and thus range in size from approximately 2,500 km^2 (which is similar to a baseline data grid cell of 0.5 x 0.5 degrees), to more than 62,500 km^2 (which is similar to a climate model grid cell of 2.5 x 2.5 degrees). For the basin yield analysis, the river runoff projected for these basins is aggregated up to 126 major river basins of the World used in the IFPRI IMPACT model (Figure 24). Data on reservoir storage are available for these larger basins.

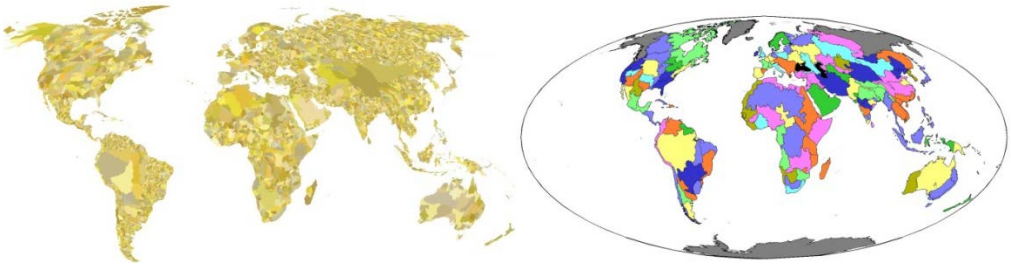

Figure 24. Map of the 8951 river basins (left) and 126 river basins (right) used in this study

2.1.2 Baseline temperature and precipitation data

Historical data is needed in this study for two reasons: (1) to model historical PET, runoff, and irrigation water demand that serve as a basis for calculating changes in those variables, and (2) to develop bias corrected GCM projections so that they are

consistent with observed data. Baseline precipitation, temperature, and daily average temperature range data for the 1961-1990 period were from the University of East Anglia's Climate Research Unit (CRU) Time Series (TS) 2.1 data set. These three variables are needed for estimation of PET using the Modified Hargreaves formulation, and for runoff using CLIRUN-II. The CRU TS data sets are the standard reference baselines for the World Meteorological Organization, and provide a monthly time series of these variables on a 0.5 x 0.5 degree grid. For the 8,951 basins, mean annual precipitation ranges from <1 to over 6,000 millimeters (mm), and temperature ranges from -21 to 30 degrees Celsius (Figure 25).

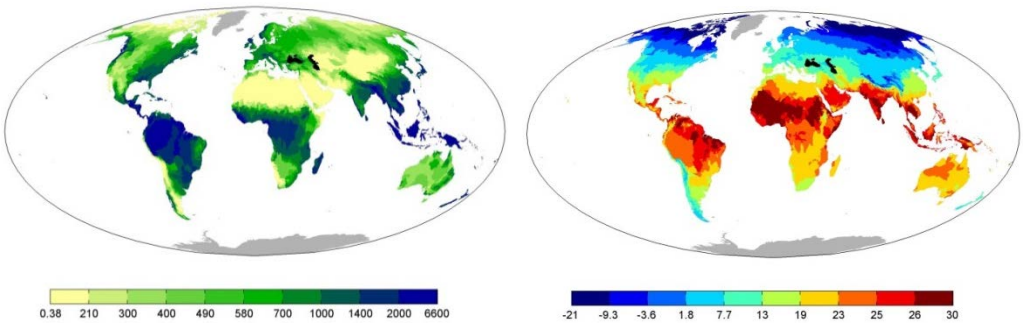

Figure 25. Observed mean annual precipitation (left; mm) and temperature (right; degC) over the period 1961-1990; CRU TS 2.1 dataset spatially averaged to the 8,951 river basins

2.1.3 Potential evapotranspiration calculation

PET is one of the key inputs to both the irrigation water demand and runoff models, and represents the amount of water lost through evaporation and transpiration assuming that sufficient water is available over the period in question. PET depends on several variables, including temperature, wind speed, solar radiation, and the range of daily temperatures, and can be estimated using one of many methods, including Penman-Monteith, Harmon, Hargreaves, and more recently, Modified Hargreaves (see Allen et al. 1998 for a description of these methods).

If sufficient data are available, the U.N. Food and Agriculture Organization (FAO) recommends using Penman-Monteith, which requires the full set of PET input variables described above. When only precipitation and temperature information are available, however, FAO recommends using a less data intensive method. Because GCMs do not reliably reproduce certain key climate variables needed for the Penman-Monteith method (wind speed, most notably), in this study, we use the Modified Hargreaves approach (Allen et al. 1998, Droogers and Allen 2002). Modified Hargreaves relies on precipitation, temperature, and average daily temperature range data, along with the latitude of the basin centroid, which is used to estimate solar radiation.

2.1.4 Baseline runoff data

Rainfall runoff models simulate the relationship between precipitation (rain and snow) and natural, unmanaged runoff. As such, these models require natural runoff data to calibrate the simulated runoff outputs. This analysis relies on two sources of baseline runoff data. The first is a global gridded dataset of historical average monthly runoff from the Global Runoff Data Center (Fekete et al., 2002). This dataset is derived from a water balance model that relies on observed discharge information (Figure 26). The GRDC dataset preserves the accuracy of measured point discharge, and employs a gridded river network at a 0.5 x 0.5 degree resolution to represent river pathways and to link continental landmasses to oceans through river channels (Strzepek et al. 2011). The dataset provides 12 monthly mean values for each grid cell, and is currently the best globally available source of terrestrial runoff data. Other datasets are being developed (e.g., McMahon et al. 2007), but are not yet available for use in a global runoff study.

The second source of runoff data is from the International Food Policy Research Institute (IFPRI), and provides a time series of monthly runoff data for the 282 IFPRI Food Producing Regions (FPUs) of the world, which are intersections between the 126 major global river basins and country boundaries. The data are available from 1950 and 2000, and as described below, are used before applying the basin yield analysis to bias correct the spatially aggregated runoff outputs from CLIRUN-II.

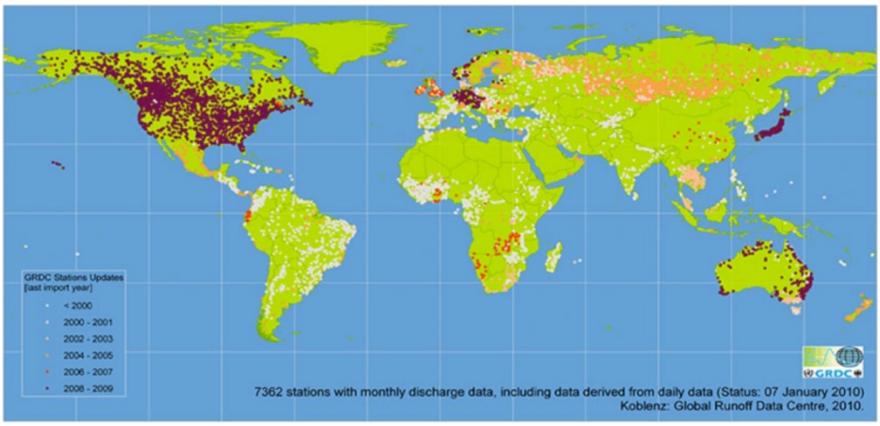

Figure 26. Locations of gauging stations used in the GRDC database. Source: (Fekete et al., 2002, GRDC, 2007)

As with temperature and precipitation, baseline gridded GRDC runoff data were processed to the 8,951 basins through spatial averaging. Runoff data at the 282 FPUs are aggregated through spatial averaging to the 126 basins. The resulting global spatial patterns of runoff depth in these two datasets map closely to each other and to precipitation (Figure 27).

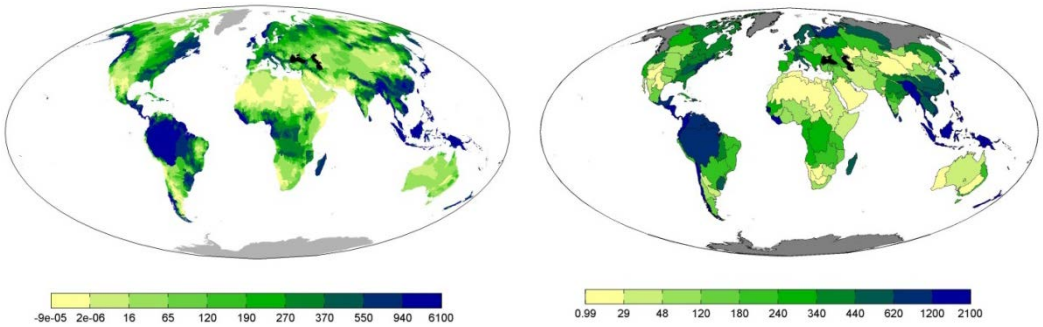

Figure 27. Annual Observed Runoff (mm) in the 8,951 river basins (GRDC, at left) and in the 126 river basins (IFPRI, at right). GRDC is the sum across the 12 available average values; IFPRI data is the annual average over the 1961-1990 baseline period.

2.2 Climate ensembles: description and processing approach

The central focus of this study is on the patterns and sources of uncertainty in a broad set of within-and between-model climate models. The characteristics of these ensembles, and the processing procedures necessary to (a) normalize for differences in emissions and climate sensitivity assumptions and then (b) combine the resulting projections with baseline data of a different spatial resolution, are described below.

2.2.1 Description of model ensembles

This study relies on model ensembles from both IPCC's CMIP3 and CMIP5 archives (available in 2007 and 2014, respectively), as well as two independently generated large ensembles (mentioned above) that rely on an emissions scenario employed by other CMIP3 models. We consider 17 different ensembles that include a total of 220 model runs, which are organized by their model series (either CMIP3 or CMIP5), emissions scenario, and whether they are in a between- or within-model ensemble (Table 1).

Of the total, the CMIP3 series contains seven of the 17 ensembles that include 144 runs, and the remaining 10 CMIP5 ensembles cover 96 runs. There are five between-model ensembles, one for each of the available emissions scenarios, and 12 within-

model ensembles. Model spatial resolutions vary from 1 x 1 degree to 4 x 5 degrees, and we processed data for all ensembles over the 1900 to 2099 period (this period includes the modeled baseline). Note that the 40-member CCSM3 ensemble extends from 1900 to 2061, and the 17-member ECHAM ensemble extends from 1950 to 2100. Also of note, the spatial resolution of the large CCSM3 ensemble is 2.8 x 2.8 degrees, differentiating it from the higher resolution but otherwise identical model included in the CMIP3 archive, which has a resolution of 1.4 x 1.4 degrees.

Table 1. Characteristics of climate model ensembles relied upon in this study

Model Series	Ensemble Name/ Emissions Scenario	Ensemble Type	Members	Time Period Available
CMIP3	**SRES B1**	Between	17	1900-2099
	SRES A1B	Between	22	1900-2099
	SRES A2	Between	17	1900-2099
	NCAR CCSM A1B	Within	40	1900-2061
	NCAR A1B	Within	7	1900-2099
	MPI ECHAM5 A1B	Within	17	1950-2099
	MPI ECHAM5 A1B	Within	4	1900-2099
CMIP5	**RCP4.5**	Between	23	1900-2099
	RCP8.5	Between	20	1900-2099
	CCSM RCP4.5 and 8.5	Within	6 each	1900-2099
	CSIRO RCP4.5 and 8.5	Within	10 each	1900-2099
	CAN RCP4.5 and 8.5	Within	5 each	1900-2099
	GISS-R RCP4.5	Within	6	1900-2099
	GISS-H RCP4.5	Within	5	1900-2099

The five emissions scenarios that encompass the 220 runs include the B1, A1B, and A2 storylines from the Special Report on Emissions Scenarios (SRES) report employed in the 2007 4th Assessment Report, as well as two Representative Concentration Pathway (RCP) scenarios at stabilization levels of 4.5 watts/m^2 and 8.5 watts/m^2 of forcing. By the end of the 21st century, the ranking of total volume of global emissions among these scenarios from lowest to highest is B1, RCP4.5, A1B, A2, and RCP8.5, as reflected in the temperature projection outcomes of these five between-model ensembles (Figure 28).

Note that this ordering does not necessarily hold for periods other than the end of century, as the shapes of the emissions trajectories also vary over time.

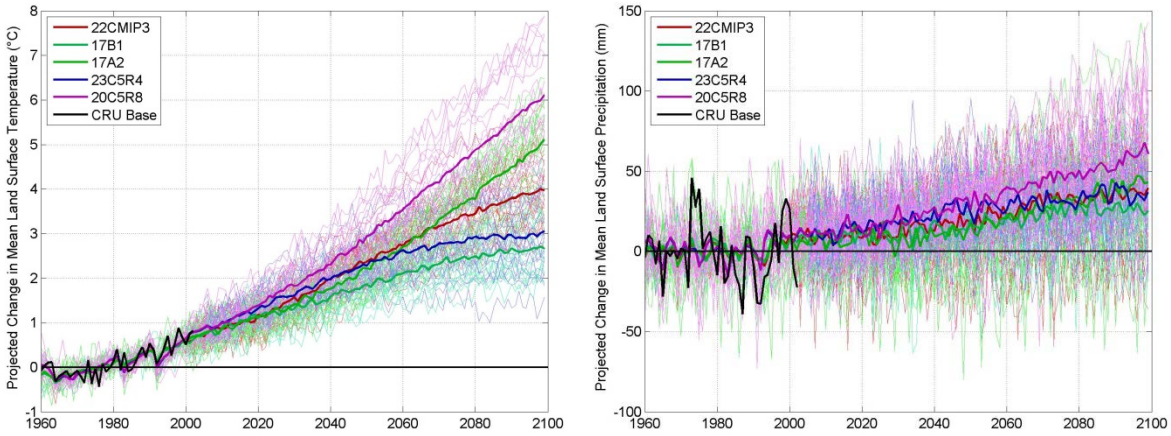

Figure 28. Projected change in mean land surface temperature (left) and precipitation (right) from the 1961-1990 baseline, across the members of between-model CMIP3 and CMIP5 ensembles included in this study. Black line is average CRU TS2.1 observational baseline.

