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]

Abstract: Evaluating the suitability of data from global climate models (GCMs) for use as input in water supply models is an important step in the larger task of evaluating the effects of climate change on water resources management such as that of water supply operations. The purpose of this paper is to present the process by which GCMs were evaluated and incorporated into the New York City (NYC) water supply's planning activities and to provide conclusions regarding the overall effectiveness of the ranking procedure used in the evaluation. A suite of GCMs participating in Phase 3 of the Coupled Model Intercomparison Project (CMIP3) were evaluated for use in climate change projections in the watersheds of the NYC water supply that provide 90% of the water consumed by NYC. GCM data were aggregated using the seven land-grid points surrounding NYC watersheds, and these data with a daily timestep were evaluated seasonally using probability-based skill scores for various combinations of five meteorological variables (precipitation, average, maximum and minimum temperatures, and wind speed). These are the key variables for the NYC water supply because they affect the timing and magnitude of water, energy, sediment, and nutrient fluxes into the reservoirs as well as in simulating watershed hydrology and reservoir hydrodynamics. We attempted to choose a subset of GCMs based on the average of several skill metrics that compared baseline (20C3M) GCM results to observations. Skill metrics for the study indicate that the skill in simulating the frequency distributions of measured data is highest for temperature and lowest for wind. However, our attempts to identify the best model or subgroup of models were not successful because we found that no single model performs best when considering all of the variables and seasons. **DOI: 10.1061/(ASCE)WR.1943-5452.0001054.** *This work is made available under the terms of the Creative Commons Attribution 4.0 International license, http://creativecommons.org/licenses/by/4.0/.*

Author keywords: Evaluation GCM models; Global climate models (GCMs); Probability-based skill score; Fourth assessment report in Coupled Model Intercomparison Project (AR4, CMIP3); Adaptation; Water supply.

Downloaded from ascelibrary.org by Uppsala Universitetsbibliotek on 07/04/19. Copyright ASCE. For personal use only; all rights reserved.

Introduction

Water utilities are increasingly incorporating climate change into their planning activities using several methodologies. New York City's Department of Environmental Protection (NYCDEP) has undertaken a Climate Change Integrated Modeling Project (CCIMP) to evaluate the potential effects of climate change on New York City's (NYC's) water supply. This project uses a suite of global climate models (GCMs) and an integrated system of watershed and reservoir models (NYCDEP 2013). The watershed and reservoir models require many meteorological variables: precipitation, average, maximum and minimum temperatures, and wind speed, which are referred to in this note as *Ppt*, *Tave*, *Tmax*, *Tmin*, and *Wind*, respectively. These variables are needed as inputs to the models simulating reservoir hydrodynamics, watershed hydrology, and vegetation (Anandhi 2016; Anandhi et al. 2011, 2013, 2016). They affect the timing and magnitude of hydrologic inputs, the fluxes of dissolved and particulate nutrients into the reservoirs, and the reservoir hydrodynamics and mixing. In previous studies, we evaluated a methodology that would rank GCMs based on the accuracy of their historical climate simulations (i.e., baseline or 20C3M) in relation to snow water equivalent simulations that are a component of hydrologic models used by NYCDEP (Anandhi et al. 2011).

The expected impacts of climate change on the NYC water supply will affect both the quality and quantity of water stored in the supply. Water quality issues have at times limited the use of different reservoirs, and the NYCDEP must make operational decisions considering both quality and quantity. The novelty of this study is that we simultaneously evaluate a suite of metrological variables that are needed as inputs for models that affect both reservoir water quality and quantity. This significantly increases the number of meteorological variables that must be considered. We demonstrate the use of skill scores (Johnson and Sharma 2009; Raisanen 2007) for evaluating GCM performance for the complete set of meteorological variables that are needed to force the watershed and reservoir models used in the CCIMP, and to document how NYCDEP has used this methodology as part of the CCIMP. We are not aware of any other water supply that has undertaken such a broad evaluation of GCM performance using the skill score methodology.

Study Region and Data

Our focus is on the Catskill and the Delaware subsystems of the New York City water supply system, which are located west of the Hudson (WOH) River. Together, the WOH watersheds provide 90% of NYC's daily water demand and are the largest unfiltered water supply system in the United States. The system consists of six reservoir watersheds [Cannonsville, Askokan, Nerversink,

[1]Assistant Professor, Biological Systems Engineering and Center for Water Resources, College of Agriculture and Food Sciences, Florida Agricultural and Mechanical Univ., Tallahassee, FL 32307 (corresponding author). Email: anandhi@famu.edu

[2]Senior Research Scientist, Section of Limnology, Dept. of Ecology and Genetics, Uppsala Univ., EBC Norbyvägen 18 D, 75236 Uppsala, Sweden. Email: don.pierson@ebc.uu.se

[3]Professor, Dept. of Geography, Hunter College and CUNY Institute for Sustainable Cities, City Univ. of New York, New York, NY 10065.

