SUPPLEMENTAL DATA

ASCE *Journal of Water Resources Planning and Management*

Evaluation of Climate Model Performance for Water Supply Studies: Case Study for New York City

Aavudai Anandhi, Donald C. Pierson, and Allan Frei

DOI: 10.1061/(ASCE)WR.1943-5452.0001054

www.ascelibrary.org

Evaluation of Climate Model Performance for Water Supply Studies: Case Study for New York City

Aavudai Anandhi[1], Donald C. Pierson[2], and Allan Frei[3]

Contents

---

[1]Biological Systems Engineering, Florida Agricultural and Mechanical University, Tallahassee, FL, 32307; Center for Water Resources, Florida Agricultural and Mechanical University, Tallahassee, FL, 32307 (corresponding author). Email: anandhi@famu.edu
[2] Limnology Department, Uppsala University Evolutionary Biology Centre, EBC Norbyvägen 18 D, 752 36 Uppsala, Sweden. Email: don.pierson@ebc.uu.se
[3] Hunter College, City University of New York, New York, NY, 10065; CUNY Institute for Sustainable cities, City University of New York, New York, NY, 10065.

GCM evaluation methods

An important part of the GCM evaluation was to test the ability of GCMs in simulating the climate variability and extremes (Randall et al., 2007). However, there is no preferred criteria to choose a subset of GCMs, although several studies have made this choice based on the ability of the GCMs to simulate the local climate (Ashofteh et al., 2015; Elhakeem et al., 2015; Singh et al., 2015; Whateley et al., 2016). The number of GCMs selected in these studies range from one to nine. Evaluation methods such as skill score, correlation coefficients, mean, median, standard deviation, anomalies, root mean square error, bias, extreme indices, empirical orthogonal functions, and principal component analysis have been used in previous studies (Anandhi and Nanjundiah, 2015; Errasti et al., 2010; Frei et al., 2003; Meehl et al., 2007a; Meehl et al., 2007b; Perkins et al., 2007). A good review of available methods and details of earlier studies can be found in Johnson and Sharma (2009) and in Table 8 in Errasti et al. (2010).

Water utilities are increasingly incorporating climate change into their planning activities using several methodologies. Many studies use climate information from only a single GCM (Ashofteh et al., 2015; Fortier and Mailhot, 2015), whereas others incorporate results from multiple climate models (Elhakeem et al., 2015; Islam and Gan, 2014). The latter method allows one to estimate a range of possible outcomes for any particular climate change scenario. Another approach is to use the ensemble mean from multiple climate models (Benestad, 2003; Smith et al., 2009; Tebaldi and Knutti, 2007). Using this approach, only one outcome is produced.

In this study we used daily output from the Intergovernmental Panel for Climate Change Fourth Assessment Report (IPCC AR4) GCM simulations of 20th century climate (20C3M scenario). We use the older GCMs participating in the phase 3 of the Coupled Model Intercomparison Project (CMIP3) rather than more recent phase 5 GCM simulations (CMIP5) because the evaluations

presented here were a component of a larger study that began in 2008, when CMIP5 GCMs were not available. We believe that the approach adopted here for water supply studies can also be used in evaluating CMIP5, since CMIP5 GCMs are "strongly tied to their predecessors" (Knutti et al., 2013; McMahon et al., 2015). We have published several papers that document using a subset of GCMs (Anandhi et al., 2011a; Anandhi et al., 2013a; Matonse et al., 2013; Matonse et al., 2011; Mukundan et al., 2013; Pradhanang et al., 2013; Samal et al., 2013). Here, we document the methods used to choose the subset of GCM data that were used for the NYC water supply climate change simulations

Study region and data

The New York City (NYC) municipal water supply is derived from three large watershed regions: the Croton, the Catskills and the Delaware. The Catskill and the Delaware systems are ~193km north of NYC and west of the Hudson River, and are referred to as the West of Hudson (WOH) watersheds. They provide at least 90% of NYC's daily water demand and are an unfiltered water supply. Water quality is maintained through the protection of natural ecosystem services in the watersheds. The quantity and quality of water in the WOH watersheds are constantly monitored using a system in situ automated monitoring stations as well as periodic manual observations. Water entering the distribution system is chlorinated and treated with UV light.

The WOH watersheds are part of the Eastern Plateau Climate Division of New York, and the area consists of six reservoir watersheds (Cannonsville, Askokan, Nerversink, Schoharie, Rondout, and Pepacton; Figure 1a,b) which encompasses an area of ~4100 km^2. The climate is humid continental and has cold winters and abundant rainfall and snowfall. This region experiences a pronounced seasonal cycle of temperature and a relatively uniform distribution of precipitation throughout the year [Figure-2, in Anandhi et al. (2011b)]. In this region, snowfall historically accounts for about

20% of total precipitation, which is 1000-1200 mm per year. The spatial distribution of temperature is characterized primarily by a southeast-to-northwest gradient. The spatial distribution of precipitation is influenced by the Atlantic Ocean, the Great Lakes, various storm tracks (e.g. coastal versus continental storm systems) and orography (Burns et al., 2007; Frei et al., 2002). Potential future climate change has been examined in some parts of the WOH region (Burns et al., 2007; Frei et al., 2002) and the larger region, e.g. Eastern North American or ENA (Anderson et al., 2010; Christensen et al., 2007; Giorgi and Bi, 2005; Giorgi and Francisco, 2000; Mahlstein and Knutti, 2010 ; Tebaldi et al., 2004; Tebaldi et al., 2005), but never specifically for the full WOH watershed area.

Two methods of spatially averaging GCM results-- CCG and CLG-- are compared in this study for five meteorological variables: precipitation, average, maximum and minimum temperatures, and wind speed (Ppt, Tave, Tmax, Tmin and Wind) during the time-period from 1960-2000. CCG uses a single grid location close to the WOH watershed (purple star in Figure-1a), and for CLG we pooled data from seven grid locations surrounding the WOH watershed (yellow squares in Figure-1a, approximately at 2.5º grids) for the land areas in the sub-continent USA. Two types of observed data (OD1 and OD2) were used in this study. OD1 is a gridded dataset obtained from Maurer et al. (2002) with a daily time-step that has a grid resolution of 1/8-degree. For this application, data from seven grid cells are extracted (yellow boxes in Figure-1). OD2 is based on the daily time-series of air temperature and precipitation collected from meteorological stations (18 stations for precipitation and 3 for temperature) that were either within or adjacent to the six WOH watersheds and wind data collected from a shore-based station near each reservoir. (Figure-1b). Thiessen polygons method was used for averaging precipitation from nearby weather stations, while inverse distance squared weighted averaging for temperature and wind for OD2 (NYCDEP, 2004). For

temperature, the environmental lapse rate was applied to correct differences between the weather station's elevation and the mean watershed elevation while averaging. More details about the datasets can be obtained from previous research (Anandhi et al., 2011b; Anandhi et al., 2013b). In both these datasets, winds are measured at 10m height and temperatures at 2m height.

Daily simulations from 20 GCMs participating in the World Climate Research Programme's (WCRP's) Coupled Model Intercomparison Project phase 3 (CMIP3) multi-model dataset for baseline scenario (20C3M) were downloaded for the five meteorological variables and several of the ensemble members (Table S1). The grids surrounding the study region were extracted and then interpolated to a common 2.5º grid using bilinear interpolation. Data from seven grid cells are extracted (yellow boxes in Figure 1) at similar location to OD1.