These 17 model ensembles are used for different purposes in this study depending on the particular question being asked. As documented in the results below, the majority of analysis is conducted on three of the larger within-model ensembles (40-member CCSM3, 17-member ECHAM, and 10-member CSIRO RCP4.5), and three of the between-model ensembles (22-member A1B, 23-member RCP4.5, and 20-member RCP8.5). This set of within-model ensembles is selected to ensure that an adequate number of members is available to allow for statistical comparisons, and the between-model ensembles are selected to provide a linkage between the emissions scenario utilized in two of the large within-model ensembles (SRES A1B), and the outcomes projected in the more recent set of CMIP5 models.

2.2.2 Climate sensitivity and emissions normalization

Two important drivers of climate model outcomes are the assumed emissions trajectory (described above) and internal assumptions about climate sensitivity. Climate sensitivity is the model-calculated global temperature response to a doubling of atmospheric carbon equivalents, and is a key uncertainty in climate modeling. In order to focus on uncertainties derived from model structure and internal variability, we controlled for differing emissions and climate sensitivity by normalizing each GCM run to the mean global temperature and precipitation trajectory of the 22-member SRES A1B CMIPs3 ensemble. The normalization procedure involved five steps for temperature (in Kelvin) and precipitation (in mm) and for each of the 220 GCM runs (see Figure 29):

(1) develop a 2000 to 2099 trajectory for the globe of grid cell area-weighted means;

(2) divide the trajectory of absolute projections (in Kelvin and mm) by the mean model baseline to develop a trajectory of ratios relative to the modeled base;

(3) fit a fourth order polynomial to each trend of ratios;

(4) divide the fitted trajectory of each GCM run by that of the fitted A1B CMIP3 mean to develop a new trajectory of deviations from the mean; and

(5) apply the ratios from Step 3 to the original GCM run to normalize.

The results of this procedure are 220 GCM runs that share a common mean global temperature and precipitation trend (Figure 30).

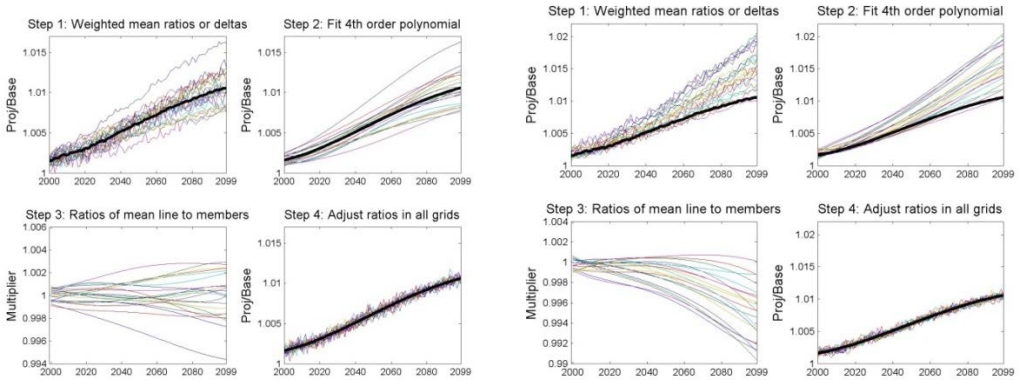

Figure 29. Climate sensitivity adjustment procedure for 22 A1B SRES CMIP3 models (left) and application of the A1B CMIP3 trend to the ensemble of 20 CMIP5 models under RCP 8.5 (right)

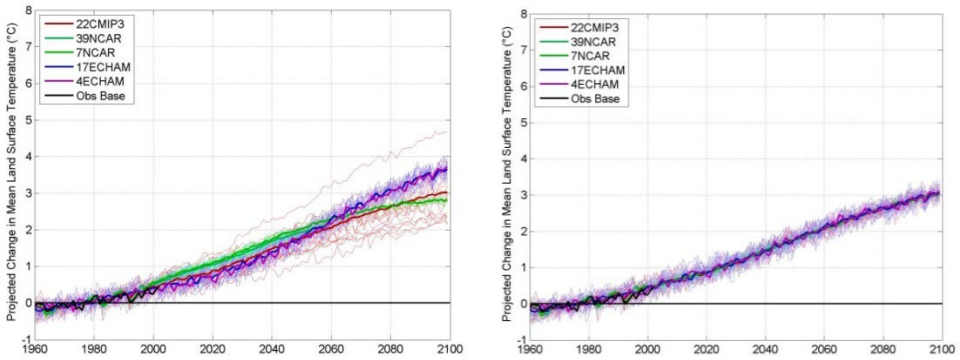

Figure 30. Projected changes in global temperature through 2100 under five A1B ensembles (left) and the climate-sensitivity adjusted trends (right)

2.2.3 Combining the climate model projections with baseline climate data

Because the spatial resolution of the 8,951 basins is generally finer than that of GCMs, it was necessary to match the lower resolution GCM output with the higher resolution basin scale. Available approaches for downscaling include statistical downscaling, or the use of empirical relationships; dynamical downscaling, or the use of regional climate models; and spatial techniques, such as linear interpolation or krigging. Given the number of climate model runs employed in this analysis and the fact that analytically correct dynamical and statistical downscaling techniques require extensive modeling and analysis for each GCM run, we employ a spatial technique in this study.

To allow for a common starting grid resolution, the 220 GCM projections were first re-gridded from their native spatial grid scale, which ranges from approximately 1x1 degree to 4x5 degree, onto a common 2 x 2 degree grid. This procedure involved spatially averaging the raw resolutions, which have varied latitude band widths near the poles, to the common resolution. This initial step allowed the temperature and precipitation patterns among the GCMs to be readily compared across both land and ocean surfaces.[1] For the water resources analyses, we then mapped these 2 x 2 degree projections for all land surfaces directly onto the 0.5 x 0.5 degree resolution of the baseline climate data. This approach captures the range of potential climate change impacts at a higher resolution without downscaling the GCMs themselves. We then aggregating these gridded data to the basin scale by overlay basin boundaries with the 0.5 x 0.5 degree grids, and then aggregating cells based upon their weighted area in each basin. For illustration, Figure 31 shows the 0.5 x 0.5 degree grid, the 2 x 2 degree grid, and the basins over the Horn of Africa.

[1] The resolution of 2 x 2 degrees was selected prior to the release of the higher resolution CMIP5 GCMs, 2 x 2 being finer than the majority of GCM resolutions in the CMIP3 ensemble. As a result, some information will be lost in averaging for any GCMs with resolutions under 2 x 2 degrees, which would affect the spatial detail of the higher resolution runoff analysis, but would be unlikely to affect broad analytical findings. For future work, a common resolution of 1 x 1 degree would be more appropriate.

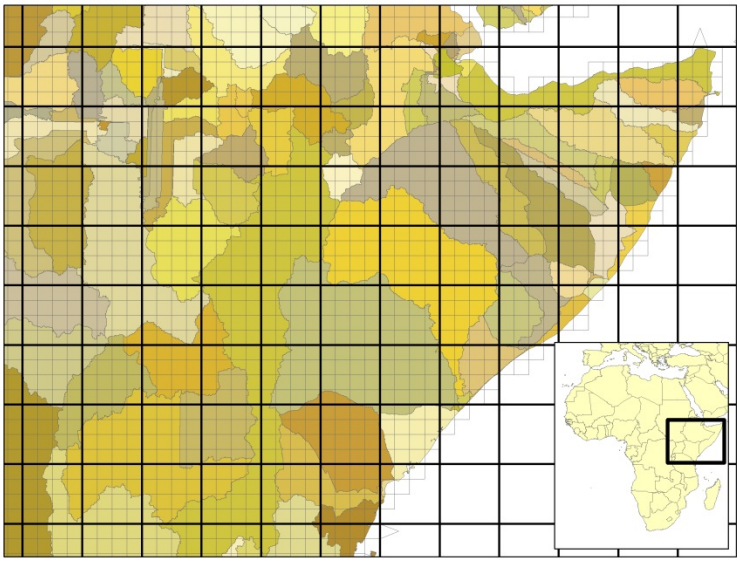

Figure 31. The horn of Africa showing an overlay of river basins, 0.5 x 0.5 degree grids, and 2 x 2 degree grids

Next, basin-scale projected changes in precipitation and temperature for the 2050 and 2090 eras were combined with basin-scale monthly precipitation and temperature data to generate absolute projections. This procedure requires that the projected changes be adjusted for bias within each model run, where bias is any statistical differences between the observed climate baseline and the modeled baseline. To do this, we first averaged the basin-scale modeled baseline and projections by month, so that within each grid cell, variable, and GCM run, there were 12 mean monthly outputs for the 30-year baseline and each 20-year future era. The changes in each variable were generated using the delta method, which subtracts the mean monthly modeled baseline from the projected values to produce delta temperature and precipitation.[2] Lastly, we added these projected changes to the gridded basin-scale CRU data to generate absolute monthly projections for each GCM run.[3]

[2] The delta approach has been applied widely in water planning studies (e.g., Hamlet et al., 2010, Sutton et al. 2011).

[3] Note that if the change in precipitation for a particular month and grid cell is negative and greater in absolute value than the observed value, the resulting absolute projection would be negative. For example,

2.3 Runoff modeling

This study employs CLIRUN-II to model changes in runoff under each of the GCM runs. CLIRUN-II (Strzepek and Fant 2010, Strzepek et al. 2011) is a one-dimensional infiltration and runoff estimation tool that uses historic runoff as a means to estimate soil characteristics. It is the latest model in a family of hydrologic models developed for the analysis of climate change impacts on runoff. Kaczmarek (1993) presented the theoretical development for CLIRUN, a single-layer, lumped, watershed rainfall runoff model, which he applied to the Yellow River in China (Kaczmarek, 1998). A snow-balance model and suite of PET models were added and the model was re-named WatBal (Yates 1996), which has subsequently been used on a wide variety of spatial scales from small and large watersheds to globally (e.g., Huber-Lee et al., 2005, Strzepek et al., 2005). CLIRUN-II builds on the CLIRUN and WatBal frameworks by addressing the issue of modeling extreme events at the monthly and annual level. CLIRUN-II follows the framework of the six-parameter (SIXPAR) hydrologic model (Gupta and Sorooshian, 1983, 1985) by adopting a two-layer approach, and employs unique conditional parameter estimation procedures.

2.3.1 Model inputs and structure

CLIRUN-II requires monthly precipitation, temperature, PET, and observed runoff. Baseline climate variables and observed runoff are used for calibration, and both the baseline and projected climate variables are subsequently used for generation of

---

if modeled base precipitation in a grid cell and month is 80 mm and projected value is 30 mm, then the projected change would be a decrease of 50 mm. If observed precipitation in that month and grid cell is 40 mm, then the absolute precipitation would be -10 mm. In these cases, projected precipitation is set to 1 mm.

simulated modeled runoff outputs. Runoff is treated as a lumped watershed, with average climate inputs and soil characteristics over the watershed, and runoff simulated at the mouth of the basin. Reported outputs include surface runoff, subsurface runoff, baseflow, and total runoff, where the total is the sum of the first three.

In CLIRUN-II, water enters the model via precipitation and leaves through ET and runoff generation (Figure 32). Differences between inflow and outflow accumulate as storage in the soil or groundwater. The model treats soil moisture as a two-layer system: a soil layer (upper layer) and a groundwater layer (lower layer), which correspond to quick and slow runoff responses to effective precipitation (i.e., precipitation plus snowmelt). Quick runoff is the portion of the effective precipitation that enters the stream system directly as surface runoff; direct runoff depends on the soil surface, and is modeled differently for frozen soil versus non-frozen soil (driven by temperature). Slow runoff is then generated by the remaining effective precipitation, which infiltrates into the soil layer and leaves as subsurface runoff, groundwater, or soil storage through a set of nonlinear equations. Subsurface runoff is linearly related to soil water storage, and percolation is nonlinearly related to both groundwater and soil storage. The soil layer percolates to groundwater, and baseflow is produced as a linear function of groundwater storage.

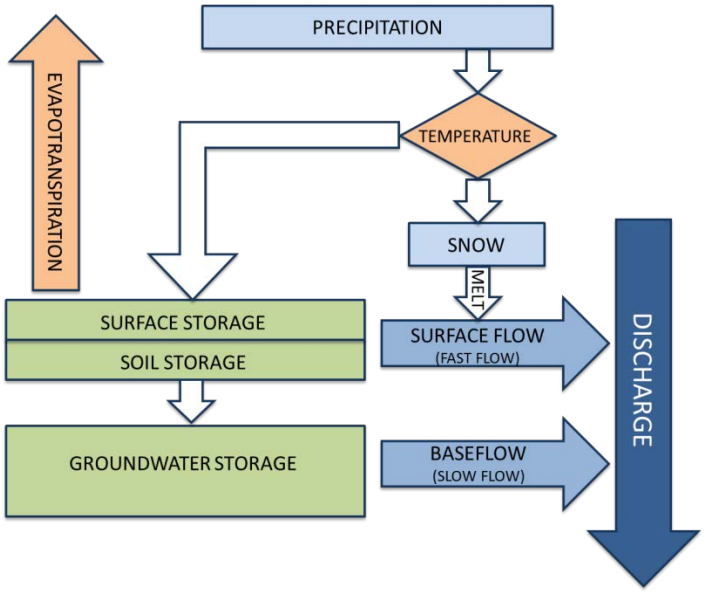

Figure 32. Schematic of water flows in CLIRUN-II

2.3.2 Calibration

CLIRUN-II simulates natural runoff with six calibration parameters, and requires natural runoff data to calibrate the model over an historic period. We calibrated the model by minimizing the squared differences between the 12 monthly GRDC runoff values and the 12 monthly averaged CLIRUN-II model outputs from 1971-1980 simulation period; this period was selected to best represent the source period for the 12 months of GRDC runoff data. Note that there are several limitations of using the GRDC dataset for calibration: (1) the dataset provides monthly averages and so yields no information on extremes; (2) gridded data in dry areas with no gauged data are unreliable; (3) the period of station data availability varies, so there may be temporal inconsistencies in the gridded data. As a result of these issues, calibration performance is closely tied to availability of runoff data (Figure 33). The choice of CLIRUN-II also introduces uncertainties; prior research suggests that there can be large differences

between the results of alternative hydrological models (e.g., Haddeland, et al 2011, Schewe et al. 2013).