Note. This manuscript was submitted on April 14, 2017; approved on October 2, 2018; published online on May 17, 2019. Discussion period open until October 17, 2019; separate discussions must be submitted for individual papers. This technical note is part of the ***Journal of Water Resources Planning and Management***, © ASCE, ISSN 0733-9496.

Downloaded from ascelibrary.org by Uppsala Universitetsbibliotek on 07/04/19. Copyright ASCE. For personal use only; all rights reserved.

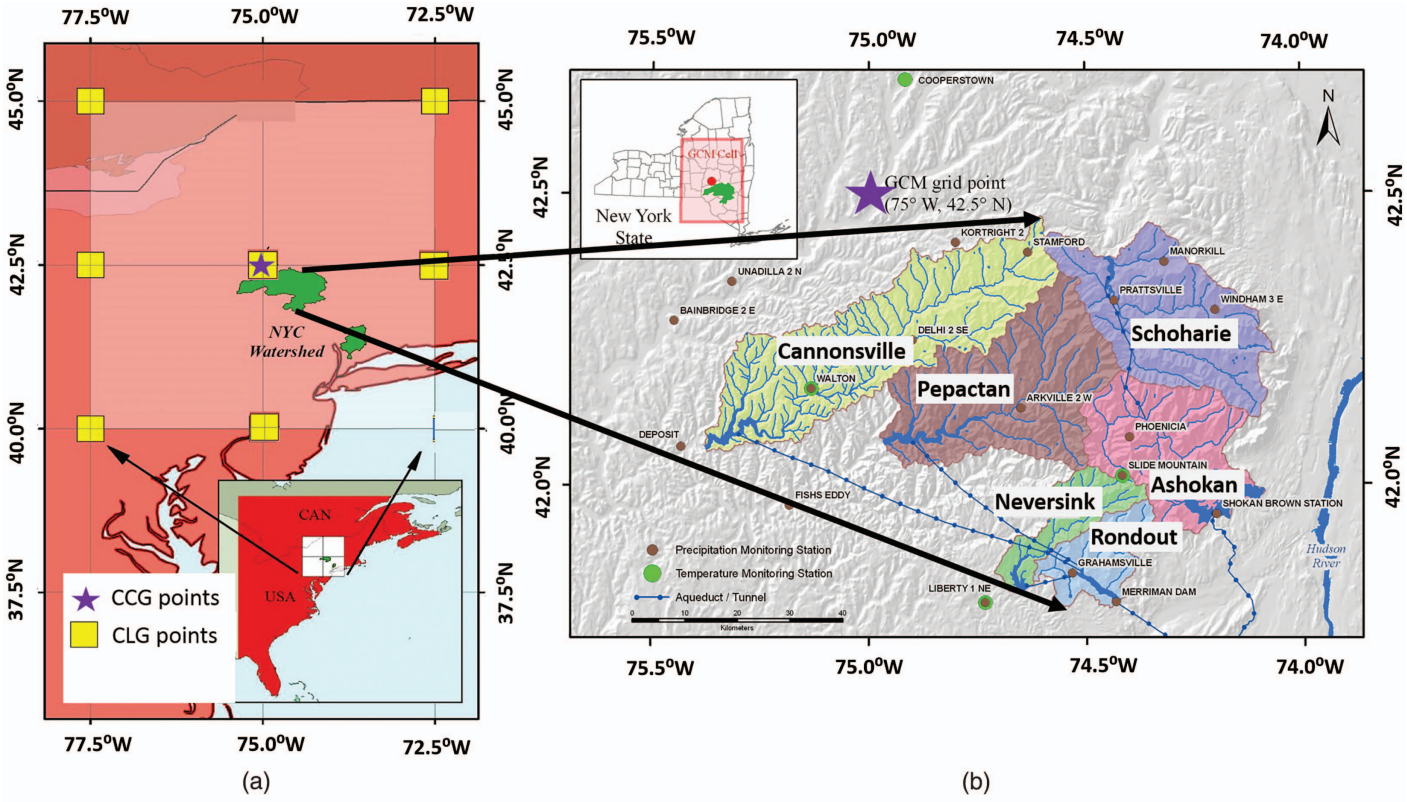

Fig. 1. (Color) (a) Two spatial scales: the seven closest land grids (CLG) used in the main this paper and closest to the grid cell closest to the center of the west of Hudson WOH watershed (CCG); (b) WOH reservoir watersheds; and (c) methodology followed in this study.