Table S1. Names of the GCMs and their acronym, country of origin, and realization numbers for the five variables used in the study are listed. Acronyms are used throughout the text to refer to models.

S.N	GCM I.D *	Acronym	GCM realization numbers or run numbers					Country
			Ppt	Tave	Tmax	Tmin	Wind **	
1	BCCR-BCM2.0	Bcc	1	1	1	1	1	Norway
2	CCSM3	Ncc	1,3,5,6,7,8,9	1,3,5,6,7,8,9	-	-	-	USA
3	CGCM3.1(T47)	cc4	1,2,3,4,5	1,2,3,4,5	1,2,3,4,5	1,2,3,4,5	1,2,3,4,5	Canada
4	CGCM3.1(T63)	cc6	1	1	1	1	1	Canada
5	CNRM-CM3	cnr	1	1	1	1	1	France
6	CSIRO-Mk3.0	cs3	1,2,3	1,2,3	1,2,3	1,2,3	1,2	Australia
7	CSIRO-Mk3.5	cs5	1,1a	1,1a,2,3	1,1a,2,3	1,1a,2,3	1,2,3	Australia
8	ECHAM5/MPI-OM	mpi	1,4	1,4	1,4	1,4	1,4	Germany
9	ECHO-G	miu	1,2,3	1,2,3	1,2,3	1,2,3	2	Germany/Korea
10	FGOALS-g1.0	iap	1,2,3	1,2,3	1,3	1,3	1,2,3	China
11	GFDL-CM2.0	gf0	1	1	1	1	1	USA
12	GFDL-CM2.1	gf1	2	2	2	2	2	USA
13	GISS-AOM	ga0	1	1	1	1	1	USA
14	GISS-ER	gir	1	1	1	1	1	USA
15	INGV-SXG	ing	1	1	1	1	1	
16	INM-CM3.0	inm	1	-	-	-	1	Russia
17	IPSL-CM4	ips	1,2	1,2	1,2	1,2	1,2	France
18	MIROC3.2(hires)	mih	1	1	1	1	1	Japan
19	MIROC3.2(medres)	mim	1,2	1,2,3	1,2,3	1,2,3	1	Japan
20	MRI-CGCM2.3.2	mri	1,2,3,4,5	1,2,3,4,5	1,2,3,4,5	1,2,3,4,5	1	Japan
	Total no. scenarios		**44**	**45**	**38**	**38**	**30**	

*As provided by Lawrence Livermore National Laboratory's Program for Coupled Model Diagnosis and Intercomparison (PCMDI): http://www-pcmdi.llnl.gov/ipcc/model_documentation/ipcc_model_documentation.php

Note: Ppt, Tave, Tmax, Tmin, and Wind in the table are acronyms for precipitation, average, maximum and minimum temperature and wind, respectively. Winds were measured at 10m height and temperatures at 2m height. ** Wind in GCM is calculated from zonal and meridional winds.

Justification for using precipitation, temperatures and wind climate variables

Ppt, Tmax, Tmin, Tave and Wind are important climate variables in the study region that are used as inputs to an integrated system of watershed and reservoir models [NYCDEP, 2013]. Ppt influences the hydrology (e.g. runoff, evapotranspiration, ground water) of the region. This affects the timing, magnitude, and total fluxes of nutrients and turbidity entering the reservoirs. Temperature affects snow accumulation, snowmelt, and evapotranspiration rates, and in turn the timing and delivery of streamflow, nutrients, and turbidity. Temperature can also influence the results of ecological model simulations of plant phenology [Anandhi, 2016]; the growing season length [Anandhi et al., 2013a; Anandhi et al., 2013b]; the timing of wet, dry, warm and cold spells [Anandhi et al., 2016]; and the occurrence of freeze events [Anandhi et al., 2013a]. Temperature based models describe the heat energy received by the plant over a given time period and relate the accumulation of heat energy to plant development or growth processes [Anandhi, 2016]. Similarly, temperature-based models can be used to simulate the freeze events such as first fall freeze and last spring freeze, which can then be used to estimate growing season length [Anandhi et al., 2013a; Anandhi et al., 2013b]. The spells represent periods of excessive warmth, cold, wetness, or dryness [Anandhi et al., 2016]. Extended periods with either excessive or low rainfall, or either high or low temperatures cause stress on plants. Such stress has many ecological and hydrological consequences such as affecting plant growth, development and yield as well as changes in growing season and implications on the water cycle . Wind affects both surface waves and the internal seiche movement, which will in turn influence sediment resuspension. Wind effects also promote vertical mixing which in turn affect vertical water density differences and the timing of the onset and the loss of thermal stratification. Wind impacts vertical mixing in the

reservoirs, which will affect light and nutrient availability and phytoplankton growth and succession.

Methodology

The methodology flowchart in Figure 1c is described in detail in this section.

Step 1. Identify purpose for the GCM evaluation.

Our purpose for GCM evaluation is to use derived outputs from GCMs as inputs for hydrology and water quality models, a water system operations model, and reservoir hydrothermal and water quality models.

Step 2. Identify variables required for evaluation.

The meteorological variables required by each of these models are different. For example, our hydrology model requires precipitation and average air temperatures while our reservoir hydrothermal models require those variables along with wind speed and solar radiation variables.

Step 3. Identify the temporal scale and domain required for evaluation.

Temporal scale refers to the timescale (e.g. daily, monthly, seasonal, annual) of values that are included in the analysis. Temporal domain refers to the time of year (e.g. January, winter, annual) and depend on the purpose of the study. For our analysis we considered GCM evaluation at a daily temporal scale and seasonal temporal domain (summer, winter, autumn, spring).

Step 4. Identify characteristics of variable for evaluation.

The characteristics of the variable refer to the statistical properties-- such as mean, median, extreme events, variance, etc.-- that are important to the study. For our study, the entire frequency distribution of the variable was identified as the characteristics of variable for evaluation.

Step 5. Identify appropriate performance metric for ranking models.

Different metrics focus on different characteristic of the variable. We chose the probability-based skill score (PSS) for our study as it is appropriate for studying the entire frequency distribution of the variable.

Step 6. Identify spatial scale.

The spatial scale of the variable can be identified in two ways (CLG, CCG). The synoptic scale of the variable can be used as the spatial scale for evaluation and it may vary with variable/region of study. The synoptic scale, also known as the large scale or the cyclonic scale in meteorology, is a horizontal length scale of the order of 1000 kilometers (about 620 miles) or more. In circumstances where the synoptic scale is not known in advance, evaluation can be carried out in various spatial scales ranging from one, four, seven or more surrounding GCM grids before selecting a spatial scale. In our study, we used the second approach and examined the performance for two spatial scales in relation to the study region, namely CLG and CCG, and finally choose CLG as our spatial scale.

Step 7. Calculate performance index.

The index was calculated for each variable, GCM and temporal domain.

Step 8. Rank the GCMs.

Based on performance index values for each variable / temporal domain, the GCMs were ranked.

Step 9. Identify the criteria for model selection.

The criteria for model selection may be broadly divided into criteria to eliminate poorly performing models and criteria to select models that clearly perform better than the majority. Criteria may depend on the resources available. Thresholds of performance index values or the number of models have both been used.

