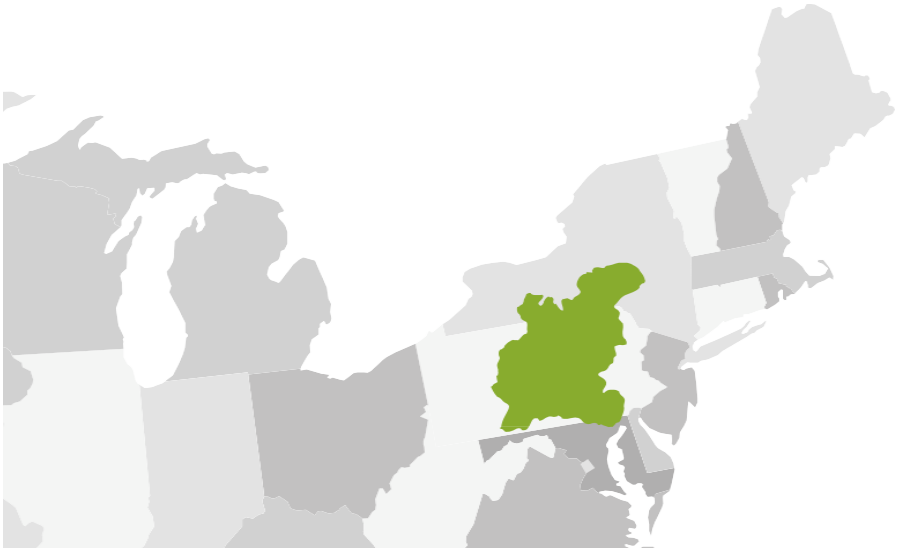
SEPTEMBER 2020

Project Hyperion - Narrative Case Study Report: *Susquehanna*

Kripa Jagannathan (kajagannathan@lbl.gov) and **Andrew Jones** (adjones@lbl.gov)

Lawrence Berkeley Laboratory



Contributors:

- Paul Ullrich, University of California, Davis (Project Team Leader)
- Bruce Riordan, Climate Readiness Institute (Engagement Facilitator)
- Abdolhossain Liaghat, Pennsylvania Department of Environmental Protection
- Abhishekh Srivastava & Richard Grotjahn, University of California, Davis
- Alan Rhoades, Lawrence Berkeley Laboratory
- Chaopeng Shen, Wen-Ping Tsai & Colin Zarzycki, Pennsylvania State University
- Gary Shenk, Chesapeake Bay Program
- John Balay, Susquehanna River Basin Commission
- Kevin Reed, Stony Brook University
- Simon Wang & Binod Pokharel, Utah State University
- Smitha Buddhavarapu, Lawrence Berkeley Lab

Contents

List of Tables

List of Figures

Introduction

This narrative case study report is a synthesis of key discussions and preliminary scientific results for the Susquehanna region, undertaken as part of the Hyperion project (2016-19). Project Hyperion (now continuing as the HyperFACETS project) is a basic science project that aims to advance climate modelling by evaluating regional climate datasets for decision-relevant metrics. While there has been an explosive growth in the number of regional climate datasets available to users, there is limited understanding of the credibility and suitability of these datasets for use in different management decisions. Hyperion aims to address this need by developing comprehensive assessment capabilities to evaluate the credibility of regional climate datasets, understand the processes that contribute to model biases, and improve the ability of models to predict management relevant outcomes.

Since decision-relevance is a core motivation for the project, Hyperion is designed on the principles of co-production. The project brings together scientists from nine research institutions and managers from twelve water agencies in four watersheds: Sacramento/San Joaquin, Colorado Headwaters, South Florida, and Susquehanna. The project structure explicitly allows for both the groups to co-develop the science plan and research questions, in addition to co-producing the science itself. The scientists include atmospheric and earth system scientists as well as hydrologists. The water managers, depending on the agency, have functions including planning, operating and managing water quality, water supply, stormwater management, flood control, and water infrastructure design.

This narrative report provides an overview of the co-production process in Hyperion (Chapter 1), the regional hydro-climatic context and challenges (Chapter 2), broad climate information needs of water management agencies (Chapter 3), and short summaries of the key scientific activities undertaken for the region (Chapter 4). This information is based on the project's co-production engagements and preliminary scientific results. Some of the preliminary results may be updated or refined as they go through the peer-review process. While this report is based on the perspectives of water management agencies that were part of Hyperion, we hope that the insights and methodologies that were developed are broadly applicable to other agencies in the region as well.

1. Co-production in Hyperion

In Hyperion, as far as possible, the research questions, approaches and results are co-produced through regular structured and unstructured engagements between scientists and managers (Figure 1). Structured engagement methods include workshops, remote and in-person focus-group discussions, and quarterly project update calls. There are also continual less-structured, informal conversations over telephone calls and emails.