Schoharie, Rondout, and Pepacton; see Fig. 1(b)], which encompasses an area of approximately 4,100 km^2.

Historical Measurements

Meteorological measurements of *Ppt*, *Tmax*, *Tmin*, *Tave*, and *Wind* were used in the skill score comparisons described subsequently. Two types of observed data (OD1 and OD2) were used in this study for the skill score comparisons. OD1 is from the daily 1/8-degree gridded reanalysis product produced by Maurer et al. (2002). Data for the five meteorological parameters was taken from seven grid cells surrounding the NYC WOH watershed [closest land grids (CLG) boxes in Fig. 1(a)]. These were then averaged to give a single daily value representative of the entire watershed area. OD2 is based on measurements made at meteorological stations (17 precipitation, and 3 temperature) distributed within the WOH watersheds [locations provided in Fig. 1(b)]. Spatial averages of air temperature and precipitation were used to calculate basin average values for each reservoir watershed (details provided in Supplemental Data). Wind data were collected from a single shore-based station near each reservoir [Fig. 1(b)]. These were also averaged to give a single WOH value.

Baseline (20C3M) GCM Scenarios

Data associated with multiple realizations of the baseline scenario (20C3M) from 20 GCMs were downloaded, and from these data the five meteorological variables were extracted (Table S1). The number of useable GCM realizations ranged between 30 and 45 depending on the climate variables evaluated. The GCMs were from research groups participating in the World Climate Research Programme's (WCRP's) Couple Model Intercomparison Project Phase 3 (CMIP3) multimodel simulations. The grids surrounding the study region were extracted and then interpolated to a common 2.5° grid using bilinear interpolation (yellow boxes in Fig. 1).

Methodology

The methodology followed in this study is briefly described in this section [Fig. 1(c)] and is described in greater detail in Supplemental Data. Basic steps include the following: identifying the purpose of GCM evaluation for the water utility (e.g., estimating changes in water quality); identifying the climate variables that play a role in the processes of concern (e.g., wind in reservoir mixing); determining the spatial scales (e.g., watershed) and temporal scales (e.g., seasonal) of interest; and identifying and estimating the performance metrics (e.g., skill score) to rank the GCM's performance.

In order to quantify the relationship between the observed meteorological data and that obtained in the 20C3M GCM scenarios, metrics of similarity were estimated using both parametric (e.g., mean) and nonparametric (e.g., various percentiles) statistical measures (described in Supplemental Data) as well as skill scores (SS) based on probability distribution functions (PDF). PDF-based SS are calculated from the overlapping area between the PDFs associated with observed measurements and the same meteorological variable obtained from 20C3M GCM scenario. SS is estimated mathematically using equations in Anandhi and Nanjundiah (2015) and ranges between 0 (no overlap of PDFs; GCM derived and observed PDFs are dissimilar) and 1 (complete overlap of PDFs; GCM derived and observed PDFs are same). More details of SS may be obtained from Perkins et al. (2007), Anandhi and Nanjundiah (2015) and Supplemental Data.

Results and Discussion

Comparison of CMIP3 Models to Observed Data

The SS ranged from 0.65 to 0.95 for *Ppt* in all four seasons at CLG scale using OD1 dataset (Fig. 2). The solid red line is the mean while the shaded region represents the variation in PDFs for a