Skill of CMIP3 models

The skill of the models was evaluated using basic statistical measures and probability distribution function (PDF) based skill scores (*SS*). For each meteorological variable, season, and spatial scale, GCM simulations were evaluated by comparing them individually with the two observation datasets (OD1 and OD2). The four seasons used are: winter (DJF), spring (MAM), summer (JJA), and fall (SON), representing the months of December through February, March through May, June through August, and September through November respectively. The statistical measures were composed of parametric (mean, standard deviation, minimum, maximum) and non-parametric [percentiles: 5^{th}, 25^{th}, 50^{th}, 75^{th}, 95^{th}; interquartile range (IQR)] measures. PDF based *SS* are calculated from the overlapping area between the PDFs (observed and GCM) (Anandhi and Nanjundiah, 2015). In the first case (referred to as CCG), the *SS* was obtained from the PDF from the single GCM grid closest to the centroid of the watershed (represented as star in Figure-1) and the PDF from OD2. In the second case (referred to as CLG), the *SS* was obtained from PDF from all the seven GCM land grids surrounding the WOHs and the PDF from OD1. In cases when there was more than one time-series for an observation or GCM, a PDF was constructed by the combined dataset from all timeseries to develop the representative distribution. For example, for each GCM in CLG, a combined dataset from all seven grids was used to construct the representative distribution.

SS is estimated mathematically using equations in (Anandhi and Nanjundiah, 2015). The advantage of this method is its simplicity and visual intuitiveness. The *SS* is the overlapping area between the PDFs (observed and GCM), and ranges from 0 to 1. The highest possible *SS* equals 1 and occurs only when the modeled and observed PDFs are same (complete overlap of PDFs). GCMs with a *SS* of 0 indicates that there is no overlap of model and observed PDFs. *SS* can be

used in evaluating multiple variables at different spatial and temporal scales. The use of SS to evaluate GCM models implicitly assumes that the match between a model's baseline simulation and observed historical data will be indicative of the models' ability to simulate future climate conditions. While there is no strong evidence that this is the case, it is logical to base the GCM evaluation on comparison to historical data since future data was not available and the convergence of GCM future climate conditions were not estimated. More details of *SS* can be obtained from Anandhi et al. (2011b) and Anandhi and Nanjundiah (2015).

Ranking procedure

In this study, the skill of CMIP3 models in simulating the variables (Ppt, Tave, Tmax, Tmin, Wind) for CLG were represented using (1) PDFs (Figure-2) of the observed (black bold line), the range of GCM simulations (shaded portion), the multi-model ensemble mean (red bold line) and the median (red dotted line); (2) boxplots of statistical measures, observed mean statistic (triangle, Δ) and GCM ensemble mean statistics (circle, o) in the four seasons (Figure-3a,b); and (3) bar graphs of skill scores of all GCMs in the four seasons (Figure-S1). For each combination of meteorological variable and season, ranking was carried out for both the statistics and the skill scores. For each variable and season, GCMs were arranged in the descending order of *SS,* with the GCM having the highest *SS* given rank 1; when ranking by the basic statistical measures, the GCM's whose statistical measures had the smallest difference was assigned rank 1. The ranked GCMs were then compared. In this study, the skill of CMIP3 models in simulating the variables (Ppt, Tave, Tmax, Tmin, Wind) Similar analysis was carried out for CCG scale using OD2 dataset. Only boxplots of statistical measures, observed mean statistic (triangle, Δ) and GCM ensemble mean statistics (circle, o) in the four seasons are shown (Figure-S2 to S6).

Results and Discussion

Comparison of CMIP3 models (CLG scale) to observed data (OD1)

GCM Models tend to overestimate the number of small Ppt events (1 to 3 mm/day, Figure-2) and small to medium Ppt events (Figure-3, minimum; 5^{th} to 75^{th} percentiles) but underestimate larger events (3 to 16 mm/day, Figure-2, Figure-3, 95^{th} percentile, maximum except in winter and spring). Observed Ppt was better simulated by most GCMs in summer and fall seasons, though they overpredicted the mean observed Ppt during winter and spring as well as the interquartile range in most seasons (except winter) (Figure-3, mean). Almost all models underpredicted the median and standard deviation of Ppt in all seasons (Figure-2, median). The *SS* ranged from 0.65 to 0.95 for Ppt in all the four seasons. Similar overestimation of small events were observed in Australia (Perkins et al., 2007) and India (Anandhi and Nanjundiah, 2015). Overestimation of small events 'GCM drizzle' can contribute to the overestimation of total precipitation because the small amount of water associated with these events covers the entire area of the grid cell in the model, while in reality small events rarely show a homogenous large-scale distribution. Between-model variability was highest for smaller events and during JJA. This could be because summer Ppt is typically convective in nature, and is probably less reliably simulated by GCMs than synoptic features since local convection is a sub-grid-scale process (Toews and Allen, 2009). Biases (differences between GCM simulations and observations) in Ppt can influence hydrological model results in a number of ways. For example, the overestimation of small events and underestimation of large events may affect the timing, magnitude, and total fluxes of nutrients and turbidity entering the reservoirs. Therefore, it seems likely that those GCMs that underestimate the magnitudes of large events will underestimate the magnitude of nutrient flux and turbidity (Anandhi et al., 2016).

In general, the statistical distributions of the temperature variables (Tave, Tmax, Tmin) were reasonably well captured by models when compared with Ppt and Wind. Similar results were observed across OD1. Among the temperatures, Tave was better simulated than Tmax and Tmin. The models tended to underestimate the number of cold days, especially during winter season (Tmax and Tmin in Figure-3, minimum, 5^{th} to 25^{th} percentiles), and overestimate the number of warm days during winter season (Tmax and Tmin in Figure-3, maximum, 75^{th} and 95^{th} percentiles). The largest temperature biases, as well as the largest between-model variability, were found in summer (Row 3 in Figure-2). In all seasons, the models tended to overestimate the frequency of lower Tmax values and underestimate the frequency of higher Tmax values. The reverse was observed for Tmin in summer and fall seasons. The *SS* ranged from 0.55 to 0.95 for Tave, 0.3 to 0.95 for Tmax, 0.4 to 0.95 for Tmin, in the four seasons (Figure-4a). GCM temperature biases could affect snow accumulation, snowmelt, and evapotranspiration rates, and, in turn, the timing and delivery of streamflow, nutrients, and turbidity. It could also influence the results of ecological model simulations in terms of plant phenology (Anandhi, 2016), the length of growing seasons (Anandhi et al., 2013a; Anandhi et al., 2013b), the timing of warm and cold spells (Anandhi et al., 2016), and the occurrence of freeze events (Anandhi et al., 2013b). The tendency of GCMs to under predict minimum daily air temperature could affect reservoir hydrodynamics and mixing because nighttime cooling of the reservoir affects convective mixing, and additionally diurnal and seasonal variations in reservoir thermal structure.