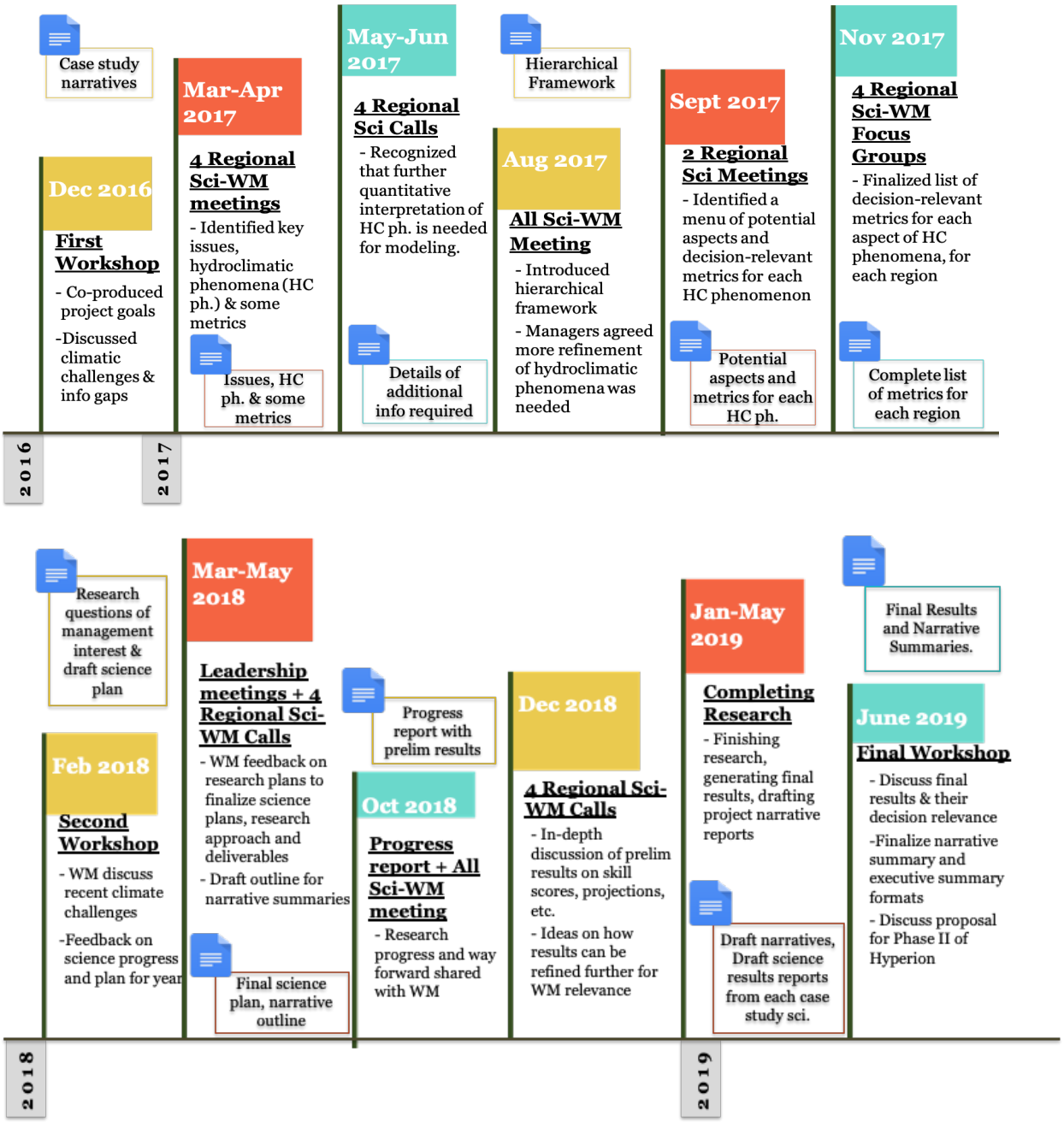


Figure 1: Co-production process and timeline
Summarizes key engagement activities along with important outcomes at each stage (depicted by the blue document icon). 'Sci' refers to Scientists, 'WM' refers to Water Manager and 'HC ph.' refers to Hydroclimatic Phenomena.

2. Regional hydro-climatic context & challenges

The Susquehanna river basin (SRB) is spread over parts of New York, Pennsylvania, and Maryland. The river empties into the Chesapeake Bay and provides more than one-half of the freshwater flowing into it. The basin provides water resources for domestic & municipal uses, power production, agricultural & industrial activities, as well as for recreational & environmental uses. The basin also serves as a passage for several migratory fish species.

SRB is one of the most flood-prone regions in the country; flooding (flash and riverine) and related stormwater management are key challenges for the region. In addition, some portions of the basin have also experienced significant droughts (the basin had 5 droughts in the past 3 decades), and therefore maintaining a sustainable water supply is also an issue. Furthermore, because some of the biggest water users (i.e. power plants) are on the river, there exist water sustainability concerns that need to be addressed. Water quality, especially in the Chesapeake Bay, is of high importance for the health of the Bay. Nutrient export to the Bay and salt-water intrusion, and their impact on water quality, are also important challenges. Streamflow and temperature are important hydrologic metrics which in-turn depend on metrics such as magnitude of rainfall, runoff, etc. Baseflow and stream temperature responses to drought are also of great importance.

Most of the planning studies in SRB such as state water plans, flood control planning, drought assessments, water supply & availability studies, and other hydrological monitoring and modeling studies, do not typically use climate projections and are based mainly on existing and historical conditions. However, this is beginning to change. One example is the case of water quality assessments, wherein the EPA mandates that climate change be considered in Total Maximum Daily Load (TMDL) analysis. Some of the system impact models relating to water quality analysis, take into account climate model data, as one of the input parameters. Climate data is also used in design and management of stormwater ponds and fish passages. Hydroelectric dams and water supply agencies are some of the biggest users of climate information. In addition, other local-level agencies have shown interest in using climate change projections (even if they are currently not using such information).

Some key information gaps for the region are that there is a need to quantify uncertainty and for higher spatial resolution in data especially for planning at the local-level. Future condition flow duration curves and statistics (e.g. projected 7-day, 10-year low flow) can be important tools that are currently not available. Projections of future water availability and of changes in extreme event occurrences (such as droughts) are lacking. Accurately predicting the variability in intensity and ‘flashiness’ of rainfall is also a constraint. In existing hydrologic models used at the Bay program, the choice of Evapotranspiration (ET) methods has a large impact on simulated sediment load, but it’s not easy to know which formula for computing ET is the most appropriate. Land use responses to changing climate signals are also a key information gap.

3. Climate information needs for water management

3.1. Overview

From the previous section, and based on focus group discussions in the project, droughts, water supply, and flooding were identified as key climate-related issues in the Susquehanna river basin. In addition, water quality was also identified as an important issue. Therefore, several peak as well as low streamflow metrics emerged as decision relevant. Since a portion of the Susquehanna region is heavily influenced by snow, snowpack metrics are also of importance. In addition, several extreme high and low precipitation metrics were identified as important. The issue of stationarity in current planning was brought up many times and recognized as a constraint that the managers are dealing with. In this context, current planning does not always incorporate information on variability/standard deviation of different runoff/precipitation metrics.

In terms of spatial scales of relevance, for several metrics, stakeholders suggested that basin scale spatial resolution may be sufficient. Currently several metrics are measured at one candidate location (rain gauge or well) per county. Depending on the type of decision, the temporal scale of planning with climate change information may range from 10 to 50 to even 100 years.