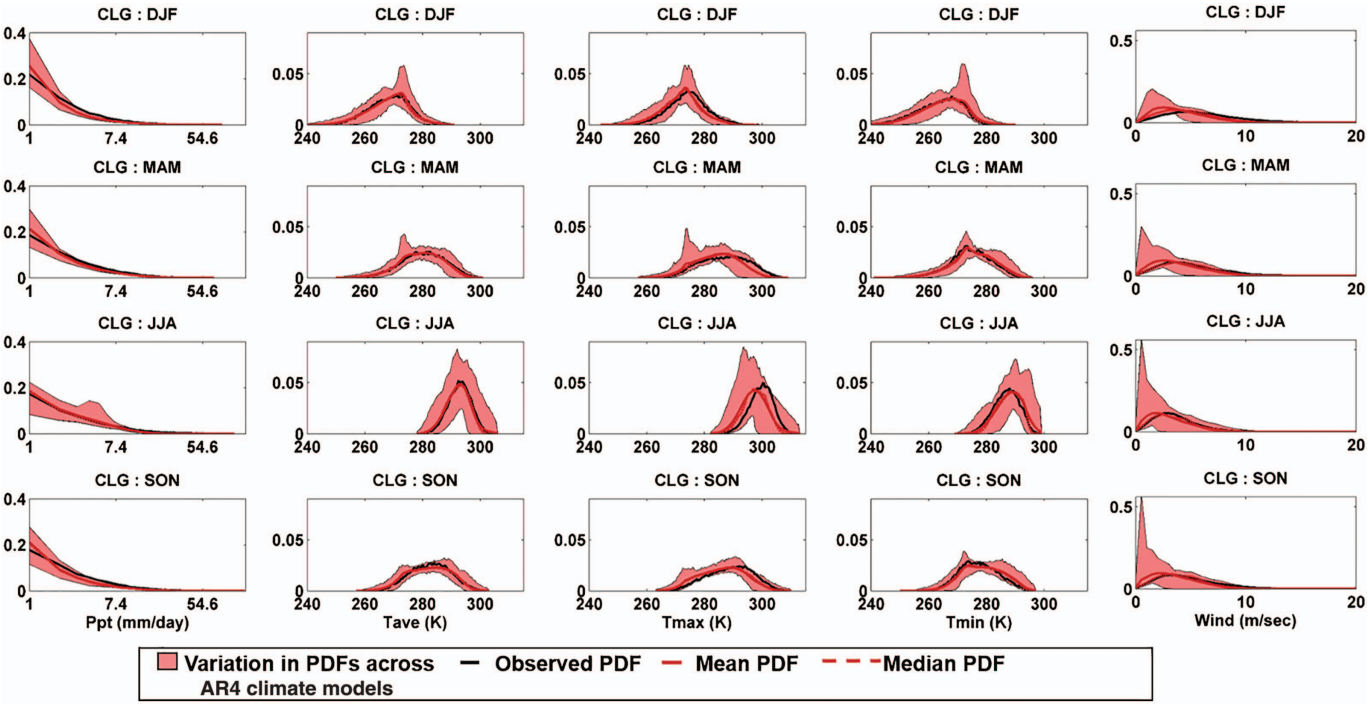

Fig. 2. (Color) Probability distribution functions (PDFs) of daily precipitation (*Ppt*), average temperature (*Tave*), maximum temperature (*Tmax*), minimum temperature (*Tmin*), and wind speed (*Wind*). The *x*-axis for precipitation is in log scale.

J. Water Resour. Plann. Manage., 2019, 145(8): 06019006

Downloaded from ascelibrary.org by Uppsala Universitetsbibliotek on 07/04/19. Copyright ASCE. For personal use only; all rights reserved.

meteorological data simulated by the different AR4 climate models in the CLG region (seven land-grid points surrounding NYC watersheds) for four seasons [December-January-February (DJF), March-April-May (MAM), June-July-August (JJA), and September-October-November (SON)]. In each panel, the black bold line represents the PDF obtained using daily observed data (OD1) for the study region. Differences between the PDFs of observed *Ppt* and that derived from most of the GCMs are larger