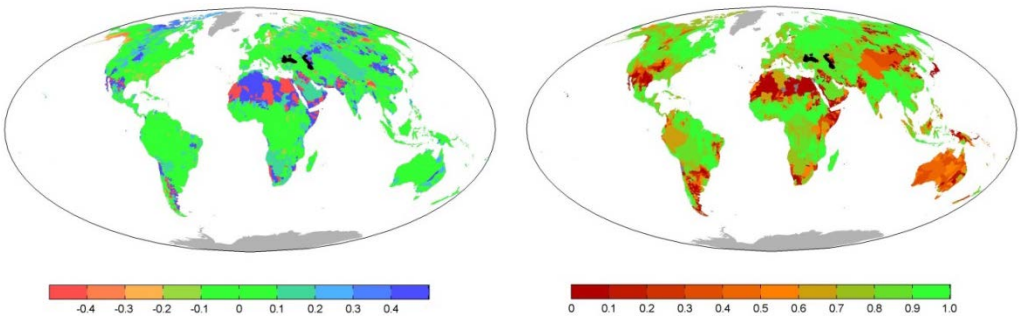

Figure 33. CLIRUN-II Calibration Error [(Modeled - Observed)/Observed] on left, and R^2 value on right. In error map, blue indicates that modeled outputs overestimate runoff and reds indicate underestimates.

2.4 Irrigation water requirements modeling

Globally, irrigation water requirements (IWRs) are the largest consumptive use of water, and will be significantly affected by projected rising temperatures and changing and more variable precipitation. Because of the strong dependence of crop water demands on temperature, climate change will have a more one-directional (i.e., increasing) effect on crop water demand than on runoff. As a result, we would expect the IWR signal to emerge from the noise sooner in this variable than in water resource variables that depend more directly upon precipitation.

Although detailed crop modeling of irrigation water requirements was far beyond the scope of this work, simplified methods can provide an understanding of water requirements suitable for a global scale analysis. One such method was developed by FAO (Allen et al. 1998) and is employed by IFPRI (Rosegrant et al. 2002), and relies on data that are available at a global scale. These include soil characteristics, temperature, precipitation, and PET, as well as crop-specific information including planting and harvest dates and seasonal timing of water demands. Soil characteristics and timing of

water demands are available through FAO (Allen et al. 1998), and a database of global planting and harvest dates for a wide range of crops are available through the University of Wisconsin (Sacks et al. 2010).

In this analysis, we focus on maize and wheat, which are the largest two global crops by growing area, and cover 18% and 15% of total global cropland in 2013, respectively (FAO 2014). Crop coverage data is provided through the Harvested Areas and Yields dataset, which is available at a 1/12 x 1/12 degree resolution through University of Wisconsin (described in Monfreda et al. 2008; Figure 34). Because this crop coverage dataset also includes rainfed areas, the area was reduced to irrigated regions only using an FAO dataset of global gridded irrigation data (described in von Velthuizen et al. 2007; Figure 35). These data are spatially aggregated up to the level of the 8,951 global river basins.

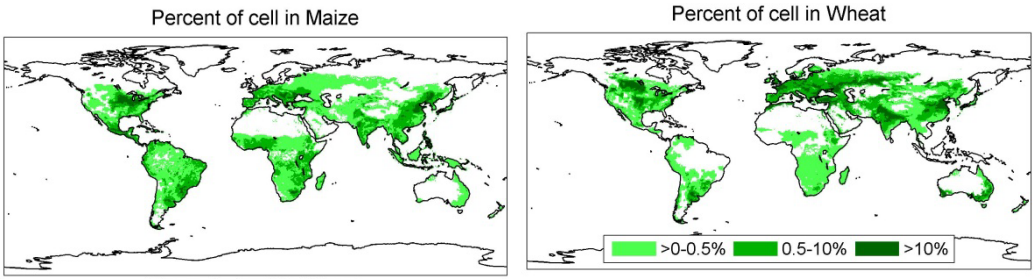

Figure 34. Global distribution of maize and wheat growing areas, where the shading of a cell represents the spatial fraction of that cell used to grow either maize or wheat (University of Wisconsin; Monfreda et al. 2008)

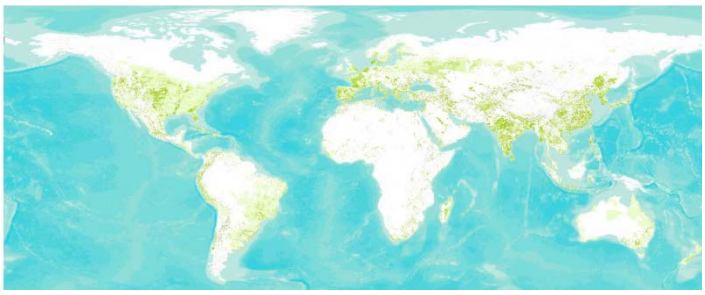

Figure 35. Distribution of irrigated areas globally, where darker shades of green represent a greater percentage of the grid cell area being used for irrigation (FAO; von Velthuizen et al. 2007)

Overall, the process outlined by FAO (1998) and IFPRI (2002) for calculating IWR for a particular crop involves first estimating total monthly crop water demand (crop evapotranspiration, or ETc), then estimating available monthly supply (effective precipitation, or Pe), and then calculating IWR each month as the difference between these values (ETc – Pe). The first step is to calculate ETc, which requires information on monthly reference evapotranspiration (ETo, which is equivalent to PET and calculated using Modified Hargreaves), the months when the crop demands water, and the crop water use coefficient for each month (Kc values). FAO (1998) provides crop water use coefficients by season, which we convert to monthly values by interpolation. Crop water demand is calculated for each month as the crop water use coefficient multiplied by reference evapotranspiration. Mean annual baseline ETo, and the corresponding annual average ETc values for maize and wheat are provided in Figure 36.

$$ETc_{c,m} = Kc_{c,m} * ETo_m$$

Where:

$ETc_{c,m}$ = Monthly crop evapotranspiration

$Kc_{c,m}$ = Monthly crop water use coefficient

ETo_m = Monthly reference evapotranspiration

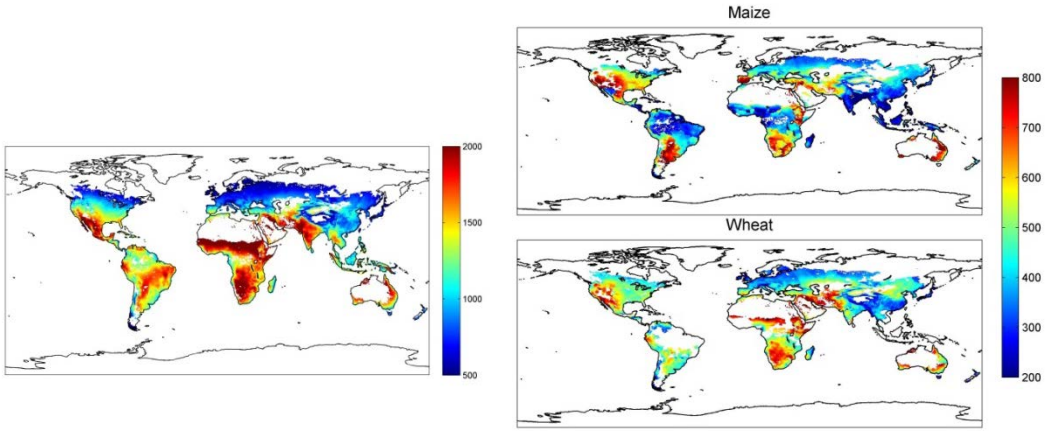

Figure 36. Annual ETo in irrigated maize and wheat growing areas (left) and corresponding ETc for each crop (right), both in mm

Next, we estimated effective precipitation as a function of soil water holding capacity, precipitation, and crop evapotranspiration. Soil water holding capacity is based on depth of irrigation, which is a soil property provided within FAO's universal soil database. Effective precipitation is capped at crop evapotranspiration, because any water over crop water demands is no longer usable as supply.

$$Pe_m = \max\left[f(DI) * \left(1.253{P_m}^{0.824} - 2.935\right) * 10^{0.001*ETc_m}, ETc_m\right]$$

Where:

Pe_m = Monthly effective precipitation

$f(DI)$ = A function of depth of irrigation that varies in form depending on the value of DI

P_m= Monthly precipitation

Lastly, we estimate the IWR as the difference between monthly crop water demand, and monthly effective precipitation. This value is specific to each crop, month, and river basin.

$$IWR_{c,m} = ETc_{c,m} - Pe_m$$

2.5 Basin yield and adaptation cost estimation

Basin yield is a measure of the annually reliable water supply from a basin. Much of the water available in a basin during a given year is lost if not stored, so storage in a basin can greatly increase its reliable supply, or yield. Basin yield is a useful broad indicator of the climate risk to basin-level water resources because it indicates a basin's ability to absorb the impact of changes in both the mean and variability of flows under climate change. Furthermore, achieving specified basin yield targets has economic implications, allowing the costs of adapting to climate changes to be estimated.

The storage yield curve has been developed as a means of relating basin yield to basin storage, and is a measure of the volume of storage needed to achieve a given level of reliable yield. Figure 37 provides an example of a storage yield curve for the Nile River at Aswan, and illustrates the information that the curve provides. The maximum yield on the curve indicates the mean annual runoff in the basin, and the minimum yield indicates the lowest flow in the runoff time series for the basin in question. That is, in a basin with no storage, the basin yield is assumed to be the lowest recorded annual flow. The shape of the storage yield curve is determined by the historical variability of basin runoff, where a steeper curve indicates a more stable system and a flatter curve a more

variable one. So all else equal, a basin with more variable flows would require more storage to achieve the same level of reliable yield.

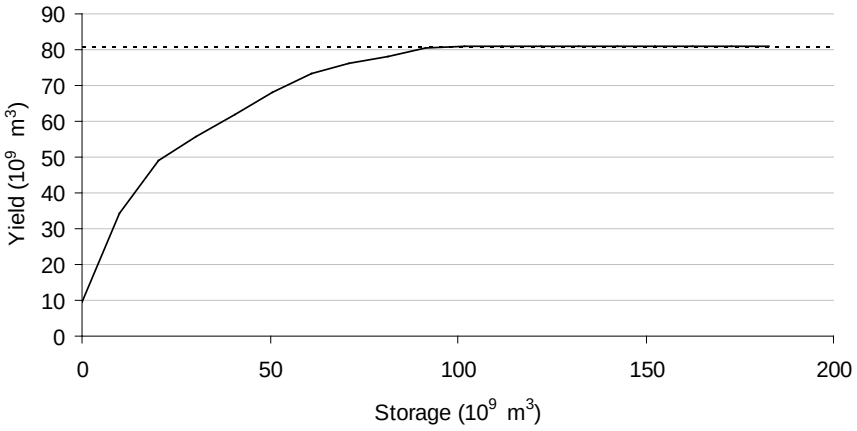

Figure 37. Storage yield curve for the Nile River at Aswan (from Strzepek et al. 2011)

2.5.1 Model formulation

To construct storage yield curves, we use the sequent peak algorithm (Wiberg and Strzepek 2005, modified from Thomas and Fiering 1962), which is an iterative procedure that identifies the minimum storage volume needed to generate various levels of reliable yield, given a basin inflow time series. In this formulation, the elements include reservoir storage, releases, evaporation and precipitation over the reservoir, and inflow, all at a monthly time step. The current analysis does not include evaporation and precipitation on the surfaces of reservoirs.

$$S_t = \{R_t + E_{t-1} - P_{t-1} - Q_t + S_{t-1} \mid if\ positive; otherwise,\ S_t = 0\}$$

Where:

S_t = reservoir storage requirement

R_t = releases

E_t = evaporation above the reservoir

P_t = precipitation above the reservoir

Q_t = inflow

t = time step

2.5.2 Generating storage yield curves under climate scenarios

Climate change can affect all elements of a storage yield curve, including the maximum yield given by the mean annual runoff, the minimum yield assumed to be the lowest annual flow, and the curve's slope given by variability of inflows. As a result, basin yield serves as an integrator of the various potential effects of climate change over both time and space. Importantly, climate change will also affect yield reliability, which can be estimated using long-established synthetic flow time series generation techniques that provide confidence intervals on the storage yield curve (Thomas and Fiering 1962).

There are two perspectives on the effect of climate change on storage yield. The first focuses on the impacts of climate change, and evaluates the change in yield given a fixed basin storage. The second view focuses on adaptation, and considers the change in storage that would be required to maintain a fixed yield (Figure 38). If yields increase, climate change may provide economic benefits, but only if the basin has existing water deficits or growing demands, or if markets exist to trade water between basins.

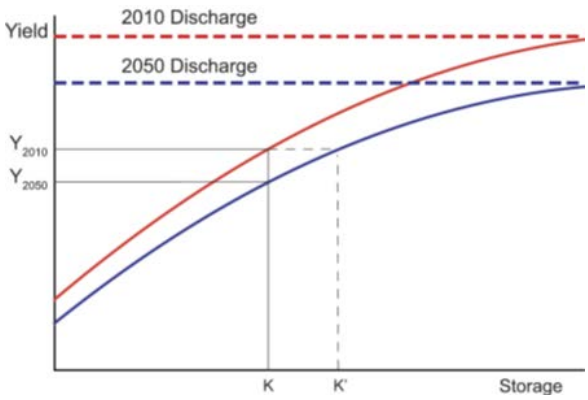

Figure 38. Example of storage yield curve (source: Ward et al. 2010). In a basin where runoff is projected to decline between 2010 and 2050, the yield provided by existing storage K falls from Y_{2010} to Y_{2050}. To maintain yield levels of Y_{2010}, storage would need to increase from K to K'.