The simulated Wind distribution compared unfavorably to the observed distribution (Figure-2). Most models overestimated smaller winds (Figure-3, minimum; Figure 2, 0-5 m/sec) and underestimated the mean and median winds, as well as the frequency of large events. The largest model biases and the largest between-model variability were found for smaller events. Models

tended to overestimate the frequency of small events (0-5 m/sec) and underestimate the frequency of large events. Similar results were observed for CCG (Figure-S6). Models overestimated the minimum wind (Figure-2; minimum) and underestimated the median and larger wind-events. Similar results were observed for CCG using OD2 dataset (Figure-S6). The *SS* ranged from 0.2 to 0.95 for Wind in the four seasons (Figure-4a). Wind distributions across the WOH region which have a significant topographical variation are variable and difficult to capture by the relatively coarse resolution of the GCMs. The tendency for these GCMs to underestimate small wind events (< 5 m/s) and overestimate large events (> 5 m/s) may also influence the hydrodynamics and mixing processes in the reservoir. Overestimated wind will lead to deeper and more intense vertical mixing, which will affect light and nutrient availability as well as phytoplankton growth and succession. Wind also affects both surface waves and the internal seiche movement, which will in turn influence sediment resuspension. Inaccurate wind forcing in the reservoir models will have differing seasonal effects depending on vertical density differences and can be expected to influence the timing of the onset and loss of thermal stratification.

The GCM ranks and SS ranking procedure

In general, no one model was consistently ranked best by SS for all of the meteorological variables (Ppt, Wind, Tave, Tmax and Tmin), or during all of the seasons (DJF, MAM, JJA, SON). The results of the *SS* ranking procedure for all ensemble members of a GCM are summarized as a function of season (Figure-S1) and *SS* arranged in descending order for each variable (Figure-4b). The closeness of the statistical measures of the GCM simulation data to that of the observations can be observed in Figures 3.

The magnitudes of skill scores did not vary between seasons for precipitation and wind, though there was a higher variability in skill during summer for Ppt. For temperature, the procedure

showed a lower magnitude of *SS* during summer. Among seasons, spring had a higher mean/median skill score for all five variables. Fall's mean/median skill scores were also high for temperature variables and wind. For each meteorological variable, different ensemble members of the same model had similar *SS* in the *SS* ranking procedure. This can indicate that the skill scores were not due to random or chaotic processes but were in fact related to model formulation. We calculate an average *SS* for each GCM and meteorological variable.

We found no clear relationship between *SS* and three model characteristics (horizontal resolution, convective scheme and flux correction). The reasons for the lack of a clear relationship could be because the climate models often shared similar types of code, used common input datasets, and are developed by scientists with similar expertise. Furthermore, some institutions have produced more than one GCM that shares many similarities. This may result in some models having similar biases (Jun et al., 2008; Knutti et al., 2010). Our results corroborate other studies carried out at multiple locations and temporal scales, which have found that it may not be straightforward to associate GCM performance with model characteristics (Anandhi et al., 2011b; Anandhi and Nanjundiah, 2015; Dai, 2006; Kim et al., 2008; Kripalani et al., 2007).

Overall rankings based on the CLG dataset showed that cc6, cc4, gao, ing and cs0 had the highest skill scores in this region; however, the ranking for individual variables was different (Table 1). Inm and ncc GCMs were eliminated while averaging because the model simulations for all the five meteorological variables were not available. Cs5 seemed to have consistently low ranks in the region for Ppt and temperature variables. Gir had very different statistics (which were outside the range of figures and so not shown) compared to the rest of the models for Ppt. Ranking based on the CCG dataset were similar to that described above for CLG (only statistics shown for CCG, S2 to S6).

There is no obvious way to choose a subset of GCM models from the results of this study that is clearly superior since there is a gradual, not abrupt, decrease in model skill as one goes from highest to lowest skill score. Furthermore, different models performed better for different meteorological variables and performance measures. This can greatly complicate choosing a subset of models if the models simulations depend on multiple meteorological drivers. The simplest way to choose a subset of models is to identify how many models are appropriate for the variable(s) of interest and to choose from these based on the combined *SS* rankings that include all needed meteorological variables. However, based on our results, we recommend using as many GCM datasets as possible, as we were not able to identify a clear subset of models that was superior for all the meteorological variables used in our water supply simulations. Combining the *SS* ranking procedure that utilizes the entire distribution of the meteorological variables with the ranking procedure that utilizes extreme of the distribution can be useful in some cases.

Since no one model was best for the various combinations used in this study, one of the subjective decisions required for the type of analysis presented here is the weighting that should be given to the meteorological variables included in the overall evaluation. As an example, precipitation is generally a key variable for water supply but is not simulated as well as air temperature in GCMs. GCMs can have good skill in simulating large scale patterns of mean precipitation such as the zonal mean distribution, but they lack skill in accurately simulating regional distributions of precipitation (Wehner et al., 2010). Further, the regional orography is smoothed and not correctly represented by the large grid sizes used in GCMs (Anandhi and Nanjundiah, 2015; Pan et al., 2011), and the relevant cloud microphysical processes are probably inaccurately simulated, in climate models (Wehner et al., 2010).

Depending on the location and research question of other variables (such as temperature and wind speed) may be as or even more important when deciding (Zion et al., 2011). For example, for winter precipitation, the state of precipitation (i.e. liquid or solid) and snow melt-- both of which depend on temperature-- will influence the magnitude and timing of streamflow and associated constituent loads entering the reservoirs (Mukundan et al., 2013; Pradhanang et al., 2013). Turbidity loading in reservoirs also depends on temperature and precipitation (Rossi et al., 2016). If the application is related to reservoir water quality, then thermal, dynamical, and/or biological processes which influence phytoplankton growth may be critical (Samal et al., 2012). In such a case, temperature and wind speed may be more important than precipitation. The relative ranking given to each meteorological variable in the skill score ranking will therefore depend to some extent on the modelling application to which the data will be input

Despite decades of work on developing various decision support tools for water managers, there remains a disconnect between the availability of future climate scenarios and the use of this information in water quality management. Furthermore, there is a need for tailoring the information in climate scenarios to make them more applicable for use in models that support decision making (Bolson et al., 2013; Kirchhoff et al., 2015). This study addresses one part of this gap. The methodology used by NYCDEP to screen GCMs for use in water supply simulations is presented, so that it can be of use to others involved in climate change evaluations where there is a need to limit their choice of GCM data. We found that there were limitations in the skill score approach since the ranked order of the GCMs vary with spatial, temporal scales, versions of the GCMs and the meteorological parameter of interest. For example, GCM rankings at daily time-scales could be different than rankings at monthly scales; CMIP3 GCM rankings could be different from CMIP5 rankings. We observed that the top-ranked GCMs at the CCG and CLG scales were

different. Consequently, we concluded in our case that we could not simply identify a sub-set of GCMs that would be suitable for all our simulation models. However, the results presented here are specific to the NYC water supply region and ongoing work in this area will make use of the CMIP5 data. The methodology presented here could be of greater value in different locations and in cases where fewer meteorological variables will be evaluated.

The evaluation procedure presented here is only one step in the preparation of future climate scenarios that were used to drive NYCDEP watershed and reservoir models. Chosen GCMs can be further processed using various downscaling approaches (Anandhi et al., 2011a; Anandhi et al., 2009; Anandhi et al., 2012; Anandhi et al., 2014; Anandhi et al., 2008; Fortier and Mailhot, 2015; Johnson et al., 2012). Then input into other models (e.g. hydrologic models) to estimate model results at finer spatial scales.