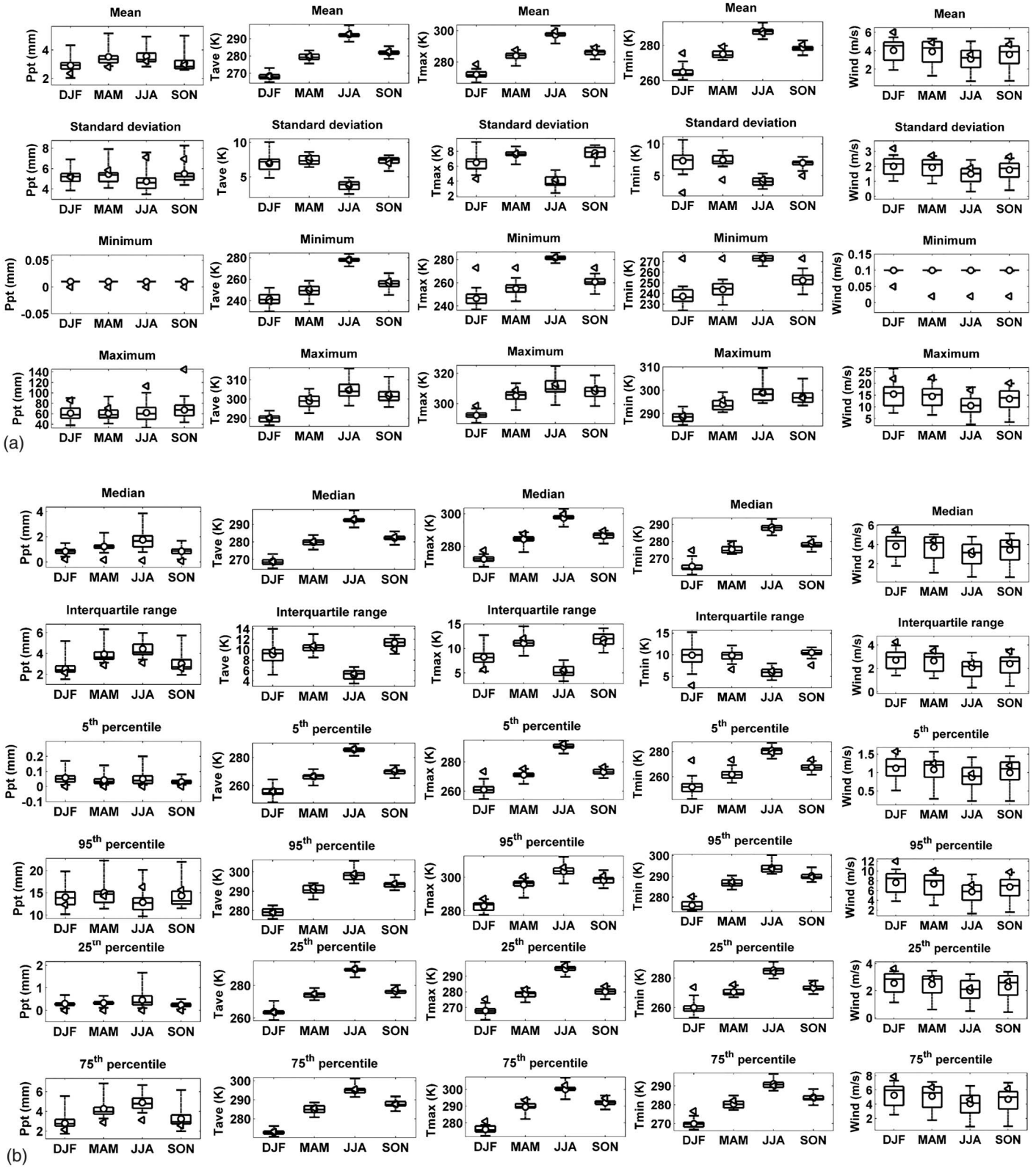

Fig. 3. (a) Statistics, namely mean, standard deviation, minimum, and maximum values of the models and observations for climate variables *Ppt*, *Tave*, *Tmax*, *Tmin*, and *Wind*; and (b) median; interquartile range; and 5th, 25th, 75th, and 95th percentile values of the models and observations for climate variables *Ppt*, *Tave*, *Tmax*, *Tmin*, and *Wind*.

Downloaded from ascelibrary.org by Uppsala Universitetsbibliotek on 07/04/19. Copyright ASCE. For personal use only; all rights reserved.

Downloaded from ascelibrary.org by Uppsala Universitetsbibliotek on 07/04/19. Copyright ASCE. For personal use only; all rights reserved.