3.2. List of decision-relevant metrics and their importance

In order for science to be actionable, resource managers need information on decision-relevant climatic metrics. Therefore, one of the first goals of Hyperion was to co-produce the decision-relevant metrics for different management decisions in each of the case study regions. From the water managers' perspective, such metrics quantitatively describe climatic phenomena that are directly related to practical management problems; changes in these quantities would necessitate shifts in water infrastructure planning and operations. From the scientists' perspective, these metrics can be used to test model fidelity for decision-relevant phenomena and hence push model development and scientific inquiry in more use-inspired directions. Table 1 represents the decision relevant metrics, along with their potential importance, that were developed through iterative engagements between Dec 2016 to Nov 2017. This table is referred from the published journal article titled "The making of a metric: Co-producing decision-relevant climate science" by Jagannathan, Jones and Ray.[1]

[1] Jagannathan, K., A. D. Jones, and I. Ray, The making of a metric: Co-producing decision-relevant climate science. *Bull. Amer. Meteor. Soc.*, doi: https://doi.org/10.1175/BAMS-D-19-0296.1.

Table 1: Examples of decision-relevant metrics for each region.
The table highlights management issues, hydroclimatic phenomena, aspect of phenomena and then each decision-relevant metric. The last column also describes some of the potential decisions or uses for these metrics that were identified by the case study water managers.

Issue	Hydroclimatic Phenomenon	Aspect of Phenomenon	Decision-relevant Metric	Decision/Use
Water Supply {Quality}	Streamflow	Peak flow	Flows that exceed 250,000 cfs or 400,000 cfs scour threshold for Conowingo Reservoir	Water quality management, specifically sediment and nutrient management for lower Susquehanna River and Chesapeake Bay.
Water Supply	Streamflow	Peak flow	10- year frequency 3-year duration high flows for Oct-March	Water quality management in terms of monitoring Chesapeake Bay water quality standards.
Floods	Streamflow	Peak flow	Probable Maximum Flood	Dam safety and flood risk management considerations.
Water Supply	Streamflow	Low flow	7-day,10-year low flow	Water quality management in terms of wastewater assimilation standards for discharge permits, and water supply planning in terms of passby flows or conservation releases for water withdrawal permits.
Water Supply and Droughts	Streamflow	Low flow	Monthly 95th percent exceedance (P95), P90, P85, P80, and P75 flows	Water supply planning for passby flows, conservation releases, and low flow augmentation associated with water use permits. Drought conditions monitoring and issuing drought watch, warning, and emergency declarations
Water Supply	Streamflow	Mean flow	Mean annual flow and harmonic mean flow	Water supply planning for passby flows and conservation releases associated with water withdrawal permits. Water quality management for calculating design flows for effluent limitations based on water quality criteria.

Issue	Hydroclimatic Phenomenon	Aspect of Phenomenon	Decision-relevant Metric	Decision/Use
Water Supply and Droughts	Streamflow	Low flow	July through November monthly median and P95 flows.	Water supply planning for water use and availability as well as consumptive use mitigation operations.
Water Supply	Streamflow	Monthly Streamflow	Percentage of annual streamflow occurring in each month	Water supply planning considering use and availability including monthly variable demands and in-stream flow needs.
Water Supply	Streamflow	Variability of Streamflow	Standard deviation of monthly or annual flows	Water supply planning considering use and availability including monthly variable demands and in-stream flow needs.
Floods	Rainfall	Extreme Rainfall	Intensity Duration Frequency (IDF) curves, (Generic 1-day, 2-day, 3-day up to 7-day duration, 2-year, 5-year, 10-year up to 100-year storms)	Flood risk management, and stormwater management and design criteria, including roadway drainage.
Water Supply and Droughts	Rainfall	Cumulative rainfall	30-,60-,90-,120-day cumulative rainfall and departure from average	Drought monitoring and declarations.
Water Supply and Floods	Rainfall	Annual cycle of Rainfall	Rainfall distribution, Focusing on shifts in wet and dry seasons	Flood risk management and water supply planning.
Water Supply and Floods	Snowpack	Annual cycle of snow accumulation and melt	SWE triangle, Focusing on peak date, accumulation rate, % of snow accumulation in different months of snow season	Water supply planning in terms of water use and availability, flood control reservoir operations, and calibration of hydrologic models.

Issue	Hydroclimatic Phenomenon	Aspect of Phenomenon	Decision-relevant Metric	Decision/Use
Water Supply	Snowpack	Monthly snowpack	Monthly water availability from snow	Water supply planning in terms of water use and availability and calibration of hydrologic models.
Floods	Snowmelt	Peak flow	Frequency of rain-on-snow events and magnitude of associated runoff and streamflow	Flood control reservoir operations and water quality management, specifically sediment and nutrient management for Chesapeake Bay.
Water Supply	Rainfall	Low precipitation	Precipitation anomalies i.e.. Rolling 90-day mean precipitation and its departures from normal.	Drought monitoring and declarations
Water Supply	Evapotranspiration (ET)	Monthly ET	March - October monthly ET	Consumptive use and water budget evaluations.

4. Key scientific activities and results from Hyperion

From the above long list of decision-relevant metrics, project Hyperion's managers and scientists collectively developed case study science plans that identified a shorter list of scientific activities and metrics that will be a focus of the project (Table A1 in the Appendix). The rest of this section presents a narrative description of three of these short-listed scientific activities: precipitation Intensity Duration Frequency (IDF) curves, streamflow modeling and extra tropical cyclones. We summarize the key motivation, methods, results and limitations from each of the three scientific activities.

4.1. Precipitation extremes and IDF curves

Summary

- This work analyzes how different climate models vary in their IDF estimates for the past and the future. It also proposes a framework that allows for examining IDF estimates for longer return periods, where the data sample size can be a limitation.

- The study finds that there is considerable variability within and across models in both predicting historical IDFs as well as in IDF projections of the future.
- A method is proposed that employs pooling of model data based upon historical performances of models. The models selected for pooling are bias corrected and then used for estimation of non-stationary IDF curves. The proposed method reduces estimation uncertainty due to enhanced sample size.