Using baseline and projected annual runoff and low-flow values, baseline and projected storage yield curves were created for each of the 126 basins (described above), for the baseline and across all 220 GCM runs, and for both the 2050 and 2090 eras. Combined with information on reservoir storage in each of the 126 basins from IFPRI (Figure 39 on left), these storage yield curves provide information on changes in basin yields given existing basin storage, and in cases where yields fall, the required increases in storage needed to maintain existing yields. Note that in basins where the ratio of existing basin storage to current mean annual runoff exceeds one (Figure 39 on right), then additional storage will provide no additional basin yields. In these basins, falling yields would require alternative, non-storage adaptation options.

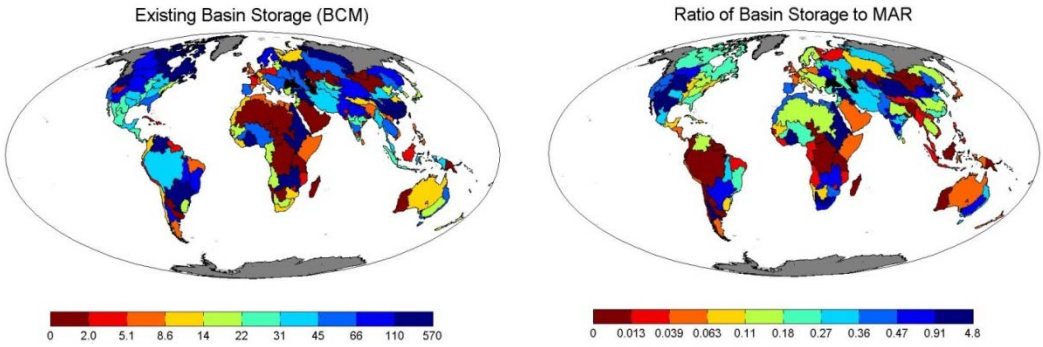

Figure 39. Existing basin storage in billions of cubic meters (BCM; left) and the ratio of basin storage to mean annual runoff (MAR; right). The ratio of basin storage to MAR is the number of years of average flow that can be stored in the basin's reservoirs.

2.5.3 Costs of maintaining current basin yield

We next estimate the economic impacts of maintaining current basin yields under a changing climate. In basins where yields decline, the adaptation response is assumed to be storage or a backstop of $1 per cubic meter of lost yield, whichever is less expensive. Costs of storage are taken from International Water Management Institute (Keller et al. 2000) and Wiberg and Strzepek (2005), who estimate volume-cost relationships for reservoirs. The backstop price per cubic meter is adopted from Ward et al. (2010), who take a similar approach to the current study in estimating the global costs of maintaining basin yields under climate change.

On the other hand, if basin yield increases under climate change, we assume that surplus water provides economic benefits only if the basin is water stressed. Operationally, if water is below a relative water stress of one on UNESCO's World Water Assessment Program (2006) water stress index, that basin is assigned no value for surplus water. Other basins are assigned values between $0/m3 and $1/m3, scaled to the basin's water stress index level (Figure 40).

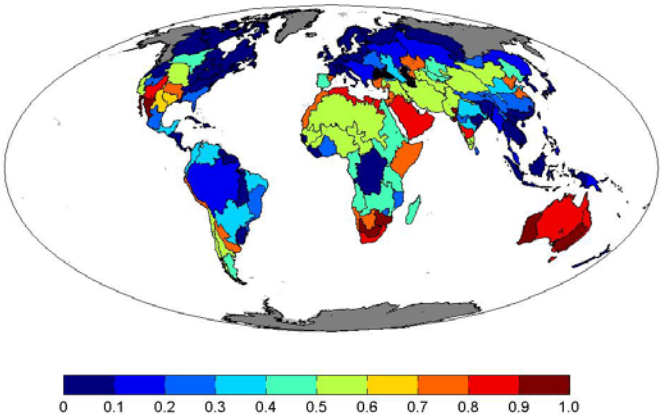

Figure 40: Assumed value (in US$) per cubic meter of surplus yield. Basins shaded in yellow and red (higher water value) are water stressed, whereas basins shaded in blue (lower water value) have little water stress.

3 RESULTS

The overall purpose of this inquiry is to analyze spatial patterns of agreement and disagreement between water availability and demand metrics, and possible lessons for water management and climate change adaptation. As the methods of evaluating levels of agreement are varied, we first present the metrics and indicators that are the focus of this assessment. We next present levels of agreement among ensemble patterns of changes in precipitation, which is the primary driver of global patterns of runoff and thus storage yield. These precipitation patterns are then compared to patterns of agreement in runoff, which although similar, differ in important ways due to dependence on temperature and the fact that runoff integrates climate outcomes over space, as previously discussed. Next, we discuss the patterns of change in IWR, which depend much more heavily on temperature and therefore shows stronger patterns of agreement. Lastly, our focus turns to patterns of agreement across ensembles in basin yield and the economic implications of maintaining historical yields. As reservoir

systems integrate the effects of climate variability and change over both time and space, this provides a broader perspective on patterns of agreement among ensembles.

3.1 Metrics, ensembles, regions, and period of focus

The metrics, ensembles, basins, and period of time that are the focus of this section are:

- **Metrics.** As the study focuses on agreement among water resource indicators, there are many potential metrics that can be used for comparison purposes. In this study, we focus on magnitude of changes in trends, level of agreement across ensembles in terms of direction of change, and signal-to-noise ratio. The signal over noise ratio provides information on when a climate signal is statistically significant, and can provide valuable information on trigger points for adaptation. For example, the adaptation strategy of flexible design (De Neufville and Scholtes 2011) involves designing infrastructure systems so that future adjustments can be made once more information becomes available (e.g., building additional hydropower turbine bays, and then adding the turbine only if a wet future occurs).
- **Ensembles.** Although we have processed 220 model runs from 17 ensembles, as noted above, we focus our comparisons on three of the larger between-model ensembles (40-member CCSM3, 17-member ECHAM, and 10-member CSIRO RCP4.5), and three of the between-model ensembles (22-member A1B, 23-member RCP4.5, and 20-member RCP8.5). Including only the three largest between-model ensembles provides an adequate number of members for statistical comparisons; this set of between-model ensembles provides a linkage between the earlier SRES

A1B used for the CCSM3 and ECHAM between-model ensembles, and two separate emissions scenarios from the CMIP5 suite of models.

- **Basins.** Although maps are presented showing both the 8,951 and 126 river basins of the world, tabular results are presented for only a subset of the 126 basins (Figure 41). These were selected to illustrate a broad cross-section of results, and include one or more major river basins in North America, South America, Africa, the Middle East, Central Asia, and Eastern Asia.
- **Period.** While results were generated for both the 2050 and 2090 eras, we focus on the 2050 (i.e., 2040-2060) period, which includes results for the 40-run NCAR CCSM3 ensemble.

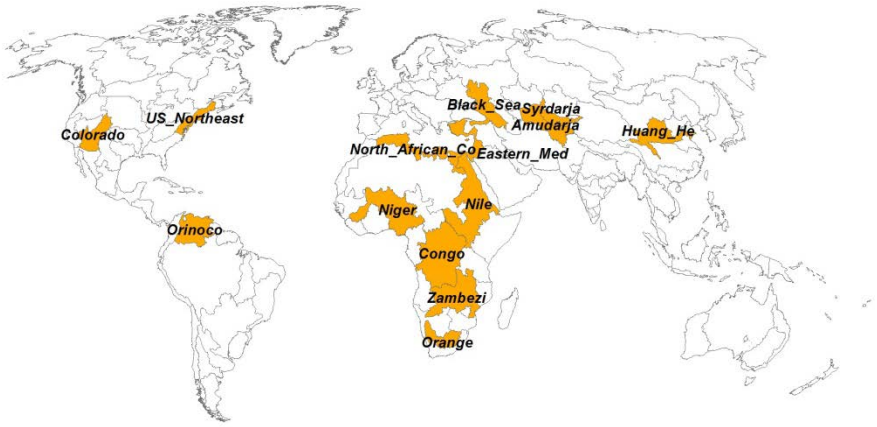

Figure 41. Selected river basins included in tabular results

3.2 Precipitation patterns

Although prior research using climate model ensembles has already examined agreement in patterns in precipitation across model ensembles (e.g., Deser et al. 2013), we provide similar analysis here for three reasons: (1) precipitation is the source of broad changes projected in other water resource variables to be discussed below, (2) patterns of change in precipitation serve as a point of comparison to the other water

resource variables, and (3) precipitation is global rather than confined to land surfaces, providing a more integrated picture of GCM patterns compared to runoff.

3.2.1 Agreement patterns over time and emissions scenarios

Our first observation is that agreement patterns in precipitation are spatially consistent over time and with emissions scenarios (Figure 42). As time progresses and emissions rise, the patterns of both agreement and persistent disagreement remain relatively fixed within models. With regard to emissions, note that the adjustment to the global mean precipitation trend means that the global trend in each decade is consistent across all ensembles. As such, the deeper colors on the RCP8.5 figure (right in Figure 42) reflect intensification of trends in both directions, rather than higher or lower precipitation globally.

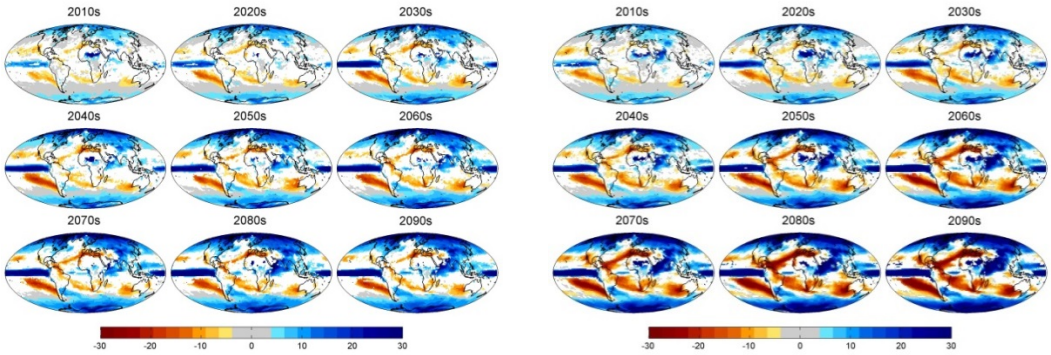

Figure 42: Percentage Change in annual precipitation, ensemble mean, 23 CMIP5 RCP4.5s (left) and 20 CMIP5 RCP8.5s (right), from the 1961-1990 baseline and the 2010s through the 2090s. In regions shaded white, fewer than 2/3 of the model members agree on the direction of change. Regions shaded gray have at least 2/3 of models agreeing on a small change of between -3% and +3%.

The regions of persistent disagreement suggest areas of the globe with meteorological patterns that are difficult to correctly model in the climate system, and the consistency of agreement patterns suggests emergent behavior characteristic of

each model. Not unexpectedly, within-model ensembles have a much greater degree of agreement in direction of change among members (Figure 43). The 40-member CCSM3 (40NCAR on Figure) and 17-member ECHAM ensembles both show very few regions on the globe where less than 2/3 of models agree on sign change.

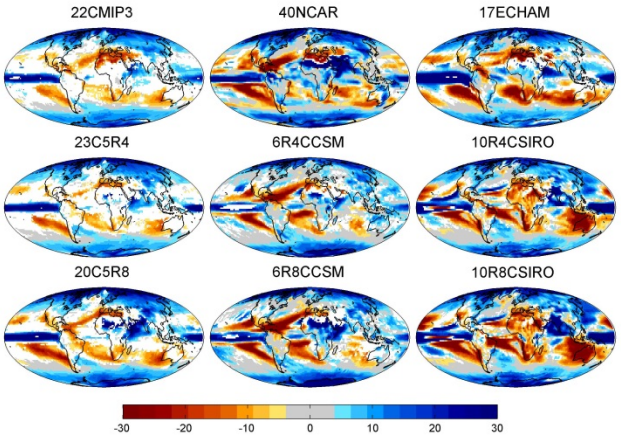

Figure 43: Percentage change in annual precipitation, ensemble mean for various between (left) and within-model ensembles between the 1961-1990 baseline and the 2050 era. In regions shaded white, fewer than 2/3 of the model members agree on the direction of change. Regions shaded gray have at least 2/3 of models agreeing on a small change of between -3% and +3%.

3.2.2 Agreement patterns across multiple within-model ensembles

In certain regions of the globe, agreement remains strong across two or more within-model ensembles, regardless of emissions scenario and CMIP series. We average the level of agreement, measured as the fraction of ensemble members that show a positive change in sign, across two ensembles of like emissions scenarios (Figure 44), and across five ensembles from available sets of CMIP5 RCP4.5 groupings, which is the broadest set of within-model ensembles available in our dataset (Figure 45). Interestingly, we agreement patterns in certain regions are maintained even when patterns from five within-model ensembles are combined. Consistently drying regions

include northern Africa and southern Europe, southern Africa, the southwestern US and central America, and Indonesia. Wetting regions include southeastern South America, the northeastern US, eastern and southern Asia, and the northern latitudes. This suggests less model uncertainty exists in these regions than in areas such as Indonesia or central Africa. Notably, most of the land surface areas of large disagreement appear to be in mid-continental (i.e., rather than coastal) regions of North and South America, Australia, Africa, and the Middle East.

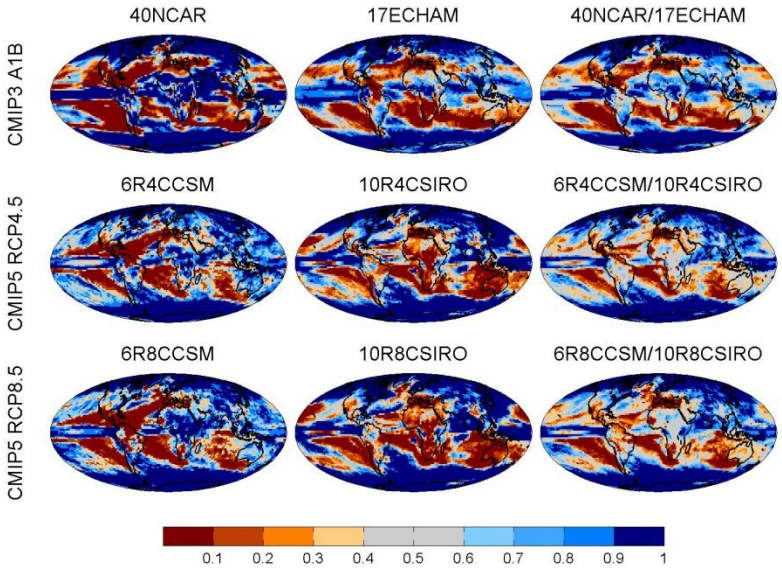

Figure 44. Fraction of ensemble members agreeing on a positive direction of change, where blue indicates agreement on wetting, and red indicates agreement on drying. Change is assessed between the 1961-1990 and 2050 era. Each rightmost map is the numerical average of its two counterparts to the left, and each row includes a set of like-emissions scenario within-model ensembles.