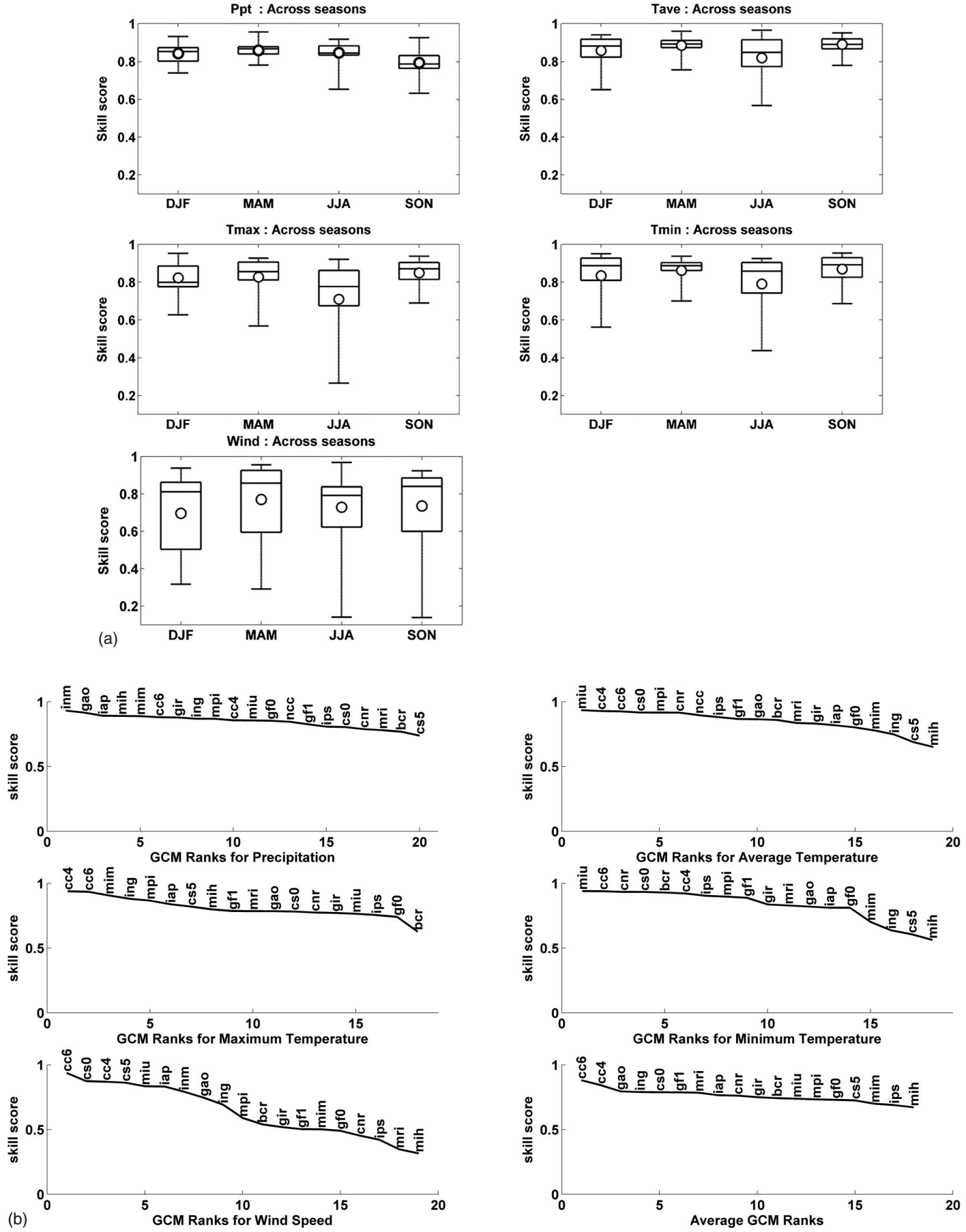

Fig. 4. (a) Summary of skill scores as a function of seasons where box and whisker plots indicate skill scores obtained for all the GCMs including all the seasons for CLG using OD1 dataset; and (b) ranking of GCMs in this study.

during summer and fall seasons (smaller skill scores). The reasons for this may be that the GCM models tend to overestimate the number of small *Ppt* events (1–3 mm/day, Fig. 2) and small to medium *Ppt* events (Fig. 3, minimum; 5th–75th percentiles). Similar overestimation of small events (GCM drizzle) were observed in Australia (Perkins et al. 2007) and India (Anandhi and Nanjundiah 2015). The boxplots in Fig. 3 are interpreted as follows: middle line shows the median value; top and bottom of box show the upper and lower quartiles (i.e., 75th and 25th percentile values); and whiskers show the minimum and maximum model values. The triangle and circle in the boxplots represent the observed and GCM ensemble mean of the statistics for seasons DJF, MAM, JJA, and SON. The *gir* GCM statistic values calculated were excluded from the plots. Box and whisker plots indicate statistics calculated for daily climate variable calculated for the various AR4 climate models across the four seasons, namely DJF, MAM, JJA, and SON for CLG spatial scale (seven land-grid points surrounding NYC watersheds) and OD1 dataset. The overestimation of small events contributes to an overestimation of total precipitation even though the models also tend to underestimate larger events in summer and fall (Figs. 2 and 3). Note that the models underpredicted the median and standard deviation of *Ppt* in all of the seasons (Fig. 2, median).

SS ranged from 0.55 to 0.95 for *Tave*, 0.3 to 0.95 for *Tmax*, and 0.4 to 0.95 for *Tmin* in the four seasons [Fig. 4(a)]. The figure is interpreted as follows: middle line shows the median value; top and bottom of box show the upper and lower quartiles (i.e., the 75th and 25th percentile values); and whiskers show the maximum and minimum percentile skill scores. The outliers are indicated by a "+." The circle in the figure represents the mean skill score each of the seasons (DJF, MAM, JJA, and SON). Among the temperatures, *Tave* was better simulated than *Tmax* and *Tmin*. The reasons for lower SS may be that the GCM models were underestimating the number of cold days and overestimating the number of warm days especially during winter season (*Tmax* and *Tmin* in Fig. 2). However, the largest *temperature* biases—as well as the largest between-model variability—were found in summer (Columns 2–4 in Fig. 2).