4.1.1. Background and Methods

IDF or Intensity Duration Frequency estimates are used for planning and management of extreme precipitation events. The curves specify the magnitude (i.e., intensity) of precipitation events across a range of durations and return periods (i.e., frequencies). These estimates provide information to support a wide variety of civil activities such as designing flood protection structures and urban drainage systems. However, there are significant uncertainties and variability in climate models' predictions of extreme precipitation. In the case of IDFs, the estimation uncertainty increases as one considers longer return periods since larger sample sizes are needed to estimate rarer events (e.g., assessing IDFs for 100-yr return period requires at least 100 years of data). Not many studies have critically examined the variability among different models in predictions of IDFs. In addition, the few studies that provide projections of IDFs for the future, either take a mean or median of IDF estimates from different models which may not address the issue of uncertainty due to small sample sizes and variability across models. Therefore, this research also proposes a new methodology that can help to reduce some of the issues associated with limited sample size for IDF estimations. The underlying hypothesis for this work is that due to data (sample size) limitations for IDF estimations, and uncertainties, new and novel methods of combining model data of IDFs may be needed to better evaluate this metric. This research focuses on the following key research questions:

1. ***How do climate models vary in their IDF estimates of the past and future? What are the differential capabilities of climate models in predicting historical IDFs?***
2. ***Do models show a statistically significant change in IDF estimates in future time periods as compared to historical?***
3. ***What framework allows for analyzing changes in IDF estimates in the wake of sample size limitation, natural variability across space, and variability across models?***

IDF estimates were computed for historical (1956-2005) and RCP8.5 simulations (2049-2098) of the NA-CORDEX models. To provide station-wise results, model data was interpolated to station locations using nearest neighbor interpolation. The reference weather station data was obtained from NOAA Atlas 14, 24-hour precipitation data from GHCN archive. Every station that had 50 years data between 1950-2005 was included. The 12 NA-CORDEX models (with 0.25x0.25deg resolution) that were evaluated are: CanESM2.CanRCM4 (A), CanESM2.CRCM5-OUR (B), CanESM2.CRCM5-UQAM (C), GFDL-ESM2M.CRCM5-OUR (D), GFDL-ESM2M.RegCM4 (E), GFDL-ESM2M.WRF (F), HadGEM2-ES.RegCM4 (G), HadGEM2-ES.WRF (H), MPI-ESM-LR.CRCM5-OUR (I), MPI-ESM-LR.CRCM5-UQAM (J), MPI-ESM-

LR.RegCM4 (K), and MPI-ESM-LR.WRF (L). Other datasets such as Variable Resolution CESM and LOCA downscaled data were also analyzed but are not presented here for brevity.

The IDF estimations are based on univariate extreme value analysis that uses the method of maximized likelihood estimation. The generalized extreme value (GEV) distribution was then fitted to the data in a non-stationary framework. IDF estimates were computed from a sample size of 50 years, for 24-hour duration events at; 2, 5,10, 25, 50, and 100-year return periods. Historical and future IDF estimates from different models were computed for each of the weather stations, and the resultant intra and inter-model variability in IDF results was examined. Since the 50-year sample size was limiting (especially for assessing longer return periods), models were then bias-corrected using quantile matching so that all models have the same historical distribution as the observations. The models (ones that accurately capture space and time variability of select precipitation metrics) were pooled together to develop a long time-series of data (i.e. if 5 models with 20 years of data can be pooled, it can yield 100 years of data).

4.1.2. Key Results

Figure 2 shows the results of a comparative skill evaluation of models in predicting historical IDFs for a 24-hour duration storm at one weather station. This analysis was done for over 50 weather stations across the region (shown in the first panel of Figure 2). The results show that there is considerable variability across models in predicting both the historical and future IDFs. Also, the estimation uncertainty is large in models. For example, model A estimates historical rainfall intensity for 24-hour duration storms of different return periods between 2 and 7 inches, whereas model L predicts the same between 2 and 3 inches. The estimation uncertainty particularly increases with higher return periods. Apart from other factors this could be due to a smaller sample size. Figure 2 also gives IDF projections for the weather station from different models. The figure suggests that there may be an increase in IDF estimates in the future, but this change (between historical and projected IDF for each model) may not be statistically significant. Further, there is also a large variability between the projections of IDFs in different models (i.e. both intra and inter model variability in projected IDFs is high), and hence a statistically significant change signal is not seen.

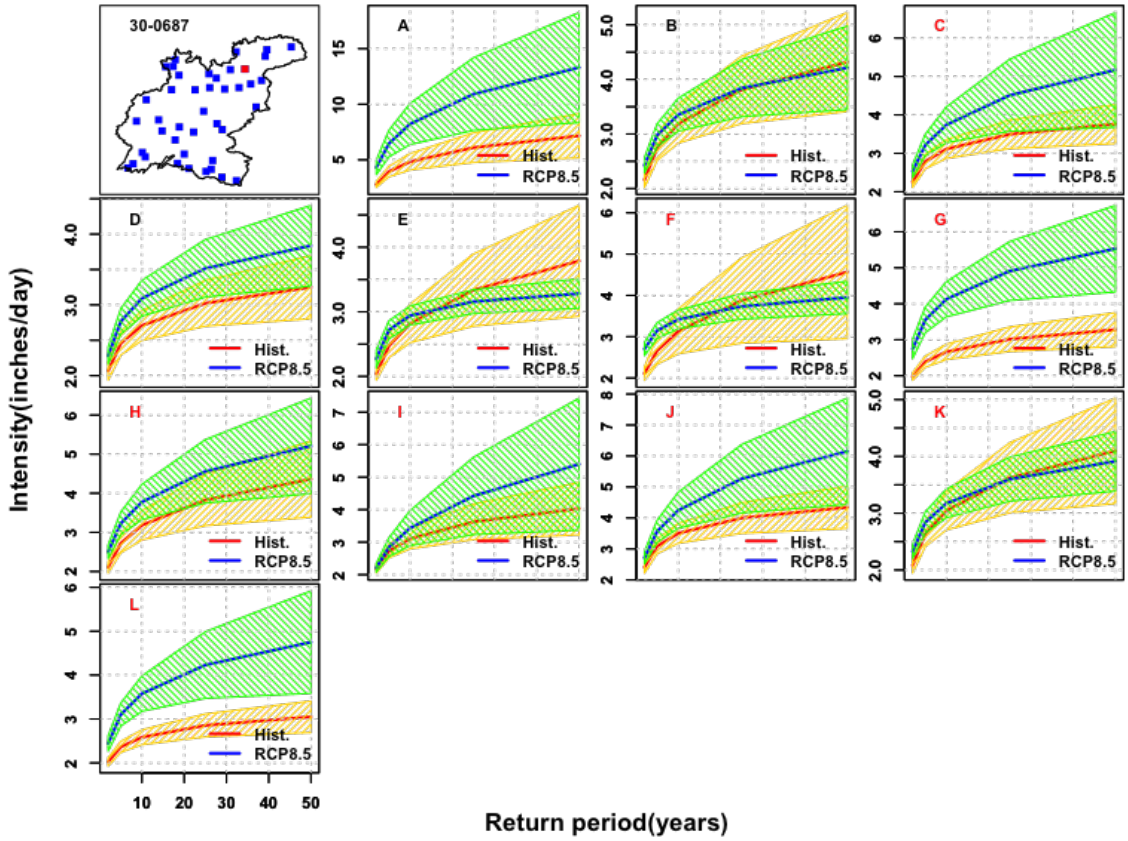


Figure 2: NA CORDEX models' predictions precipitation intensity estimates for a 24-hour duration storm
The first panel shows weather stations and their associated latitude longitude. Panels labelled A-L represent the different NA CORDEX models' predictions of historical and future precipitation intensity estimates for a 24-hour duration storm. Red is the historical IDF estimate, and the yellow shaded area is the 95% confidence interval (CI) around it. Similarly, blue is the future IDF estimate and the green area is the 95% CI around it.