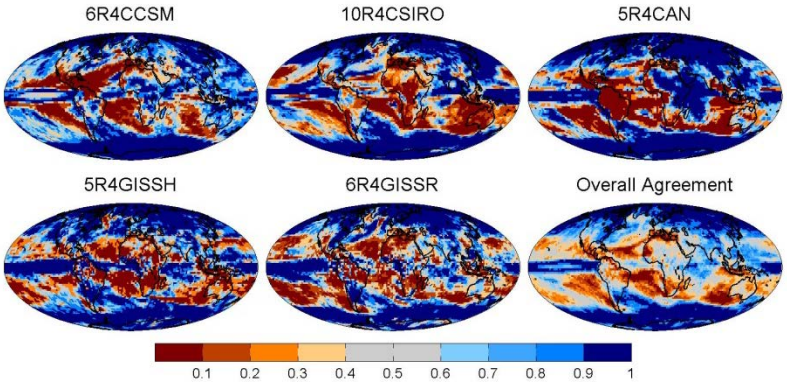

Figure 45. Fraction of ensemble members agreeing on a positive direction of change for five within-model CMIP5 RCP4.5 ensembles. Change is assessed between the 1961-1990 and 2050 era. The bottom-right map is the numerical average of the other five ensembles.

3.2.3 Patterns of agreement in latitude bands

Lastly, patterns of changes in precipitation by latitude band are very similar across both inter- and intra-GCM ensembles, suggesting that the bulk of disagreement between models results from longitudinal differences (Figure 46).

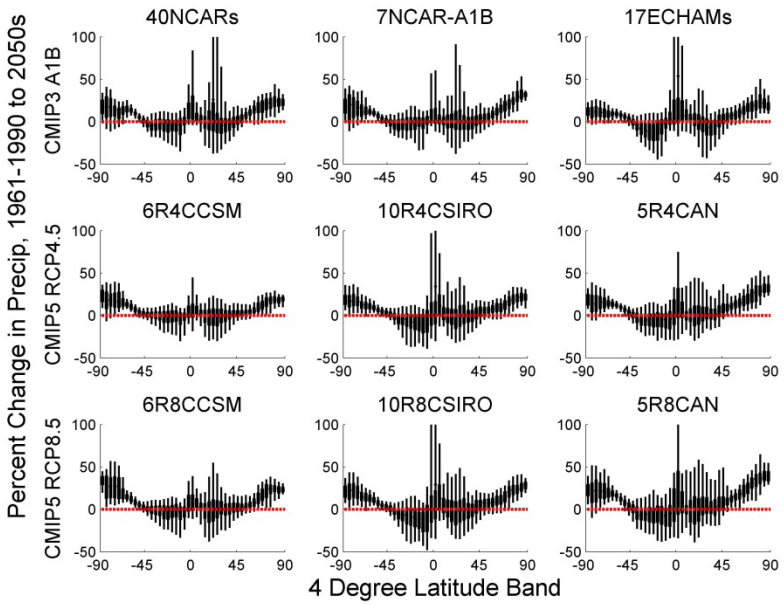

Figure 46. Percent change in precipitation between the 1961-1990 baseline and the 2050 era. Boxplots are averages of all grid cells in 4-degree latitude bands of the earth, with the boxplots being defined based on ensemble members.

3.3 Runoff

One of the central arguments of this work is that because water resource variables are dependent on temperature and integrate changes in precipitation over space, water resources variables will tend to have greater levels of agreement between model ensembles. We next compare the level of agreement between modeled runoff at the 8951 and 126 basins to levels of agreement within precipitation at the same scales, and then evaluate signal over noise ratios to identify emergence.

3.3.1 Agreement across ensembles relative to precipitation

Generally, runoff exhibits a spatially similar pattern of agreement to precipitation, but with a more robust agreement on drying in certain regions such as Australia and South America (Figure 47). As these levels of agreement are difficult to identify clearly visually in maps, Table 2 compares the percentage of sign agreement within selected river basins, with wetting or drying agreement of greater than 80% shaded in blue and red. Certain basins, such as the northeastern US or Eastern Mediterranean, have universal agreement across models, suggesting robust trends in those regions. The most apparent difference between precipitation and runoff, however, is a general increase in drying, presumably due to the incorporation of temperature. In the Orinoco (South America), Niger (western Africa), Amudarja (Central Asia), and Huang He (China), this drying moves individual ensembles into the 80% drying category when precipitation is translated to runoff.

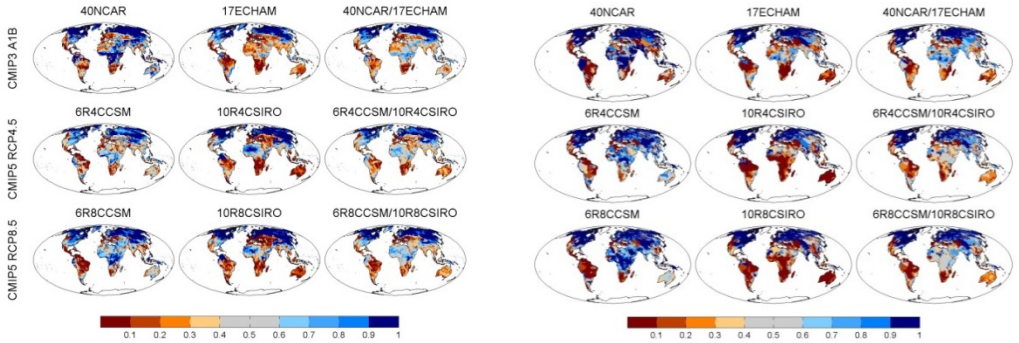

Figure 47. Fraction of ensemble members showing a positive change in annual precipitation (left) and runoff (right) from the 1961-1990 baseline to the 2050 era. Each rightmost map is the numerical average of its two counterparts to the left, and each row includes a set of like-emissions scenario within-model ensembles.

Table 2. Percent of ensemble members showing a positive change in precipitation and runoff, change from 1961-1990 to the 2050 era. Cells shaded in blue indicate that greater than 80% of members agree on a wetting trend, and in red means more than 80% agree on a drying trend. Basin locations provided in Figure 41 above.

Basin	Precipitation			Runoff		
	40NCARs	17ECHAMs	10R4CSIRO	40NCARs	17ECHAMs	10R4CSIRO
Zambezi	63%	**0%**	**0%**	28%	**0%**	**0%**
Congo	**100%**	**94%**	**0%**	**100%**	35%	**0%**
Niger	**100%**	**100%**	**10%**	**0%**	**100%**	**0%**
Nile	**100%**	**100%**	**0%**	**100%**	47%	**0%**
Orange	**100%**	**18%**	**0%**	**85%**	**6%**	**0%**
Orinoco	**100%**	35%	30%	**100%**	**6%**	20%
Amudarja	**5%**	35%	60%	**0%**	29%	**10%**
Syrdarja	**3%**	71%	60%	**3%**	59%	50%
Huang_He	**100%**	71%	70%	**100%**	**18%**	30%
Black_Sea	**3%**	47%	30%	**5%**	**94%**	70%
Colorado	**0%**	41%	**10%**	**0%**	35%	**10%**
US_Northeast	**100%**	**94%**	**100%**	**100%**	**100%**	**100%**
E. Mediterranean	**0%**	**0%**	**0%**	**0%**	**0%**	**0%**
N. African Ctries	**0%**	**0%**	**0%**	**0%**	**0%**	**0%**

3.3.2 Signal-to-noise in runoff versus precipitation

As noted above, the signal-to-noise (S-N) ratio can provide synthesized information about agreement on trends within an ensemble of model runs. We process the S-N ratio following the approach employed by Deser et al. (2012), who define the signal as the change from a baseline to a given 30-year projected period, and the noise as the interannual standard deviation over those 30 years. In our case, the signal would be the difference between the mean 1961-1990 value and the 2050 era value, and the noise would be the interannual standard deviation during the 2050 era. We then report the median S-N ratio over the ensemble for the 126 basins (Figure 48) and a selection of those basins (Table 3).

By 2050s, drying signal emerges in some arid regions during dry season (lowest 3 months of runoff for each basin; Figure 48 at right), and wetting annual signal emerges in others (Figure 48 at left) , although the effect is still modest. This absense of apparent emergence in runoff is similar to the findings of Mahlstein et al. (2012) for annual precipitation within the CMIP3 A1B ensemble.[4]

Annual 22 A1B results similar to spatial pattern of precipitation emergence documented in Mahlstein et al. (GRL 2012). As with agreement, comparing precipitation and runoff S-N ratios across individual basins suggests that runoff tends to move signals in a more drying direction due to the temperature effect (Table 3). Some emergent wetting S-N ratios in precipitation are only mild S-N ratios in runoff, and many of the drying ratios intensify.[5]

[4] Importantly, note that the high arid basin S-N ratios during the dry season is most likely attributable to near zero runoff during those months.

[5] Looking at monthly trends in precipitation and runoff over the 22 CMIP3 A1B runs reveals clearer signs of month-to-month emergence (i.e., whiskers do not overlap zero line) in runoff than precipitation (See Appendix B, Figure B-1).

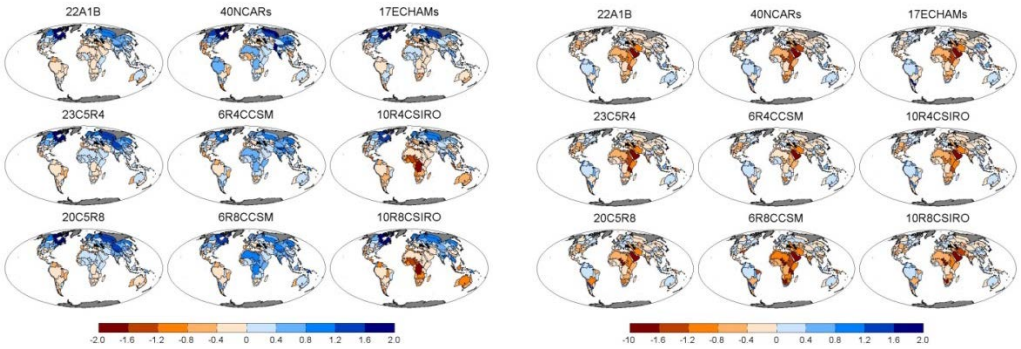

Figure 48. Median ensemble signal to noise ratio of annual runoff (left) and dry season runoff (right) across the 126 basins, from the 1961-1990 baseline to the 2050 era. The signal is the average change between periods, and the noise is the interannual standard deviation of the 2050 era. The dry season is defined as the lowest three months of the baseline period each year for each basin.

Table 3. Median ensemble ratio of signal to noise in annual precipitation and runoff for select basins, where the signal is the trend from the 1961-1990 period to the 2050 era, and the noise is the interannual standard deviation of the 2050 era. Cells in light blue show S/N ratios of 0.5 to 1, dark blue greater than 1, light red between -0.5 and -1, and dark red less than -1. Basin locations provided in Figure 41 above.

Basin	Annual Precipitation			Annual Runoff		
	40NCARs	17ECHAMs	10R4CSIRO	40NCARs	17ECHAMs	10R4CSIRO
Zambezi	0.03	-0.45	-0.64	-0.06	-0.39	-0.65
Congo	**1.16**	0.34	**-1.35**	0.44	-0.07	**-1.84**
Niger	0.44	0.41	-0.80	-0.46	0.38	**-1.51**
Nile	**1.32**	0.33	-0.59	0.41	-0.03	-0.33
Orange	0.56	-0.27	-0.51	0.09	-0.30	-0.55
Orinoco	0.77	-0.09	-0.29	0.47	-0.32	-0.66
Amudarja	-0.33	-0.12	0.08	-0.33	-0.32	-0.16
Syrdarja	-0.35	0.19	0.08	-0.29	0.06	-0.05
Huang_He	**1.48**	0.15	0.16	**1.39**	-0.34	-0.25
Black_Sea	-0.34	-0.01	-0.10	-0.24	0.35	0.14
Colorado	-0.69	-0.06	-0.33	-0.60	-0.09	-0.41
US_Northeast	0.84	0.53	0.59	0.65	0.69	0.67
E. Mediterranean	-0.72	-0.64	-0.51	-0.90	-0.69	-0.47
N. African Ctries	-0.26	-0.47	-0.48	-0.56	-0.82	-0.89

3.4 Irrigation water requirements

Irrigation water requirements are the largest global water use and therefore a strong indicator of the incremental effect that climate change will have on global water demand. Due to the strong dependence on temperature, regional trends are generally positive, and agreement across both between- and within-model ensembles is considerably broader than with precipitation or runoff (Figure 49). Because IWR is the difference between crop water demands and usable rainfall, the strongest positive trends are in regions where rainfall declines are largest. Although one would expect to see clear emergence of the signal from the noise based on these trends, the interannual variation of IWR is driven by both temperature and precipitation, and this "double noise" causes noise to generally overwhelm the signal, with some exceptions in the Middle Easter and western North America (Figure 50).