SS ranged from 0.2 to 0.95 for *Wind* in the four seasons [Fig. 4(a)]. In most cases, the GCM simulated *Wind* distribution compared unfavorably to the observed distribution (Fig. 2). The reasons for the lower SS is because models overestimated smaller winds (Fig. 3, minimum; Fig. 2, 0–5 m/s) 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. Additionally, the models tended to overestimate the frequency of small events (0–5 m/s) and underestimate the frequency of large events. Similar results were observed for OD2 data set at CCG scale (figure not shown).

GCM Ranks and SS Ranking Procedure (CLG to OD1)

The results of the probability-based SS ranking procedure at the CLG scale using OD1 dataset for all ensemble members of a GCM are summarized as a function of season [Figs. 4(a) and S1] and the SS is arranged in descending order for each variable [Fig. 4(b)]. In the figure, the AR4 climate models are ranked based on average skill scores for spatial scale CLG using OD1 dataset. For a variable and GCM, the average skill score is calculated from the skill scores of different realizations for a GCM and four seasons (DJF, MAM, JJA, and SON) in each realization. The GCM with the highest skill score is given rank 1. While calculating the average ranks, only GCMs that have all five meteorological variables were used. The closeness of the statistical measures of the GCM data to equivalent observations can be seen in Fig. 3. In general, no one model was consistently ranked best by SS for all the meteorological variables (*Ppt*, *Wind*, *Tave*, *Tmax*, and *Tmin*), or during all the seasons (DJF, MAM, JJA, and SON). Overall, the magnitudes of SS did not vary between seasons for *Ppt* and *Wind*, although there was a higher variability in SS during summer for *Ppt* [Fig. 4(a)]. For temperature, there were generally lower magnitudes of SS during summer. Overall, spring had a higher mean/median skill score for all five variables. Fall's mean/median SS 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 (i.e., cc4 and cc6). This can indicate that the skill scores were not due to random or chaotic processes but were in fact related to model formulation. Ensemble average SS showed no clear relationship between SS and three model characteristics (horizontal resolution, convective scheme, and flux correction).

Table 1. Top five GCMs with highest skill score from each meteorological variable observed in WOH watersheds

Meteorological variable	Top five GCMs with highest skill score[a]
Ppt	inm, gao, iap, mih, and mim
Tave	miu, cc4, cc6, cs0, and mpi
Tmax	cc4, cc6, mim, ing, and mpi
Tmin	miu, cc6, cnr, cs0, and bcr
Wind	cc6, cs0, cc4, cs5, and miu

[a]These rankings need not be the same for other regions, evaluation method, and CMIP5 GCMs. The details of the GCMs are available in Table S1.

Overall rankings are in Table 1. The cs5 seemed to have consistently low ranks in the region for *Ppt* and temperature variables. The *gir* had very different statistics (not shown due to being outside the range of figures) compared with the rest of the models for *Ppt*.

Our results show that when GCMs are ranked by skill score and a variety of statistics for different metrological variables, there is no obvious way to choose a subset of models that are clearly superior. First, we did not find certain models as clearly superior; instead, there was a gradual decrease in model skill along a continuum from highest to lowest skill score. Second, different models performed better for different meteorological variables and performance measures. This can greatly complicate choosing a subset of models when 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, then choose a subset from these based on the combined SS rankings that include all needed meteorological variables. For the NYC water supply watershed region and when evaluating multiple metrological parameters, we concluded that using as many GCM datasets as possible was the best strategy. 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. However, we were able to eliminate several GCM data sets that clearly underpreformed. Even though our evaluation was not able to clearly identify GCM models that preformed best for our purposes, we feel that documentation of this methodology is valuable. Results could be different when fewer meteorological parameters are needed, or in other geographical regions where the GCMs may agree to a greater extent.

Summary and Conclusions

The analysis presented in this note leads to several conclusions:

- No single GCM performed well for all the variables considered in the study.

Downloaded from ascelibrary.org by Uppsala Universitetsbibliotek on 07/04/19. Copyright ASCE. For personal use only; all rights reserved.

- The mean and median of all GCM data over the entire time period compares well with the mean and median of all the measured data (OD1 and OD2) for *Ppt* and temperature variables (*Tave*, *Tmax*, and *Tmin*).
- Winds in the region were not well simulated by the GCMs.

Based on the results of this study, one way to choose a subset of GCM datasets is identifying GCMs with the highest average skill scores across all variables. Skill scores can then be used to eliminate the worst-performing models from the ensemble set (e.g., in our case, cs5 and ips). Water quality simulations would then be based on a reduced (but still relatively large) number of GCM models, and the results will be more constrained due to the elimination of the poorly preforming GCMs. A second approach for when computational resources are limiting would be to use the skill scores to choose a smaller subset of models that would likely lead to results that are representative of the study. In our study region, the top five models were cc6, cc4, gao, ing, and cs0.