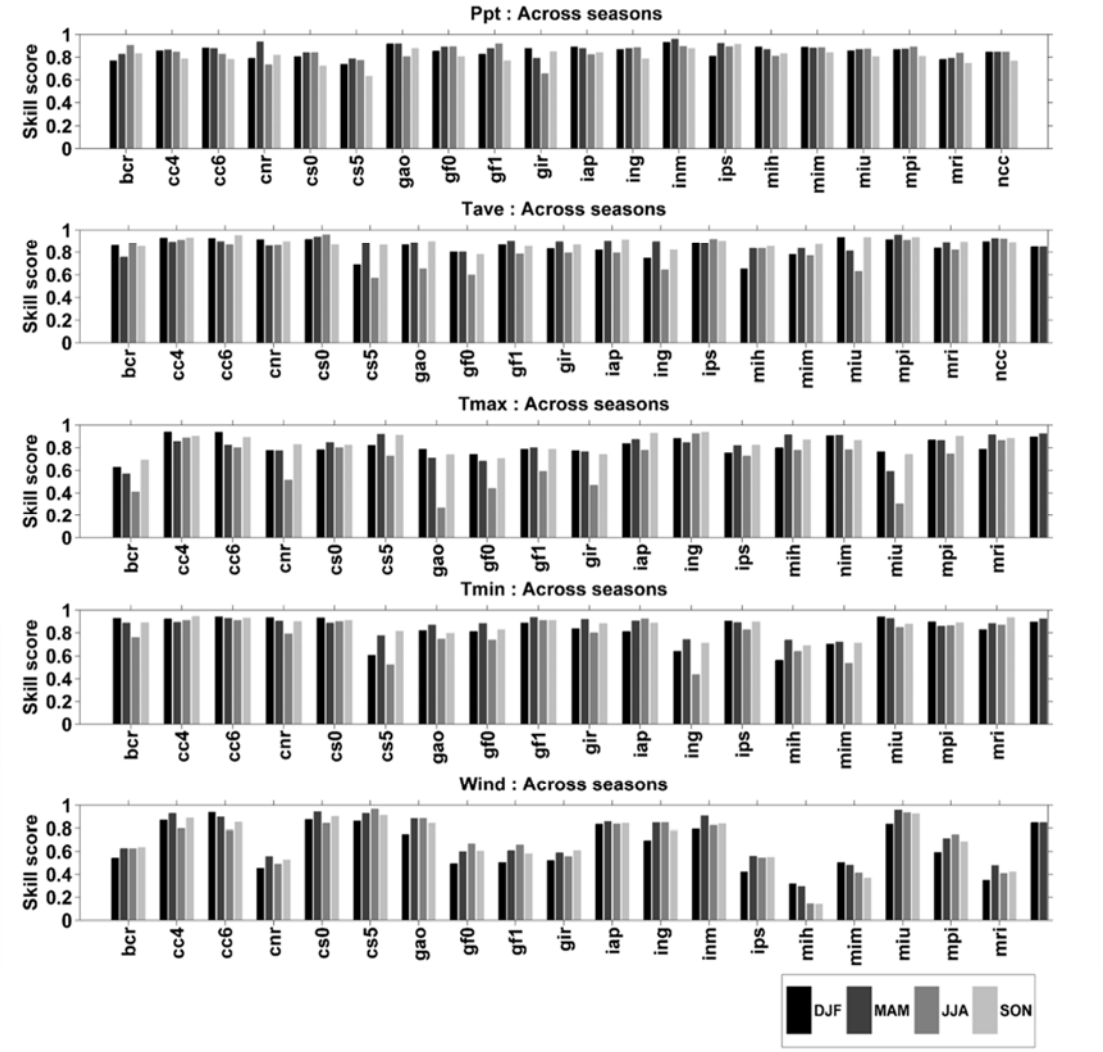

Figure S1. Skill scores of AR4 climate models. Barplots indicate the variation in skill scores for five meteorological variables (Ppt, Tave, Tmax, Tmin, Wind). Four bars are plotted for each GCM/realization combination, where each bar represents a season (DJF, MAM, JJA, and SON) for CLG using OD1 dataset.

Statistics estimated for CCG region using OD2

The GCMs were evaluated at CCG spatial scale based on a number of parametric and non-parametric measures, including the 5^{th}, 25^{th}, 50^{th}, 75^{th}, and 95^{th} percentiles; the mean; the interquartile range (IQR); and the standard deviation in addition to the skill scores (figures S2-S6).

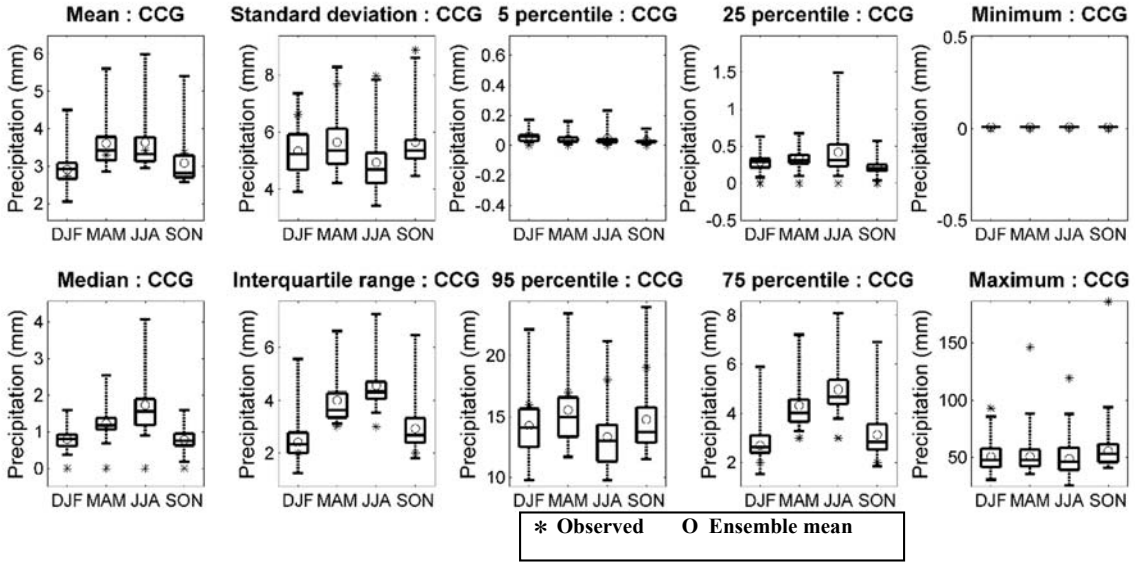

Figure S2. Precipitation statistics. The box and whisker plots indicate statistics calculated for daily precipitation for the various AR4 climate models across the four seasons, namely DJF, MAM, JJA, and SON for CCG spatial scale. These plots are interpreted as follows: the middle line shows the median value, the top and bottom of box show the upper and lower quartiles (i.e., 75^{th} and 25^{th} percentile values), and the whiskers show the minimum and maximum model values. The '*' and circle in the figure represents the observed and GCM ensemble mean of the statistics respectively for seasons DJF, MAM, JJA, SON. The GISS–er model statistic values that were calculated were excluded from the plots. The results are for CCG using OD2 dataset.