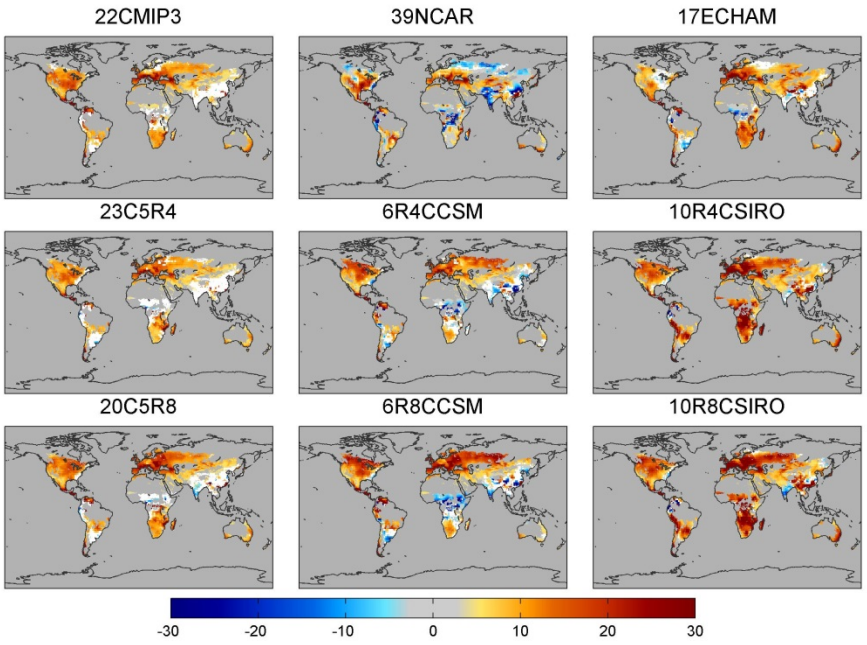

Figure 49. Percentage Change in irrigation water requirement, ensemble means, from the 1961-1990 baseline to the 2050 era. In regions shaded white, fewer than 2/3 of the model members agree on the direction of change. Regions shaded in lighter gray have at least 2/3 of models agreeing on a change of between -3% and +3%

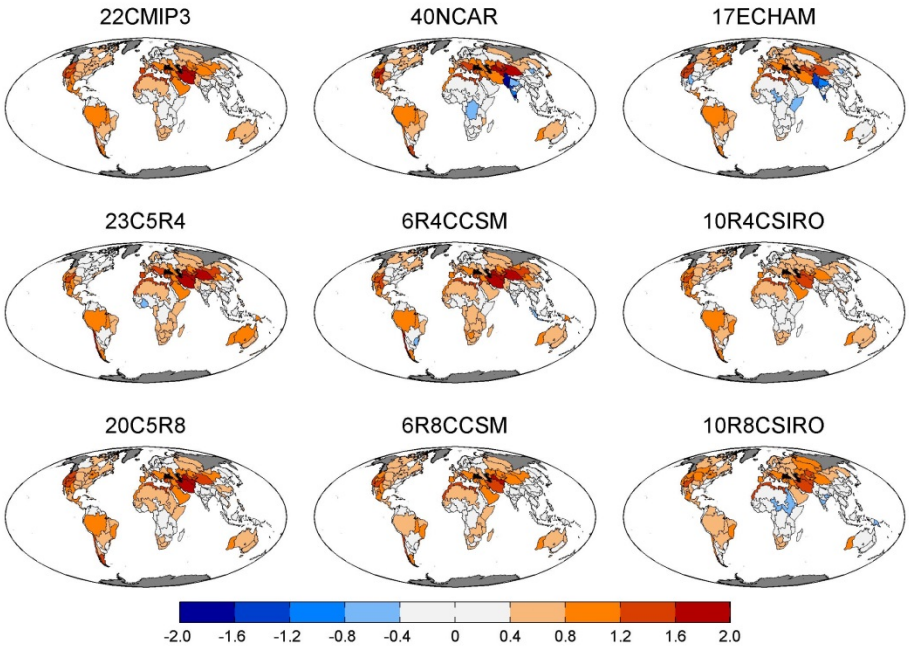

Figure 50. Median IWR signal-to-noise ratio across ensemble members for 9 ensembles. Signal is mean change between 1961-1990 and 2050 era, and signal is interannual standard deviation over yeras in the 2050 era.

3.5 Basin storage yield

Sustainable water supply yield from a river basin is influenced by mean annual runoff, interannual flow variability, and available storage infrastructure. As a result, changes in minimum flows, variability, or mean conditions under climate change would affect basin yield. Figure 51 provides an example—for a set of four between-model ensembles for the Zambezi basin in southern Africa—of how climate change can affect storage yield in a basin. Due to its shallow slope, small vertical changes in the storage yield curve can cause large increases in storage requirements to maintain fixed yields.

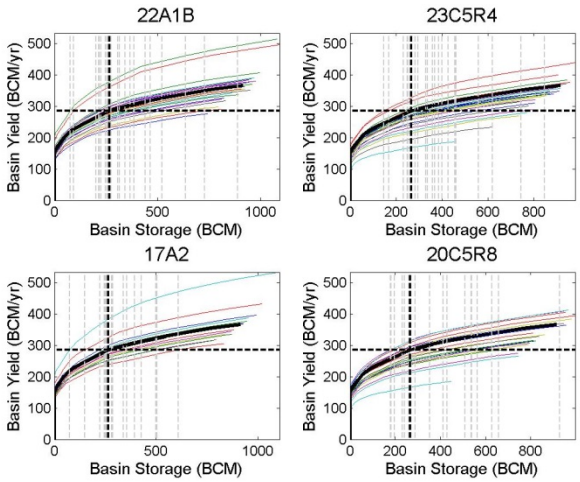

Figure 51. Storage yield curves for Zambezi Basin. Historical storage yield curve is a black solid line; existing storage is a vertical black dashed line; resulting current yields is a horizontal black dashed line; storage yield curves for each of the ensemble members are colored solid lines; and storage requirements to maintain current yields are vertical gray dashed lines.

As described in the methodologies and suggested in Figure 51, in basins with falling yields, either additional storage or another source (e.g., desalination) will be needed to meet demands. On the other hand, increases in storage yield may present opportunities for internal basin development or inter-basin transfers. Changes in yields closely mirror changes in runoff (Figure 52, on left), whereas resulting changes in storage requirements can be magnified considerably due to the nonlinear relationship between yield and storage (Figure 52, on right).

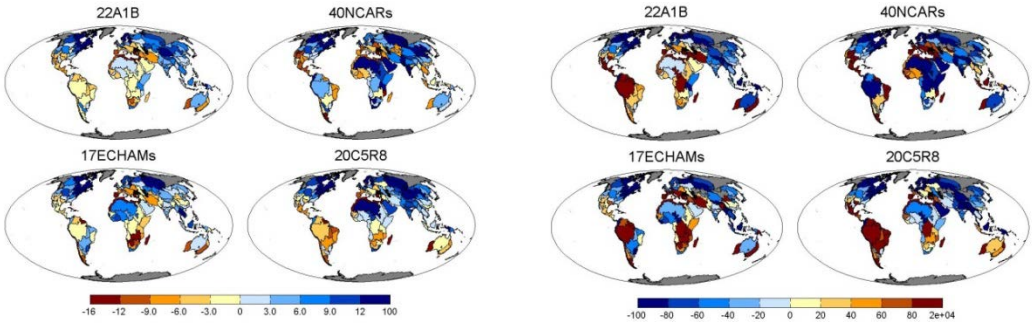

Figure 52. Median percent change in storage yield across select ensembles (left) and median percent change in storage to maintain constant yield across select ensembles (right)

As described in the methodologies section, we estimate the economic effects of climate change by assuming that basins with decreasing yields incur costs to maintain historical yields, and that basins with increasing yields gain those as economic benefits if water is scarce. Costs are the cheaper of either increased storage or a backstop of $1 per cubic meter of lost yield, whichever is less expensive. We find that the median annual global net costs of adapting to climate change in the 2050 era is $15 billion per year at a 5% discount rate (Figure 53); higher than Ward et al. (2010), although that study focused on developing countries only). However, the inter-quartile ranges of the CMIP3 and CMIP5 ensembles range from -$5 billion (RCP4.5) to +$40 billion (A2), and the range of intra-GCM ensembles is from -$25 billion (NCAR A1B) to +$80 billion (CSIRO RCP4.5). A lesson to draw from these findings is that it is critically important to consider a broad range of climate models when doing adaptation planning.[6]

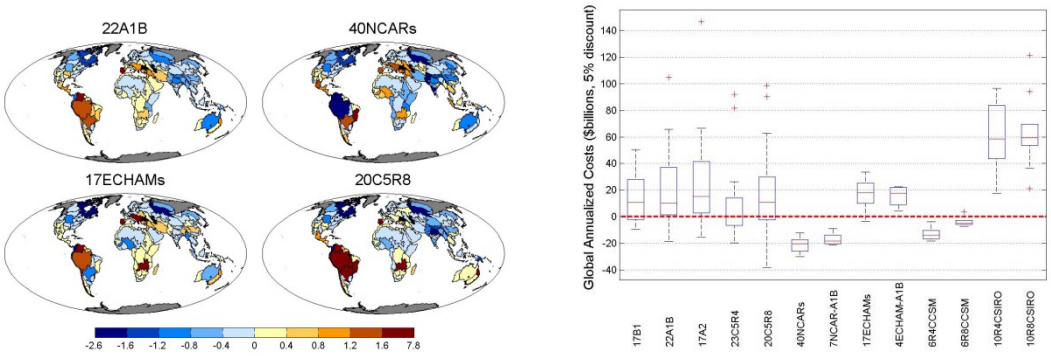

Figure 53. Median Cost Across Select Ensembles (billions of US$, discounted at 5%; left) and Boxplots of Global Costs for each Ensemble (billions of US$, discounted at 5%; right)

To evaluate agreement among adaptation cost estimates, it is more relevant to focus on geographic regions rather than the globe. As a result, we aggregate the basin-

[6] For example, the World Bank EACC study estimate costs of adaptation in developing countries to be ~$100 billion per year. The water supply component of this was approximately $9 billion/year. They used two scenarios—NCAR A2 run 1 (dry), and CSIRO A2 run 1 (wet)—to develop these estimates, which result in a range from -$20 billion to +$95 billion based on our run 1 of these two ensembles.

scale costs to the seven World Bank regions, which include East Asia and the Pacific (EAP), Europe and Central Asia (ECA), Latin America and the Caribbean (LAC), Middle East and North Africa (MENA), South Asia (SA), Sub-Saharan Africa (SSA), and all other countries (NB; see Figure 54). We see an unexpected degree of agreement in direction of economic outcome across these regions (Figure 55). In the MENA region, the interquartile range (IQR) of the three between-model and three within-model ensembles all show positive costs, and in the SSA region, only one ensemble has an IQR with economic benefits. There is a general agreement on benefits in SA, and on costs in LAC and other countries.

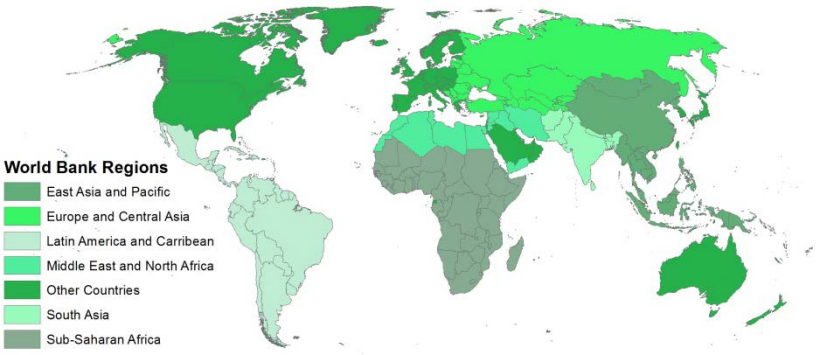

Figure 54: World Bank regions

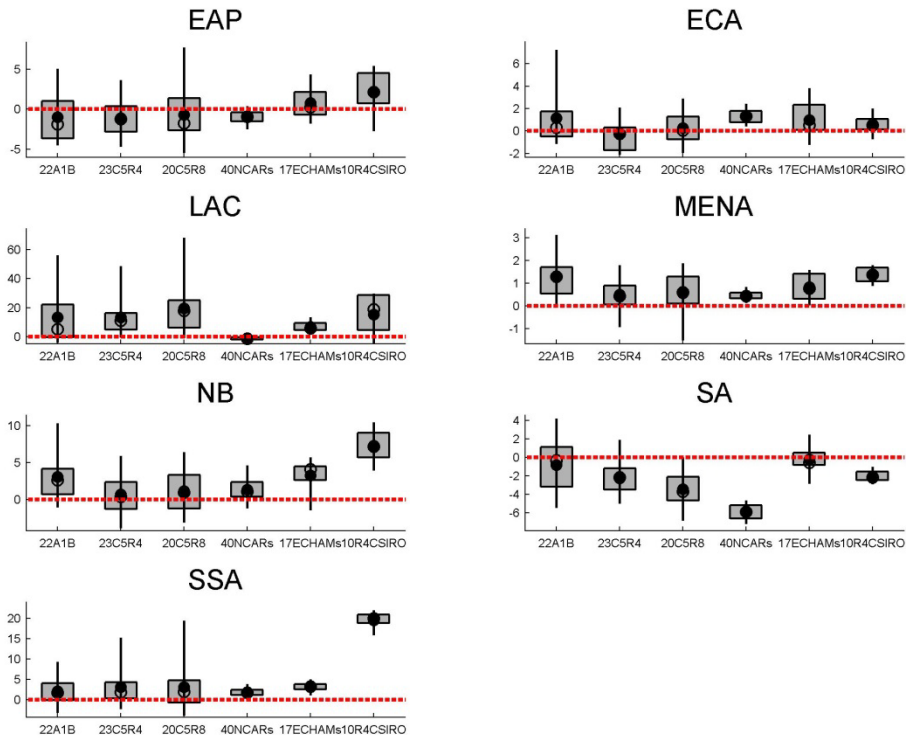

Figure 55. Distribution of adaptation net costs across three between-model and three within-model ensembles for each World Bank region. The dashed red line represents zero net costs, values above the line represent net costs, and below the line represent net benefits. The boxplots are composed of individual model runs within each ensemble.