In order to overcome the limitation of a small sample size, a methodology for pooling different models' data to create a large sample size was developed. This methodology required bias correction of data. Bias corrected and pooled results are presented in Figure 3. From the figure we can gather that bias correction in this case reduces some of the inter-model variability. Also, pooling model data reduces some of the estimation uncertainty by increasing the precision of the results due to a larger sample size (as compared to using individual models or taking the median results). This pooling enables the detection of a statistically significant change or IDF estimates is seen between historical and RCP 8.5, showing that when the models are combined together the changed signal is clearer. This approach overcomes the small sample size limitations, thereby providing a clearer picture of the change in IDF estimates that may be expected in the future.

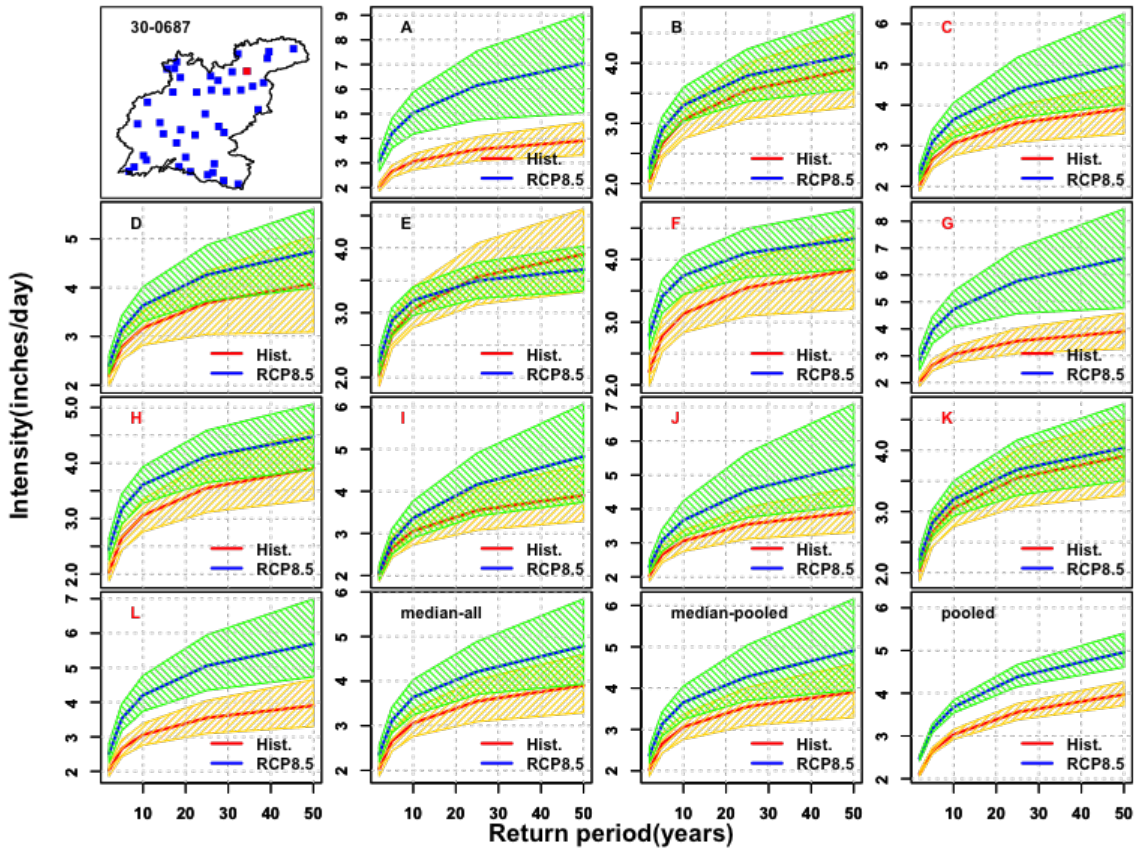


Figure 3: 24-hr IDF estimates in the bias-corrected historical and future simulations.
The red curve and yellow shading indicate IDF estimates and corresponding 90% confidence interval in the bias-corrected historical simulation. The blue curve and green shading indicate IDF estimates and corresponding 90% confidence interval in the bias-corrected RCP8.5 simulation. The reader should note that historical IDF estimates and their 90% confidence intervals are the same in all models. The red curve with orange shading varies between panels solely due to changes in the vertical scale. The ‘’median-all" panel shows the median of IDF estimates from all models, while the “median-pooled" shows the median of the IDF estimates from models that are used for pooling. “Pooled" shows the IDF estimates computed from pooling of better performing models. Models that are used for pooling are shown in red letters in the top left corner of the figures. The X-axis indicates return periods in years and the Y-axis indicates intensity in inches/day.

Figure 4 presents results for all weather stations where only models with good skill score ((C), (F), (G), (H), (I), (J), (K) and (L)) for the annual maximum precipitation (AMP) metric were pooled. This figure again showcases that a statistically significant change between historical and RCP 8.5 is seen when the models are pooled together to create a large sample of data. With this pooling method, almost all-weather stations in the region show an increase in intensity of precipitation for the 24-hour duration storm for different recurrence intervals.