Several studies in NYCDEP document the use of these results in CCIMP for simulating future changes in water quantity and quality. The second phase of the CCIMP are currently using GCM simulations from CMIP5. Other criteria (such as climate change sensitivity) may be included in the choice of models, but such analysis is beyond the scope of this study. The average ranking we used is just one way to create a single ranking, though considering and weighting the ranking for each variable is probably more informative. Future studies can build on this research by testing the performance and convergence of CMIP5 model datasets to similar ranking procedures.

Acknowledgments

We acknowledge the modeling groups, the Program for Climate Model Diagnosis and Intercomparison (PCMDI), and the WCRP's Working Group on Coupled Modelling (WGCM) for their roles in making available the WCRP CMIP3 multimodel dataset. The New York City Department of Environmental Protection supported this study as part of the CCIMP. This material is based on work partially supported from the USDA-NIFA capacity building Grant No. 2017-38821-26405, Evans-Allen Project, Grant No. 11979180/2016-01711, USDA-NIFA Grant No. 2018-68002-27920, as well as the National Science Foundation under Grant No. 1735235 awarded as part of the National Science Foundation Research Traineeship. The author thanks the three anonymous reviewers, associate editor, and editor for their helpful and constructive comments and suggestions. The support of Ms. N. Ramalingam is also acknowledged.

Supplemental Data

Figs. S1–S6 and Table S1 are available online in the ASCE Library (www.ascelibrary.org).

References

Anandhi, A. 2016. "Growing degree days—Ecosystem indicator for changing diurnal temperatures and their impact on corn growth stages in Kansas." *Ecol. Indic.* 61: 149–158. https://doi.org/10.1016/j.ecolind.2015.08.023.

Anandhi, A., A. Frei, S. M. Pradhanang, M. S. Zion, D. C. Pierson, and E. M. Schneiderman. 2011. "AR4 climate model performance in simulating snow water equivalent over Catskill mountain watersheds, New York, USA." *Hydrol. Processes* 25 (21): 3302–3311. https://doi.org/10.1002/hyp.8230.

Anandhi, A., S. Hutchinson, J. Harrington, V. Rahmani, M. B. Kirkhamd, and C. Rice. 2016. "Changes in spatial and temporal trends in wet, dry, warm and cold spell length or duration indices in Kansas, USA." *Int. J. Climatol.* 36 (12): 4085–4101. https://doi.org/10.1002/joc.4619.

Anandhi, A., and R. S. Nanjundiah. 2015. "Performance evaluation of AR4 climate models in simulating daily precipitation over the Indian region using skill scores." *Theor. Appl. Climatol.* 119 (3–4): 551–566. https://doi.org/10.1007/s00704-013-1043-5.

Anandhi, A., M. S. Zion, P. H. Gowda, D. C. Pierson, D. Lounsbury, and A. Frei. 2013. "Past and future changes in frost day indices in Catskill mountain region of New York." *Hydrol. Processes* 27 (21): 3094–3104. https://doi.org/10.1002/hyp.9937.

Johnson, F., and A. Sharma. 2009. "Measurement of GCM skill in predicting variables relevant for hydroclimatological assessments." *J. Clim.* 22: 4373–4382. https://doi.org/10.1175/2009JCLI2681.1.

Maurer, E., A. Wood, J. Adam, D. Lettenmaier, and B. Nijssen. 2002. "A long-term hydrologically based dataset of land surface fluxes and states for the conterminous United States." *J. Clim.* 15 (22): 3237–3251. https://doi.org/10.1175/1520-0442(2002)015<3237:ALTHBD>2.0.CO;2.

NYCDEP (New York City Department of Environmental Protection). 2013. *Climate change integrated modeling project: Phase I assessment of impacts on the New York City water supply.* Kingston, NY: Division of Watershed Water Quality Science and Research Bureau of Water Supply, NYCDEP.

Perkins, S. E., A. J. Pitman, N. J. Holbrook, and J. McAneney. 2007. "Evaluation of the AR4 climate models' simulated daily maximum temperature, minimum temperature, and precipitation over Australia using probability density functions." *J. Clim.* 20 (17): 4356–4376. https://doi.org/10.1175/JCLI4253.1.

Raisanen, J. 2007. "How reliable are climate models?" *Tellus* 59 (1): 2–29. https://doi.org/10.1111/j.1600-0870.2006.00211.x.

Downloaded from ascelibrary.org by Uppsala Universitetsbibliotek on 07/04/19. Copyright ASCE. For personal use only; all rights reserved.