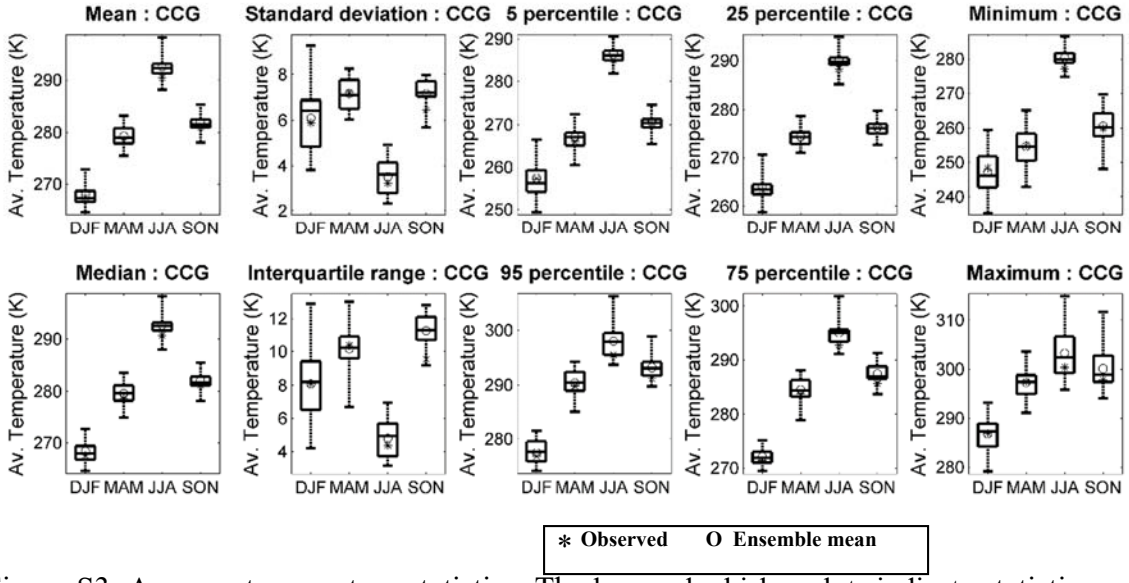

Figure S3. Average temperature statistics. The box and whisker plots indicate statistics calculated for daily average temperature from various AR4 climate models across the four seasons, namely DJF, MAM, JJA, and SON for CCG spatial scale. These plots are interpreted as follows: the middle line shows the median value, the top and bottom of box show the upper and lower quartiles (i.e., 75th and 25th percentile values), and the whiskers show the minimum and maximum model values. The '*' and circle in the figure represents the observed and GCM ensemble mean of the statistics respectively for seasons DJF, MAM, JJA, and SON.

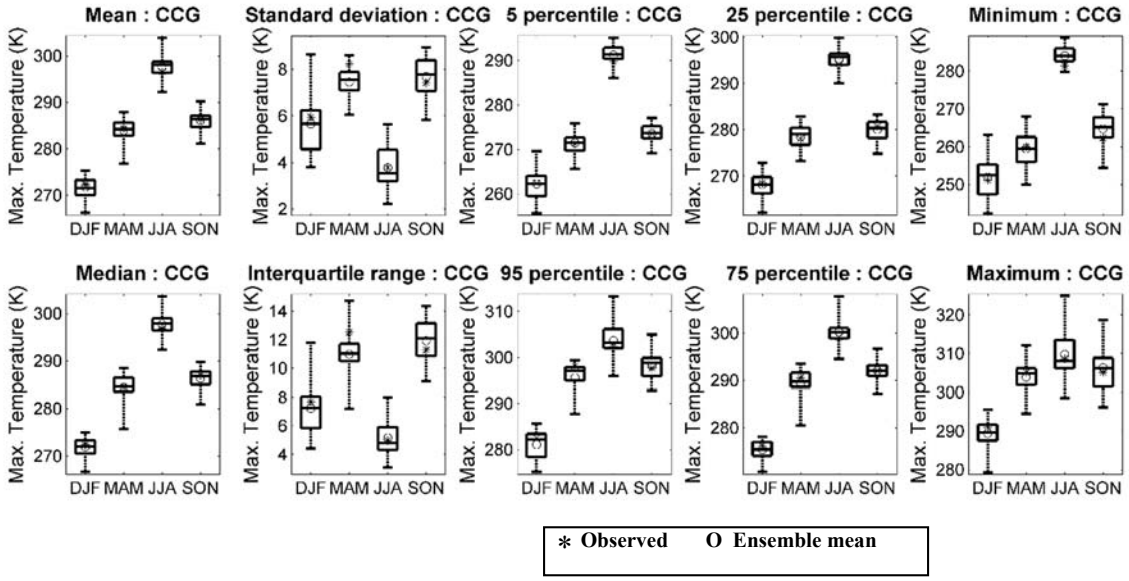

Figure S4. Maximum temperature statistics. The box and whisker plots indicate statistics calculated for daily maximum temperature calculated from various AR4 climate models across the four seasons, namely DJF, MAM, JJA, and SON for CCG spatial scale. These plots are interpreted as follows: the middle line shows the median value, the top and bottom of box show the upper and lower quartiles (i.e. 75th and 25th percentile values), and the whiskers show the minimum and maximum model values. The blue '*' and circle in the figure represents the observed and GCM ensemble mean of the statistics respectively for seasons DJF, MAM, JJA, and SON.

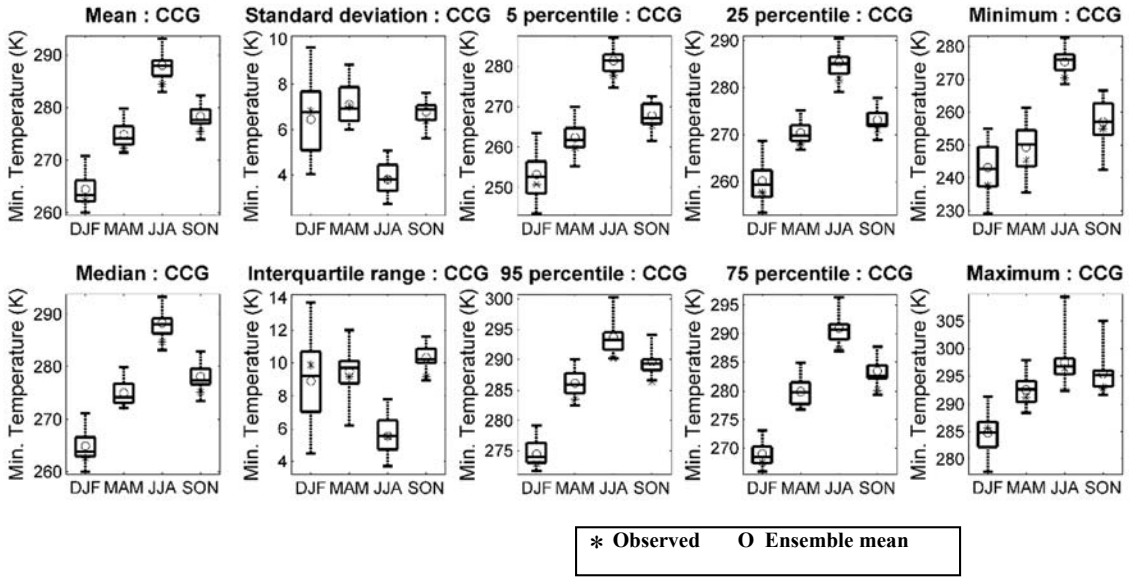

Figure S5. Minimum temperature statistics. The box and whisker plots indicate statistics calculated for daily minimum temperature calculated from various AR4 climate models across the four seasons, namely DJF, MAM, JJA, and SON for CCG spatial scale. These plots are interpreted as follows: the middle line shows the median value, the top and bottom of box show the upper and lower quartiles (i.e. 75^{th} and 25^{th} percentile values), and the whiskers show the minimum and maximum model values. The '*' and circle in the figure represents the observed and GCM ensemble mean of the statistics respectively for seasons DJF, MAM, JJA, and SON.