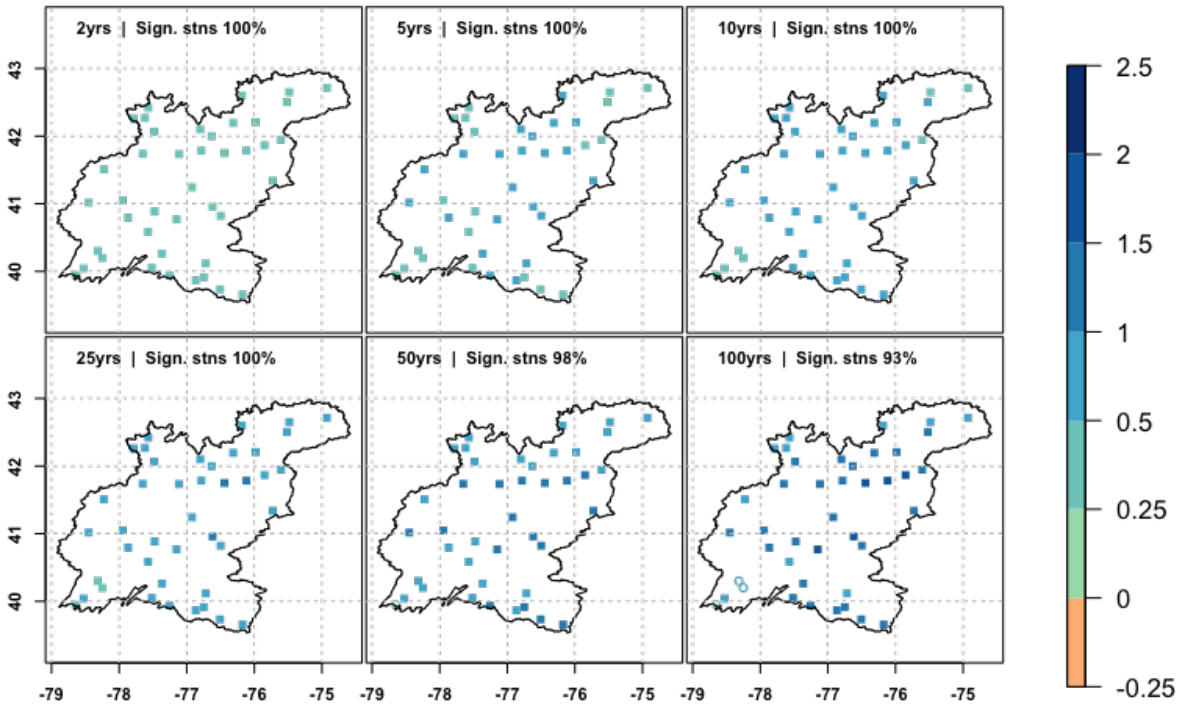


Figure 4: Change in 24-hr precipitation for 2, 5, 10, 25, 5 0& 100 year return periods from pooled models. The differences that are significant at the 90% significance level are shown as solid squares and those not significant at 90% are shown as blank circles. The significance is computed using the z-statistic as defined in section 3.2.3 of Srivastava et al. 2019. Units are in inches/day. Significant stations shows the percentage of stations at which the differences are significant. Pooled models: (C), (F), (G), (H), (I), (J), (K) and (L).

4.1.3. Discussion and Conclusions

Considering widespread variability across models, using a multi-model estimate of IDF seems to be a better option than relying on any single model. The IDF estimates based upon bias-corrected and pooled model data offer a larger sample size that enables the detection of significant increases in future precipitation estimates at more stations than any other method. This method can be applicable to any region or spatial scale, even where models do not agree well with each other and are data limited.

Although this work improves on current capabilities for assessing IDF estimates across different models, it is to be noted that the pooling method increases precision but not accuracy of results. Hence, alternate methods for evaluating accuracy of model predictions may be needed. One such method that has been pursued by this research team is examining model skill for other extreme precipitation indices. This work has assessed the skill of simulations of observed precipitation indices (P-indices) in the historical runs of regional climate models in the NA-CORDEX program. Some of the results from this work are presented in the Appendix (Figures A5-A8). Further work on the topic must also focus on understanding the causes for models performing good/ bad (resolution, dynamics, convection, parameterization). Continual refinement of the multi-model approach to estimate future precipitation changes/ IDF estimates is also needed.

4.2. Streamflow

Summary:

- This work examines the skill of streamflow simulations from a hydrological model, CLM-PAWS, forced by (a) precipitation from a variable resolution climate model versus (b) precipitation from observed gridded reanalysis data.
- Overall, streamflow simulations from CLM-PAWS (in terms of streamflow probability density function, median flow, and 33-percentile flow) matched well with observed USGS gauge data, and the variable resolution CESM model forced hydrologic model is as skillful as gridded climate reanalysis data forced hydrologic model.
- However, both the simulations underestimated peak flows, suggesting that neither of the forcings were able to fully capture the extreme precipitation events, indicating potential biases in the hydrological model.

4.2.1. Background and Methods

Many water management decisions such as water supply or dam operations, are based on estimates of streamflow. Therefore, understanding how well different models and datasets are able to capture streamflow, is important especially in the face of climate change. For streamflow estimations, there are two types of models in play: climate models that provide precipitation data, and hydrological models that convert the precipitation into streamflow. There are biases and uncertainties within each of these models that can impact the streamflow estimations. Therefore, exploring the streamflow estimations from hydrologic models driven by different observation and model data can be a useful way to understand how errors or biases propagate (or do not propagate) from precipitation to streamflow. This can also help better assess the value and limitations of climate data driven hydrological models, and point to areas for improvement in the models. This analysis focuses on understanding how climate simulation errors propagate from precipitation to streamflow, and asks the question:

1. ***What is the skill of streamflow simulations from a hydrologic model (CLM-PAWS) forced by (a) precipitation from a variable resolution climate model versus (b) precipitation from an observation-based gridded reanalysis dataset, as compared to historically observed streamflow data?***

In order to examine this, the Perkin skill score between observed (USGS stream gage) and simulated (CLM-PAWS, Process-based Adaptive Watershed Simulator coupled to the Community Land Model) streamflow was computed. Two simulated streamflows were used: the first was forced with NLDAS precipitation data (which is a reanalysis dataset based on observations) and the other simulation was forced with VR-CESM precipitation data (VR-CESM is a variable resolution climate model with a high resolution for the study region). The Perkin skill score examines the difference in the probability distribution function (PDF) of the two datasets. The Perkin skill score varies between 0-1 and is 1 for identical distributions and 0 for completely different distributions. A similar skill score was also employed for the precipitation