The explanation for the much stronger sign agreement in economic effects has two sources: (1) economic outcomes map to runoff, which includes the stronger sign agreement in temperature; and (2) in many basins, changes in precipitation have a magnified effect on runoff. These dynamics can be seen for a selection of basins in Figure 56, which plots percentage changes in precipitation and runoff versus annual costs for all 220 model runs. That the fitted precipitation line is generally above the runoff line, indicating that zero changes in precipitation will still result in costs due to the temperature effect. In basins such as Zambezi, Volta, and Niger, the precipitation line is steeper than the runoff line, such that marginal changes in precipitation have a magnified effect on economic outcomes, leading to greater sign agreement.

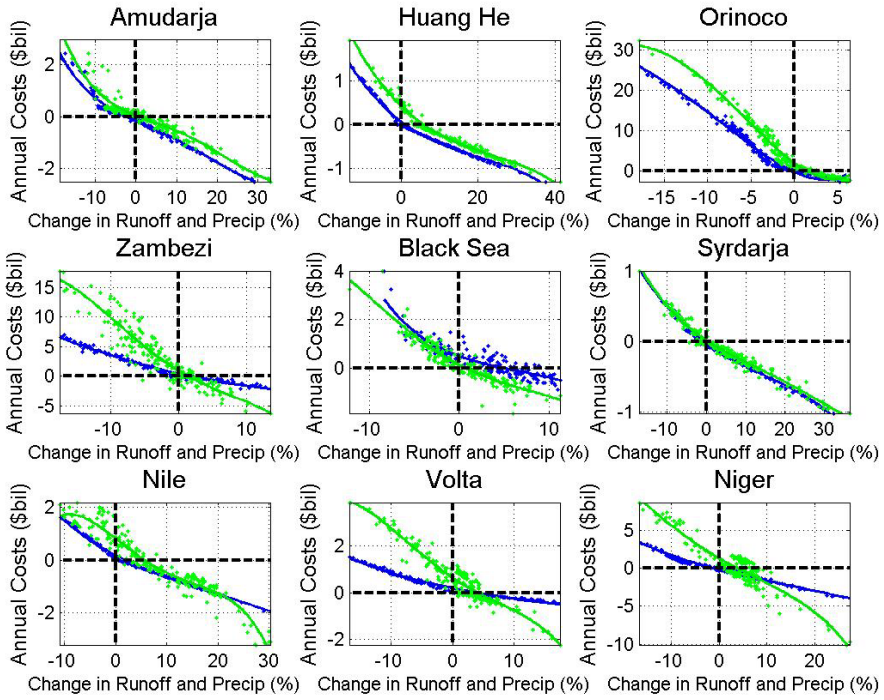

Figure 56. Percentage changes in runoff (blue) and precipitation (green) versus annual adaptation costs in selected basins, where each point represents one of the 220 model runs. The lines are a 4th order polynomial fit to the points.

4 CONCLUSIONS AND FURTHER RESEARCH

Climate change and rapidly rising global water demand are expected to place unprecedented pressures on already strained water resource systems. Successfully planning for these future changes requires a sound scientific understanding of the timing, location, and magnitude of climate change impacts on water needs and availability – not only average trends, but also interannual and decadal variability and associated uncertainties. Using a range of available within- and between-model ensembles, this research explores the spatial and temporal patterns of high confidence as well as uncertainty in projected river runoff, irrigation water requirements, and basin storage yield. A central hypothesis of our work is that by integrating precipitation effects over space and time, projections of these water resource variables will tend to

have higher levels of within- and cross-ensemble agreement than precipitation. Under each ensemble, we also develop cost estimates of adapting global water supply systems to maintain historically available supply.

4.1 Conclusions

We begin this research focusing on precipitation, which is the main climatic driver of uncertainty in projections of water demand and availability. Between-model patterns of high confidence and uncertainty in precipitation tend to remain fixed over time and with increased forcing (Figure 42), suggesting 'emergent behavior' in each model that has been observed in prior research (e.g., Strzepek and Schlosser 2010). Projected changes in precipitation and temperature drive modeled changes in river runoff, and we observe strong spatial patterns of multiple-ensemble agreement and disagreement in both precipitation and runoff trends (Figure 47). Regions with robust cross-ensemble drying trends include southern Europe, northern Africa, western Australia, southern Africa, eastern Brazil, and northern Mexico; and wetting trends occur in the northeastern US, Canada, northern regions of the globe, and parts of southeast Asia.

Basin yield has the advantage of integrating changes in both the mean and variability of projected runoff over time. We find that relative to changes in precipitation, patterns of changes in basin yield are both magnified and systematically drier due to the dependence of river runoff on land surface dynamics and temperature. Due to the temporally integrating effects of basin yield and monetary discounting, the costs of maintaining historical yields show still stronger patterns of agreement across GCM ensembles, particularly when focusing on agreement within broad geographic regions (Figure 55). If the robust patterns of projected increases in irrigation water

requirements (Figure 49) were incorporated into basin supply needs to more fully capture the effects of climate change, the model agreement observed here would be broader still. The fact that agreement exists across such a broad range of multiple-member GCM ensembles suggests a high degree of confidence in direction of change in water availability in these regions, and provides clearer signals for longer-term investment decisions in water infrastructure. By showing the variability in adaptation costs across within-model ensembles, this research also re-affirms the importance of considering a broad range of climate models when doing adaptation planning. Reliance on the outputs of too few models can generate highly misleading results, particular in regions where the direction and magnitude of projected changes are highly uncertain.

4.2 Further Research

There are several avenues for future research. First would be a more complete evaluation of the time of signal emergence in runoff, irrigation water demand, and changes in storage yield. The influence of temperature on the water system likely means earlier emergence times, which has implications for adaptation planning. Such a study would also incorporate new recently available large between-model ensembles. The current work focuses primarily on two ensembles: the 40-member NCAR and 17-member ECHAM sets. Broader conclusions about agreement between ensembles would be possible using new ensembles, such as the 30-member CESM1 (CAM5) set. Another important direction for future work is an assessment of whether select ensembles are more appropriate for certain regions based on their statistical performance relative to the observed climate of that region (i.e., the model skill of the ensemble). Although our screening assessment comparing basin yield outputs using modeled and observed

climate inputs suggests that model skill is generally too poor for clear region-ensemble couplings to emerge (Appendix B), a more thorough investigation is needed.

Future research would also integrate projected changes in irrigation water requirements and supplemental irrigation requirements stemming from changes in rainfed yields into an analysis of basin water supply under an uncertain future. Estimates of economic outcomes could be developed using ranges of population trajectories and environmental flow assumptions. In addition, the broader study would evaluate implications for basin yield and adaptation costs incorporating a wider set of projected global changes, including rising food demands, population increases, and environmental flow requirements. The present analysis evaluates how historical yields would be affected under climate change given current population and per capita water use. Because storage yield curves are non-linear, further research would evaluate the incremental impacts of climate change on future rather than current yield requirements.

Acknowledgements

We are grateful for technical contributions provided by Diane Ivy and Yohannes Gebretsadik.

REFERENCES

Alcamo, J. M. Flörke, and M. Märker. 2007. "Future Long-Term Changes in Global Water Resources Driven by Socio-economic and Climatic Changes," *Hydrological Sciences Journal* 52 (2): 247-275.

Allen R G, Pereira L S, Raes D and Smith M 1998 Crop evapotranspiration—guidelines for computing crop water requirements. *FAO Irrigation and drainage*, paper 56.

Arnell, N. W. 2004. "Climate Change and Global Water Resources: SRES Emissions and Socio-economic Scenarios." *Global Environmental Change* 14: 31–52.

Boehlert, B., E. Fitzgerald, J. Neumann, K. Strzepek, and J. Martinich. *In review*. The Effect Greenhouse Gas Mitigation on Drought Impacts in the U.S. *Weather, Climate, and Society.*

De Neufville, R. and S. Scholtes. 2011. *Flexibility in Engineering Design.* MIT Press: Cambridge, MA.

Deser, C., Phillips, A., Bourdette, V. & Teng, H. 2012. Uncertainty in climate change projections: the role of internal variability. *Clim. Dynam.* 38**,** 527–547.

Droogers P and Allen R 2002 Estimating Reference Evapotranspiration Under Inaccurate Data Conditions *Irrigation and Drainage Syst.* **16** 33-45

FAO (UN Food and Agriculture Organization). 2014. FAOSTAT: World Crop Areas for 2013. Accessed on September 1, 2014 from http://faostat.fao.org/.

Fekete B, Vorosmarty C and Grabs W 2002. High-resolution fields of global runoff combining observed river discharge and simulated water balances, *Global Biogeochemical Cycles*, **16:3** 15-1 to 15-10.

GRDC (Global Runoff Data Center) 2007 Major River Basins of the World (Koblenz: Federal Institute of Hydrology)

Gupta, V. J., and S. Sorooshian. 1983. "Uniqueness and Observability of Conceptual Rainfall–Runoff Model Parameters: The Percolation Process Examined." *Water Resources Research* **19** (1): 269-276.

Gupta, V. K., and S. Sorooshian. 1985. "Relationship between Data and the Precision of Parameter Estimates of Hydrologic Models." *Journal of Hydrology* (JHYDA7) **81** (1-2): 57-77.

Haddeland, Ingjerd, and Coauthors, 2011: Multimodel Estimate of the Global Terrestrial Water Balance: Setup and First Results. J. Hydrometeorology, **12**, 869–884.

Hamlet, A., P. Carrasco, J. Deems, M. Elsner, T. Kamstra, C. Lee, S. Lee, G. Mauger, E. Salathe, I. Tohver, and L. Binder. 2010. *Final Report for the Columbia Basin Climate Change Scenarios Project*. University of Washington: Climate Change Impact Group. December.

Hawkins, E. and R. Sutton, 2009: The Potential to Narrow Uncertainty in Regional Climate Predictions. *Bull. Amer. Meteor. Soc.*, 90, 1095–1107. doi: http://dx.doi.org/10.1175/2009BAMS2607.1

Hawkins, E. and Sutton, R. 2012. Time of emergence of climate signals. *Geophys. Res.*

Lett. **39,** L01702.

Huber-Lee, A, D. Yates, D. Purkey, W. Yu, C. Young, and B. Runkie. 2005. "How Can We Sustain Agriculture and Ecosystems? The Sacramento Basin (California, USA)," *Climate Change in Contrasting River Basins: Adaptation Strategies for Water, Food and Environment*, ed. J. C. J. H. Aerts, P. Droogers. CABI Publishing.

IPCC. 2013. *Climate Change 2013: The Physical Science Basis. Contributions of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Chnge* [Stocker TF, D. Qin, GK Plattner, M. Tignor, SK Allen, J Boschung, A. Nauels, Y Xia, V Bex and PM Midgley, (eds)]. Cambridge University Press, Cambridge United Kingdom and New York, NY, USA, 1535 pp.

Keller, A, R. Sakthivadivel,; D Seckler. 2000. Water scarcity and the role of storage in development. Colombo Sri Lanka: International Water Management Institute (IWMI), vii, 20p. (Research report 39).

Kaczmarek, Z. 1993. "Water Balance Model for Climate Impact Analysis." *Acta Geophysical Polonica* 41 (4): 423-437.

Kaczmarek, Z. 1998. *Human Impact on Yellow River Water Management Interim Report IR-98-016*. Austria: International Institute for Applied Systems Analysis.

Konzmann, M., Gerten, D., Heinke, J. 2013: Climate impacts on global irrigation requirements under 19 GCMs, simulated with a vegetation and hydrology model. *Hydrol. Sci. J.* **58**, 1–18.

Mahlstein, I., R. W. Portmann, J. S. Daniel, S. Solomon, and R. Knutti. 2012. Perceptible changes in regional precipitation in a future climate, *Geophys. Res. Lett.*, **39**, L05701, doi:10.1029/2011GL050738.

McMahon, T.A., R.M. Vogel, M.C. Peel, G.G.S. Pegram, 2007: Global streamflows – Part 1: characteristics of annual streamflows*. Journal of Hydrology*, **347**(3–4): 243–259.

Monfreda et al. 2008. Farming the planet: 2. Geographic distribution of crop areas, yields, physiological types, and net primary production in the year 2000, *Global Biogeochemical Cycles,* **22**, GB1022, doi:10.1029/2007GB002947

Milly, P., K. Dunne, and A. Vecchia. 2005. Global Pattern of Trends in Streamflow and Water Availability in a Changing Climate." *Nature* 347-350.

Pike, J. 1964. The estimation of annual runoff from meterological data in a tropical climate. *J Hydrol* 2:116-123.

Rosegrant, RW, X Cai, and SA Cline. 2002. World Water and Food to 2025: Dealing with Scarcity. International Food Policy Research Institute: Washington DC.

Sacks, W.J., D. Deryng, J.A. Foley, and N. Ramankutty (2010). Crop planting dates: an analysis of global patterns. *Global Ecology and Biogeography* **19**, 607-620. DOI: 10.1111/j.1466-8238.2010.00551.x

Schewe, J., Heinke, J., Gerten, D., Haddeland, I., Arnell, N.W., Clark, D.B., Dankers, R., Eisner, S., Fekete, B., Colón-González, F.J., Gosling, S.,N., Kim, H., Liu, X., Masaki, Y, Portmann, F.T., Satoh, Y., Stacke, T., Tang, Q., Wada, Y., Wisser, D., Albrecht, T., Frieler, K., Piontek, F., Warszawski, L., Kabat, P. 2014. Multi-model assessment of

water scarcity under climate change. *Proceedings of the National Academy of Sciences.* **111**(9): 3245-3250.

Solomon, S., G-K. Plattner, R. Knutti, and P. Friedlingstein. 2009. Irreversible climate change due to carbon dioxide emissions. *Proc. Natl. Acad. Sci*. doi:10.1073/pnas.0812721106

Strzepek, K., A. McCluskey, J. Hoogeveen, and J. van Dam. 2005. "Food Demand and Production: A Global and Regional Perspective". *Climate Change in Contrasting River Basins: Adaptation Strategies for Water, Food and Environment*. ed. J. C. J. H. Aerts, P. Droogers. CABI Publishing, 2005

Strzepek, K. M., and C. W. Fant IV. 2010. *Water and Climate Change: Modeling the Impact of Climate Change on Hydrology and Water Availability*. University of Colorado and Massachusetts Institute of Technology.