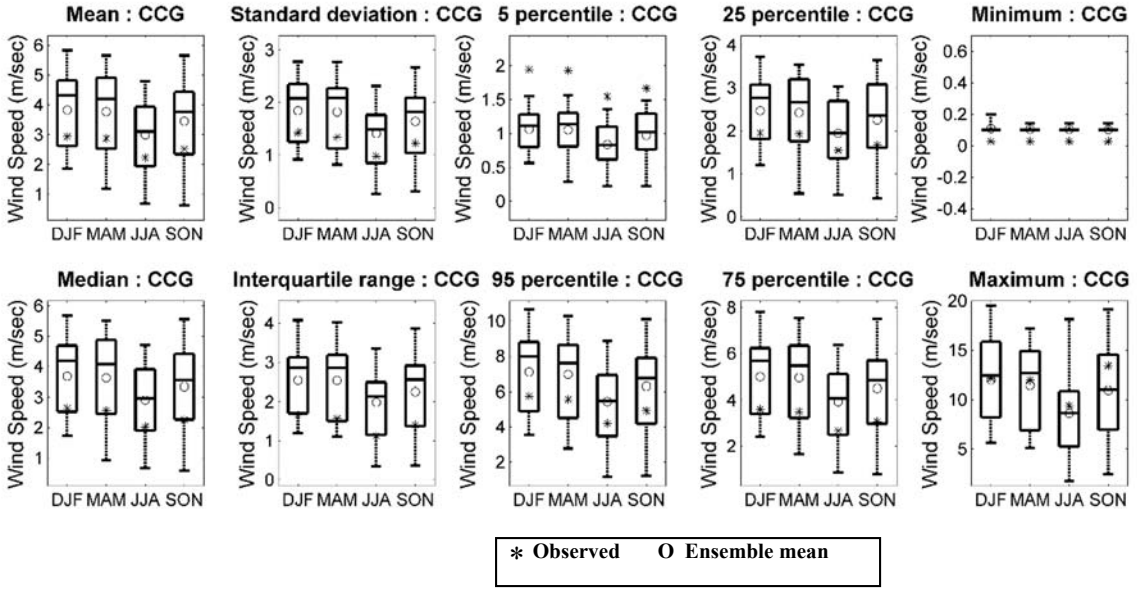

Figure S6. Wind speed statistics. Box and whisker plots indicate statistics calculated for daily wind speed calculated from various AR4 climate models across the four seasons, namely DJF, MAM, JJA, and SON for CCG spatial scale. These plots are interpreted as follows: the middle line shows the median value, the top and bottom of box show the upper and lower quartiles (i.e., 75th and 25th percentile values), and the whiskers show the minimum and maximum model values. The '*' and circle in the figure represents the observed and GCM ensemble mean of the statistics respectively for seasons DJF, MAM, JJA, and SON.

References

Anandhi, A., 2016. Growing degree days–Ecosystem indicator for changing diurnal temperatures and their impact on corn growth stages in Kansas. Ecological Indicators, 61: 149-158.

Anandhi, A. et al., 2011a. Examination of change factor methodologies for climate change impact assessment. Water Resources Research, 47(3).

Anandhi, A. et al., 2011b. AR4 climate model performance in simulating snow water equivalent over Catskill Mountain watersheds, New York, USA. Hydrological Processes, 25(21): 3302-3311.

Anandhi, A. et al., 2016. Changes in spatial and temporal trends in wet, dry, warm and cold spell length or duration indices in Kansas, USA. International Journal of Climatology, 36(12): 4085-101.

Anandhi, A. and Nanjundiah, R.S., 2015. Performance evaluation of AR4 Climate Models in simulating daily precipitation over the Indian region using skill scores. Theoretical and Applied Climatology, 119(3-4): 551-566.

Anandhi, A. et al., 2013a. Long-term spatial and temporal trends in frost indices in Kansas, USA. Climatic Change, 120(1-2): 169-181.

Anandhi, A., Srinivas, V., Kumar, D.N. and Nanjundiah, R.S., 2009. Role of predictors in downscaling surface temperature to river basin in India for IPCC SRES scenarios using support vector machine. International Journal of Climatology, 29(4): 583-603.

Anandhi, A., Srinivas, V., Kumar, D.N. and Nanjundiah, R.S., 2012. Daily relative humidity projections in an Indian river basin for IPCC SRES scenarios. Theoretical and Applied Climatology, 108(1-2): 85-104.

Anandhi, A., Srinivas, V., Kumar, D.N., Nanjundiah, R.S. and Gowda, P.H., 2014. Climate change scenarios of surface solar radiation in data sparse regions: a case study in Malaprabha River Basin, India. Clim Res, 59(3): 259-270.

Anandhi, A., Srinivas, V., Nanjundiah, R.S. and Nagesh Kumar, D., 2008. Downscaling precipitation to river basin in India for IPCC SRES scenarios using support vector machine. International Journal of Climatology, 28(3): 401-420.

Anandhi, A. et al., 2013b. Past and future changes in frost day indices in Catskill Mountain region of New York. Hydrological Processes, 27(21): 3094-3104.

Anderson, B.T., Hayhoe, K. and Liang, X.-Z., 2010. Anthropogenic-induced changes in twenty-first century summertime hydroclimatology of the Northeastern US. Climatic Change, 99: 403-423.

Ashofteh, P., Haddad, O. and Loáiciga, H., 2015. Evaluation of Climatic-Change Impacts on Multiobjective Reservoir Operation with Multiobjective Genetic Programming. Journal of Water Resources Planning and Management, 0(0): 04015030.

Benestad, R.E., 2003. What Can Present Climate Models Tell Us About Climate Change? Climatic Change, 59.

Bolson, J., Martinez, C., Breuer, N., Srivastava, P. and Knox, P., 2013. Climate information use among southeast US water managers: beyond barriers and toward opportunities. Regional Environmental Change, 13(1): 141-151.

Burns, D.A., Julian, K. and McHale, M.R., 2007. Recent climate trends and implications for water resources in the Catskill Mountain region, New York, USA. Journal of hydrology, 336: 155-170.

Christensen, J.H. et al., 2007. Regional climate projections. Climate change 2007. The physical science basis contribution of working group I to the fourth assessment report of the Intergovernmental Panel on Climate Change. Cambridge University Press, Cambridge, United Kingdom and New York, USA.

Dai, A., 2006. Precipitation characteristics in eighteen coupled climate models. Journal of Climate, 19(18): 4605-4630.

Elhakeem, A., Elshorbagy, W., AlNaser, H. and Dominguez, F., 2015. Downscaling Global Circulation Model Projections of Climate Change for the United Arab Emirates. Journal of Water Resources Planning and Management, 0(0): 04015007.

Errasti, I., Ezcurra, A., Sáenz, J. and Ibarra-Berastegi, G., 2010. Validation of IPCC AR4 models over the Iberian Peninsula. Theoretical and Applied Climatology, (In Press).

Fortier, C. and Mailhot, A., 2015. Climate Change Impact on Combined Sewer Overflows. Journal of Water Resources Planning and Management, 141(5): 04014073.

Frei, A., Armstrong, R.L., Clark, M.P. and Serreze, M.C., 2002. Catskill Mountain Water Resources: Vulnerability, Hydroclimatology, and Climate Change Sensitivity. Annals of the Association of American Geographers. , 92(2): 202-224.