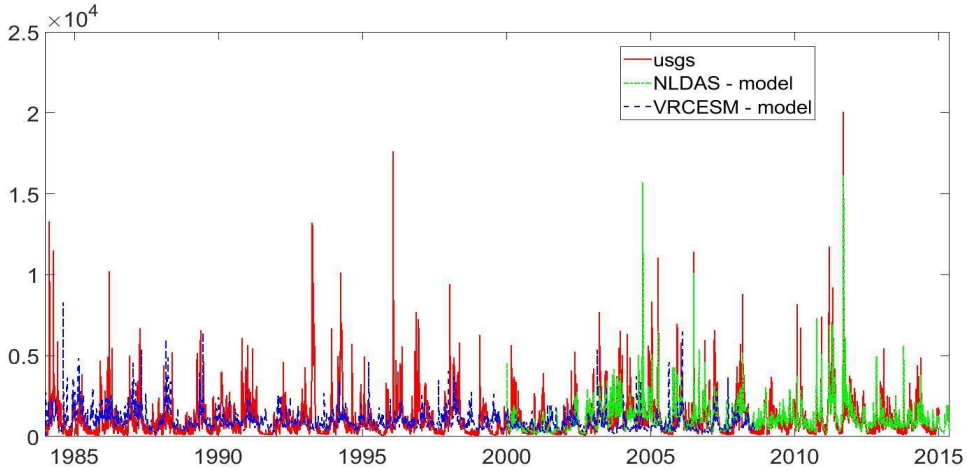
evaluation. The historical period for the USGS streamflow data was from 1984-2017, NLDAS forced streamflow simulations were for 2000-2017, and VR-CESM forced streamflow was from 1984-2014.

4.2.2. Key Results

The Perkin skill score computed for three sub-basins in the Susquehanna River Basin were all close to 1, showing that the PDF, or shape of the distribution of streamflow simulations from CLM-PAWS (with both NLDAS and VR-CESM forcings), matched well with the observed USGS data. There was also not a lot of difference between the scores for CLM-PAWS-NLDAS and CLM-PAWS-VR-CESM showing that the skill of the model forced by the reanalysis and variable resolution model data are similar. This suggests that in this region, the variable resolution CESM model forced hydrologic model is as skillful as the gridded climate reanalysis data forced hydrologic model, in capturing the distribution function of streamflow. Refer to Table 2 and Figure 5 below. It is important to note that although the models were skilled in getting the probability density function of the streamflow, the figure also shows that both these simulations still underestimate peak flows, suggesting that neither of the forcings were able to fully capture the extreme precipitation events. Since the underestimation was observed in both the reanalysis and model-forced simulations, it is possible that this bias arose from the hydrological model. Therefore, the results further suggest that CLM-PAWS may have limitations in terms of expressing the streamflow impacts of extreme precipitation.

Table 2: Perkin's skill score (0-1) for streamflow 2000-2014 for the different sub-basins

Sub-basin	NLDAS	VR-CESM	log10(NLDAS)	log10(VR-CESM)
Susquehanna-sub1	0.96	0.95	0.77	0.78
Susquehanna-sub2	0.93	0.92	0.83	0.85



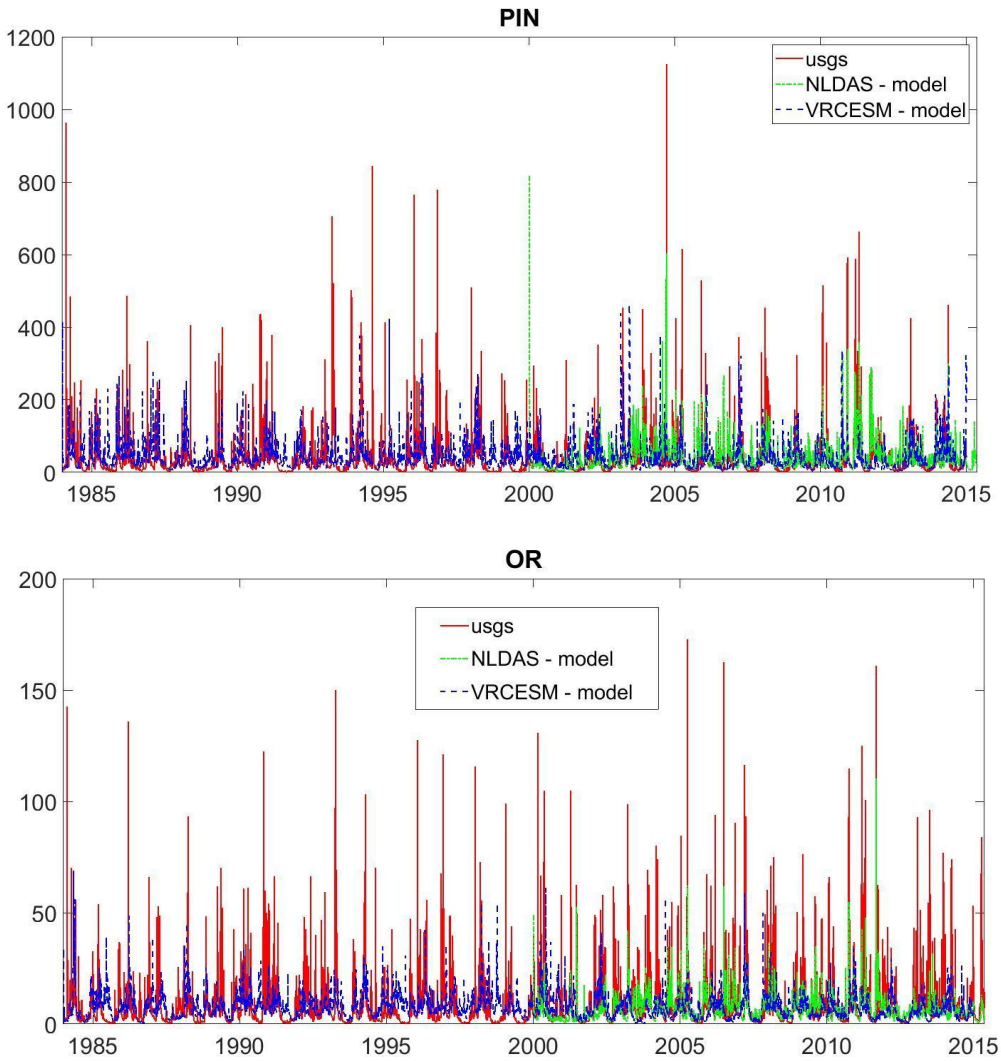


Figure 5: Streamflow estimations from observed and two modeled simulations for all of the Susquehanna river basin and 2 sub-basins.
CLM-PAWS demonstrated good performance for the streamflow simulation of this region using North American Land Data Assimilation System (NLDAS) forcing (2000-2017). On the other hand, streamflow peaks were underestimated by VR-CESM, as compared to NLDAS.