Strzepek, K. and C.A. Schlosser. 2010. Climate change scenarios and climate data, *World Bank Economics of Adaptation to Climate Change*, Discussion paper number 9, (Washington DC: The World Bank)

Strzepek, K. A McCluskey, B Boehlert, M Jacobsen, C Fant IV. 2011. Climate Variability and Change: A Basin Scale Indicator Approach to Understanding the Risk to Water Resources Development and Management. World Bank, Washington, DC

Strzepek K, Yohe G, Neumann J and Boehlert B 2010 Characterizing changes in drought risk for the United States from climate change *Environ. Res. Lett.* **5** 044012

Strzepek K, M Jacobsen, B Boehlert, J Neumann. 2013. Toward evaluating the effect of climate change on investments in the water resources sector: insights from the forecast and analysis of hydrological indicators in developing countries. *Environmental Research Letters* **8** (4), 044014

Strzepek, K., J. Neumann, J. Smith, J. .Martinich, B. Boehlert, M. Hejazi, J. Henderson, C. Wobus, R. Jones, K. Calvin, D. Johnson, E. Monier, J. Strzepek, H. Yoon. *In review.* Benefits of Greenhouse Gas Mitigation on the Supply, Management, and Use of Water Resources in the United States. *Climatic Change.*

Sutton, WR, JP Srivastava, and JE Neumann. 2013. Looking Beyond the Horizon: How Climate Change Impacts and Adaptation Responses Will Reshape Agriculture in Eastern Europe and Central Asia. Directions in Development. Washington, DC: World Bank. doi:10.1596/978-0-8213-9768-8.

Thomas, HA and MB Fiering. 1962. Mathematical synthesis of streamflow sequences for the analysis of river basins by simulation. In: A. Maass, M.M. Hufschmidt, R. Dorfman, H.A. Thomas Jr., S.A. Marglin and G.M. Fair (Editors), Design of Water Resource Systems. Harvard University Press, Cambridge, Mass.

Turc. 1954. Water balance of soils: relationship between precipitation, evapotranspiration and runoff. *Ann Agron.* 5:49-595 and 6:5-131

Von Velthuizen et al. 2007. Mapping biophysical factors that influence agricultural production and rural vulnerability. FAO and IIASA.

Vörösmarty, C. J., P. Green, J. Salisbury, and R. B. Lammers, 2000. Global water resources: Vulnerability from climate change and population growth. *Science*, **289**, 284–288.

Ward, P., K. Strzepek, W. Pauw, L. Brander, G. Hughes, J. Aerts. 2010. Partial costs of global climate change adaptation for the supply of raw industrial and municipal water: a methodology and application. *Environmental Research Letters.* **5**(4).

Wiberg D and KM Strzepek. 2005.Development of Regional Economic Supply Curves for Surface Water Resources and Climate Change Assessments: A Case Study of China *IIASA Research Report* RR-05-001. November.

World Water Assessment Programme, 2006, UN World Water Development Report 2: Water: A Shared Responsibility; Paris, UNESCO and New York, Berghahn Books, p. 116

Yates D. 1996. "WatBal: An Integrated Water Balance Model for Climate Impact Assessment of River Basin Runoff." *International Journal of Water Resources Development* **12** (2): 121-139.

APPENDIX A: SUPPLEMENTAL GRAPHICS REFERENCED IN MAIN BODY

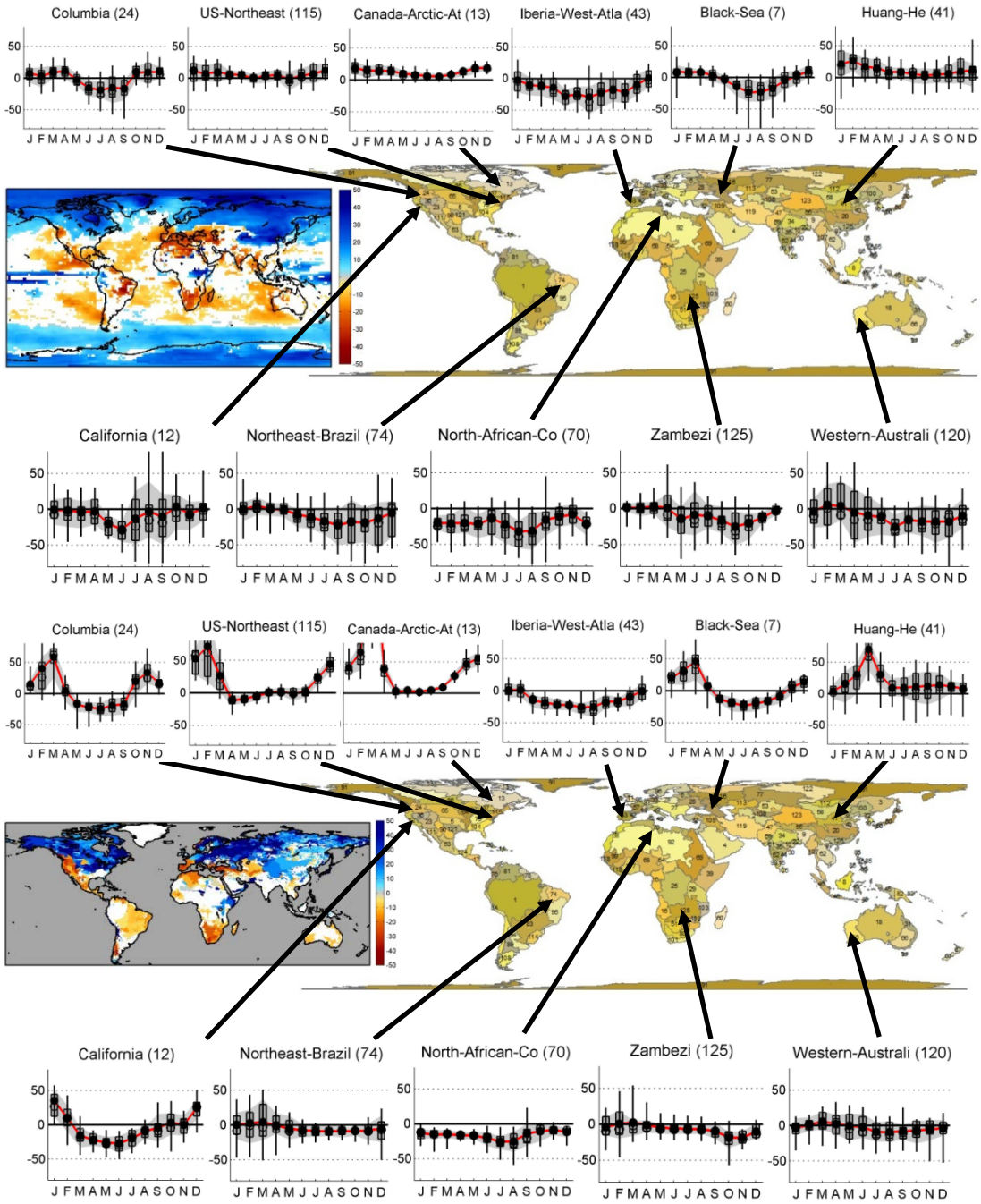

Figure A-1. Monthly patterns of percent changes in precipitation (upper graphic) and runoff (lower graphic) to the 2050s across the 22 A1B runs. Runoff tends to be both less variables and show clearer signs of month-to-month emergence (i.e., whiskers do not overlap zero line) than precipitation.

APPENDIX B. SCREENING ASSESSMENT OF MODEL SKILL

We evaluated the possibility of recommending some subset of ensembles for a given region based on their performance relative to the observed history of that region. The degree to which a model is able to replicate the statistical characteristics of history for a region is called its "skill". Table B-1 presents the skill of several ensembles at estimating basin yield for a set of 21 basins, where skill is defined based on the mean percent deviation in three yield values: (1) at zero reservoir storage, (2) at existing storage, and (3) mean annual runoff. For this assessment, we relied on runoff generated using the Turc-Pike model (Turc 1954, Pike 1964) for both the observed and modeled climate series over the 1941-1990 period. Figures B-1 and B-2 present the mean model bias in mean precipitation and the mean model bias in the coefficient of variation on annual precipitation between 1941-1990. From this preliminary assessment, the ensembles do not appear to perform well systematically, although further inquiry is warranted.

Table B-1. Median percent difference across ensemble members between 1941-1990 historical basin yield and 1941-1990 modeled basin yield. To develop basin yield, the Turc-Pike model is used to generate annual runoff estimates for this period, and then storage yield curves are constructed. The measure below is the median of differences between yield readings from the "observed" and "modeled" storage yield curves: minimum yield, yield at existing storage, maximum yield (MAR). Shaded differences have absolute errors of less than 20%.

Basin	40NCARs	17ECHAMs	6R4CCSM	10R4CSIRO	5R4CAN	5R4GISSH	6R4GISSR
Zambezi	29%	42%	-41%	-27%	>100%	>100%	29%
Nile	>100%	>100%	<-50%	-42%	>100%	>100%	>100%
Niger	>100%	>100%	-33%	**4%**	>100%	>100%	>100%
Volta	>100%	>100%	**8%**	30%	>100%	>100%	>100%

Basin	40NCARs	17ECHAMs	6R4CCSM	10R4CSIRO	5R4CAN	5R4GISSH	6R4GISSR
Congo	>100%	>100%	-37%	-22%	>100%	**2%**	69%
Orange	>100%	>100%	99%	>100%	>100%	44%	93%
Amazon	-39%	-46%	<-50%	<-50%	**9%**	<-50%	<-50%
Orinoco	<-50%	<-50%	<-50%	<-50%	**6%**	<-50%	<-50%
Murray_Australia	**-10%**	>100%	<-50%	-47%	>100%	-36%	27%
Indus	>100%	>100%	<-50%	-37%	>100%	<-50%	<-50%
Huang_He	>100%	>100%	>100%	>100%	>100%	>100%	>100%
Ganges	22%	74%	<-50%	-46%	65%	<-50%	<-50%
Tigris_Euphrates	<-50%	-33%	<-50%	<-50%	**10%**	-37%	<-50%
Baltic	28%	59%	36%	43%	42%	70%	**15%**
Syrdarja	95%	>100%	**-19%**	**-13%**	31%	**-17%**	<-50%
Black_Sea	-21%	**-3%**	-49%	-43%	**9%**	**3%**	<-50%
Colorado	>100%	>100%	>100%	>100%	>100%	>100%	-40%
Columbia	>100%	>100%	89%	97%	>100%	58%	53%
Mississippi	<-50%	<-50%	26%	28%	**-17%**	-36%	<-50%
Southeast_US	-48%	-23%	-25%	**-15%**	**7%**	**-19%**	<-50%
US_Northeast	-24%	-30%	25%	20%	**12%**	**-8%**	-29%

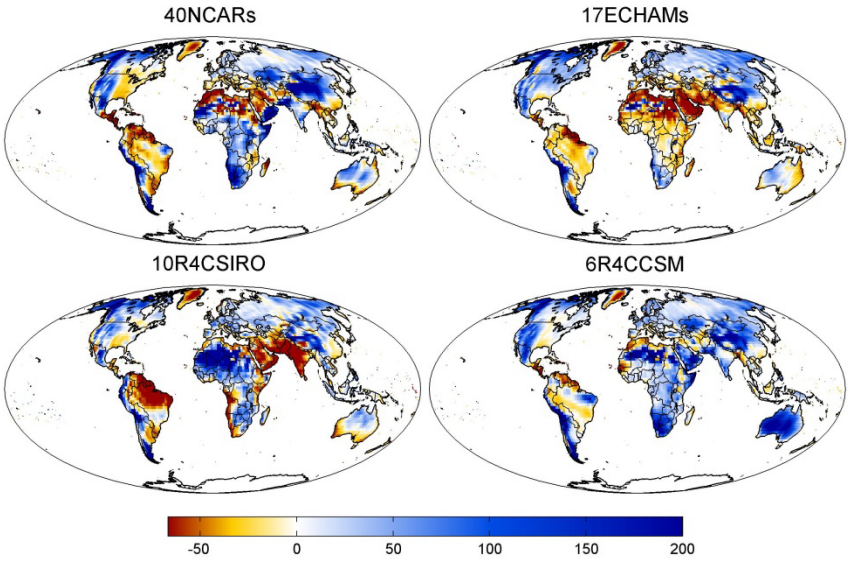

Figure B-1. Mean model bias in mean annual precipitation for four model ensembles relative to the CRU historical baseline between 1941-1990

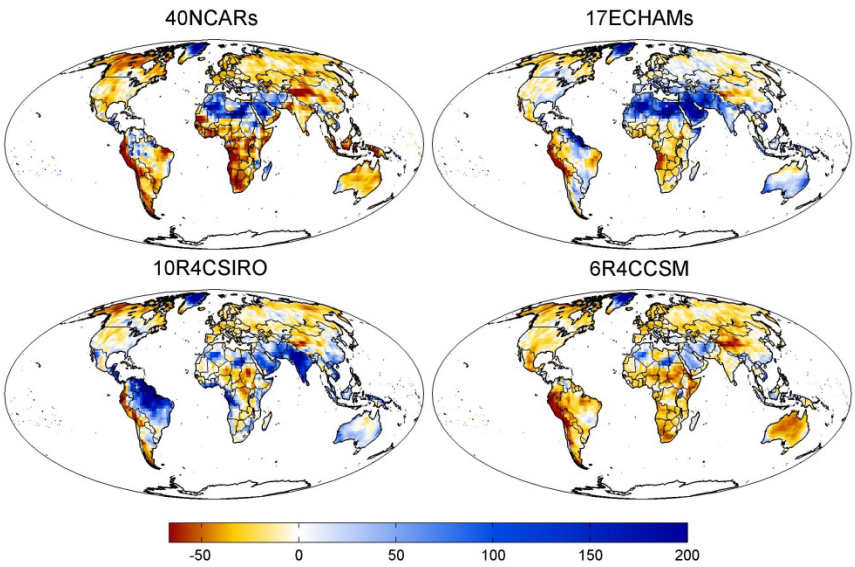

Figure B-2. Mean model bias in annual precipitation COV for four ensembles relative to the CRU historical baseline between 1941-1990.