Frei, A., Miller, J.A. and Robinson, D.A., 2003. Improved simulations of snow extent in the second phase of the Atmospheric Model Intercomparison Project (AMIP-2). Journal of Geophysical Research, 108(D12): 4369-4386.

Giorgi, F. and Bi, X., 2005. Updated regional precipitation and temperature changes for the 21st century from ensembles of recent AOGCM simulations. Geophysical Research letters, 32: L21715.

Giorgi, F. and Francisco, R., 2000. Uncertainties in regional climate change prediction: a regional analysis of ensemble simulations with the HADCM2 coupled AOGCM. Climate Dynamics, 16: 169-182.

Islam, Z. and Gan, T., 2014. Effects of Climate Change on the Surface-Water Management of the South Saskatchewan River Basin. Journal of Water Resources Planning and Management, 140(3): 332-342.

Johnson, F. and Sharma, A., 2009. Measurement of GCM skill in predicting variables relevant for hydroclimatological assessments. Journal of Climate, 22: 4373-4382.

Johnson, T., Butcher, J., Parker, A. and Weaver, C., 2012. Investigating the Sensitivity of U.S. Streamflow and Water Quality to Climate Change: U.S. EPA Global Change Research Program's 20 Watersheds Project. Journal of Water Resources Planning and Management, 138(5): 453-464.

Jun, M., Knutti, R. and Nychka, D.W., 2008. Spatial analysis to quantify numerical model bias and dependence: how many climate models are there? Journal of the American Statistical Association, 103(483): 934-947.

Kim, H.-J., Wang, B. and Ding, Q., 2008. The global monsoon variability simulated by CMIP3 coupled climate models*. Journal of Climate, 21(20): 5271-5294.

Kirchhoff, C.J., Lemos, M.C. and Kalafatis, S., 2015. Climate Risk Management.

Knutti, R., Furrer, R., Tebaldi, C., Cermak, J. and Meehl, G.A., 2010. Challenges in combining projections from multiple climate models. Journal of Climate, 23(10): 2739-2758.

Knutti, R., Masson, D. and Gettelman, A., 2013. Climate model genealogy: Generation CMIP5 and how we got there. Geophys. Res. Lett., 40: 1194–1199.

Kripalani, R., Oh, J. and Chaudhari, H., 2007. Response of the East Asian summer monsoon to doubled atmospheric CO2: Coupled climate model simulations and projections under IPCC AR4. Theoretical and Applied Climatology, 87(1-4): 1-28.

Mahlstein, I. and Knutti, R., 2010 Regional climate change patterns identified by cluster analysis. Climate Dynamics, in press.

Matonse, A.H. et al., 2013. Investigating the impact of climate change on New York City's primary water supply. Climatic Change, 116(3-4): 437-456.

Matonse, A.H. et al., 2011. Effects of changes in snow pattern and the timing of runoff on NYC water supply system. Hydrological Processes, 25(21): 3278-3288.

Maurer, E., Wood, A., Adam, J., Lettenmaier, D. and Nijssen, B., 2002. A Long-Term Hydrologically Based Dataset of Land Surface Fluxes and States for the Conterminous United States*. Journal of Climate, 15(22): 3237-3251.

McMahon, T., Peel, M. and Karoly, D., 2015. Assessment of precipitation and temperature data from CMIP3 global climate models for hydrologic simulation. Hydrology and Earth System Sciences, 19(1): 361-377.

Meehl, G., Arblaster, J. and Tebaldi, C., 2007a. Contributions of natural and anthropogenic forcing to changes in temperature extremes over the United States. Geophysical Research Letters, 34(L19709).

Meehl, G. et al., 2007b. The WCRP CMIP3 multimodel dataset: a new era in climate change research. Bulletin American Meteorological Society, 88: 1383-1394.

Mukundan, R. et al., 2013. Suspended sediment source areas and future climate impact on soil erosion and sediment yield in a New York City water supply watershed, USA. Geomorphology, 183: 110-119.

Pan, L.-L. et al., 2011. Influences of climate change on California and Nevada regions revealed by a high-resolution dynamical downscaling study. Climate Dynamics, 37(9-10): 2005-2020.

Perkins, S.E., Pitman, A.J., Holbrook, N.J. and McAneney, J., 2007. Evaluation of the AR4 Climate Models' Simulated Daily Maximum Temperature, Minimum Temperature, and Precipitation over Australia Using Probability Density Functions. Journal of Climate, 20(17).

Pradhanang, S.M. et al., 2013. Streamflow responses to climate change: Analysis of hydrologic indicators in a New York City water supply watershed. JAWRA Journal of the American Water Resources Association, 49(6): 1308-1326.

Randall, D.A. et al., 2007. Cilmate Models and Their Evaluation. Climate Change 2007: The Physical Science Basis. Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change. Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA.

Rossi, N., DeCristofaro, L., Steinschneider, S., Brown, C. and Palmer, R., 2016. Potential Impacts of Changes in Climate on Turbidity in New York City's Ashokan Reservoir. Journal of Water Resources Planning and Management: 04015066.

Samal, N.R. et al., 2013. Modelling potential effects of climate change on winter turbidity loading in the Ashokan Reservoir, NY. Hydrological Processes, 27(21): 3061-3074.

Samal, N.R. et al., 2012. Impact of climate change on Cannonsville reservoir thermal structure in the New York City Water Supply. Water Quality Research Journal of Canada, 47(3-4): 389-405.

Singh, H., Sinha, T. and Sankarasubramanian, A., 2015. Impacts of Near-Term Climate Change and Population Growth on Within-Year Reservoir Systems. Journal of Water Resources Planning and Management, 141(6): 04014078.

Smith, R.L., Tebaldi C., Nychka D. and Mearns L.O., 2009. Bayesian modeling of uncertainty in ensembles of climate models. Journal of the American Statistical Association, 104: 97-116.

Tebaldi, C. and Knutti, R., 2007. The use of the multi-model ensemble in probabilistic climate projections. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 365(1857): 2053-2075.

Tebaldi, C., Mearns, L.O., Nychka, D. and Smith, R.L., 2004. Regional probabilities of precipitation change: a Bayesian analysis of multimodel simulations. Geophysical Research Letters, 31: L24213.

Tebaldi, C., Smith R. W., Nychka D. and Mearns L.O., 2005. Quantifying uncertainty in Projections of Regional Climate Change: a Bayesian Approach to the Analysis of Multimodel Ensembles. Journal of Climate, 18(10): 1524-1540.

Toews, M.W. and Allen, D.M., 2009. Evaluating different GCMs for predicting spatial recharge in an irrigated arid region. Journal of Hydrology, 374(3-4): 265-281.

Wehner, M.F., Smith, R.L., Bala, G. and Duffy, P., 2010. The effect of horizontal resolution on simulation of very extreme US precipitation events in a global atmosphere model. Climate Dynamics, 34(2-3): 241-247.

Whateley, S., Steinschneider, S. and Brown, C., 2016. Selecting Stochastic Climate Realizations to Efficiently Explore a Wide Range of Climate Risk to Water Resource Systems. Journal of Water Resources Planning and Management, 0(0): 06016002.

Zion, M.S. et al., 2011. Investigation and Modeling of winter streamflow timing and magnitude under changing climate conditions for the Catskill Mountain region, New York, USA. Hydrological Processes, 25(21): 3289-3301.