For the most part, the models produce realistic precipitation distributions and decent streamflow predictions including the median flow and 33-percentile flow. However, despite a good skill score in distributions of streamflow, the models are still under-estimating extreme events. Both climate projections and the hydrologic model contributed uncertainties. But for water availability forecasts (which presumably are more concerned with PDF or median streamflow), these models can be trusted although more work is needed for using the model for extreme streamflow related decisions.

4.2.3. Discussion and Conclusions

Streamflow PDF predictions directly based on variable resolution CESM simulations will likely be sufficient for water availability assessment but may still under-estimate extreme events.

Further refining of models would be necessary to appropriately capture extremes. A key limitation of the Perkin skill score is that it only provides a measure for similarities in PDFs. Therefore, it is possible to get a good skill score even if variables like extreme streamflow are not well represented. In this work, only one variable resolution model skill has been evaluated. Further work is needed to assess CLM-PAWS simulations for different models (such as CORDEX) as well as for future projections.

4.3. Extratropical Cyclones

Summary:

- This work focuses on how well the CESM LENS model is able to simulate precipitation (both total precipitation and precipitation as snow) associated with wintertime extratropical cyclones (ETCs) in the northeastern US, and the minimum model resolution that is needed to capture such events.
- The model was found to credibly capture the precipitation impact from ETCs although it produces a higher number of “low-end” storms. CESM LENS not only reproduced patterns of coastal ETCs, but also genesis locations which implies dynamic credibility.
- The required minimum resolution for adequately simulating ETCs is coarser than TCs, which is consistent with previous work that demonstrates that the minimum resolution needed to simulate, and track cyclones is ~1deg.

4.3.1. Background and Methods

Coastal storms, such as tropical and extratropical cyclones (TCs and ETCs), are responsible for a substantial amount of disaster related losses in the U.S. every year. Thirty nine percent of the U.S. population lives in counties directly on the coastline; a significant amount of the nation’s critical energy infrastructure is located in these counties. Credible simulation of these events can help better understand and prepare for future changes in precipitation associated with these events. It is often understood that high resolution models may be needed for appropriately simulating regional scale extreme precipitation events, however there is no clarity on what is the minimum resolution needed to capture such events. Since increasing model resolution can be costly in terms of resources and time, better understanding the overall benefits of increased resolution can be useful for effective use of resources. The overall hypothesis is that the current class of GCMs and RCMs (1deg and finer) can do a sufficient job simulating storm-level metrics (i.e., individual cyclones) due to the spatial scales of these predominantly baroclinic storms. This research focuses on the following research questions:

1. ***Can models credibly simulate precipitation associated with wintertime ETCs in the North Eastern US?***
2. ***What is the minimum resolution needed to capture such events?***

ETCs can be tracked using Lagrangian techniques and following cyclone-specific markers such as sea level pressure minima (Figure 6). A software package named "TempestExtremes" was developed, which is a flexible open-source software for tracking storm features in climate data (Ullrich & Zarzycki, 2017; https://github.com/ClimateGlobalChange/tempestextremes).[2] For the purposes of ETCs, the efficacy of this tracking algorithm was first evaluated. Following tracking, total precipitation and snowfall-only precipitation was integrated along a track to highlight hydro-climatological and societal impacts. ETCs were then classified using NOAA Regional Snowfall Index (RSI) (Squires et al., BAMS, 2014).[3] The RSI ranks snowstorm impacts on a scale from 1 to 5. For example, RSI of 4 is a crippling storm, while 5 is ranked highest as extreme. An additional metric was also developed (Regional Precipitation Index, RPI) which evaluates the total water equivalent impact of discrete ETCs, analogous to RSI.

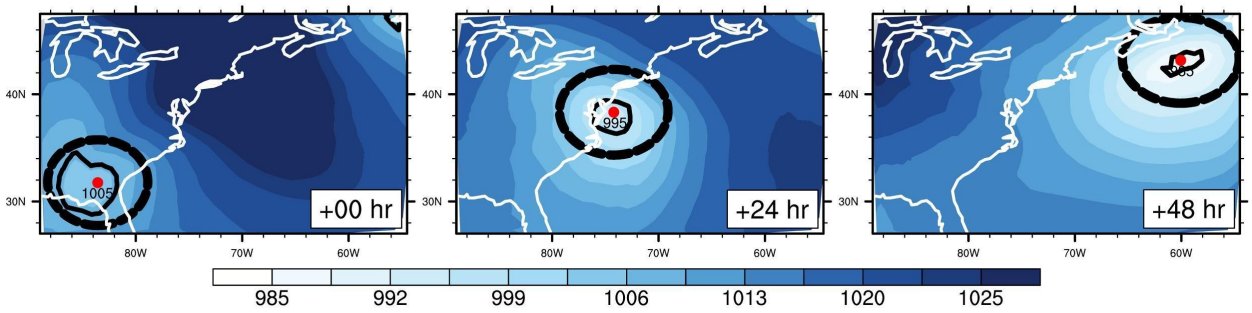


Figure 6: An example of the ETC tracking mechanics.
Sea level pressure is color contoured with the minima (center of storm) being tracked by the red dot across a 48 hour period.

This particular research has focused on examining these winter storms for present day as well as two future time slices (“mid-century” and “end of century”). Model skill evaluation has been conducted with the Community Earth System Model, focusing on a large ensemble of simulations (35 members). The performance of the model for storm metrics has been tested against both gridded reanalysis products (such as JRA-55), and present-day stats have also been compared to NOAA’s hand-curated reference dataset for regional snowstorms. Ongoing evaluation of the variable-resolution CESM model, also permits evaluation of the minimum resolution needed for adequately simulating ETCs.

4.3.2. Key Results

The efficacy of the tracking algorithm was tested in terms of how well the algorithm "found" historical storms without manual intervention. It was concluded that the technique can reasonably match historical datasets from NOAA when using observational gridded products such as reanalysis in an automated manner (Figure 7).

[2] Ullrich, P.A. and C.M. Zarzycki (2017) "TempestExtremes v1.0: A framework for scale-insensitive pointwise feature tracking on unstructured grids" Geosci. Model. Dev. 10, pp. 1069-1090, doi: 10.5194/gmd-10-1069-2017.
[3] Squires, Michael F., et al. "The regional snowfall index." *Bulletin of the American Meteorological Society* 95.12 (2014): 1835-1848.

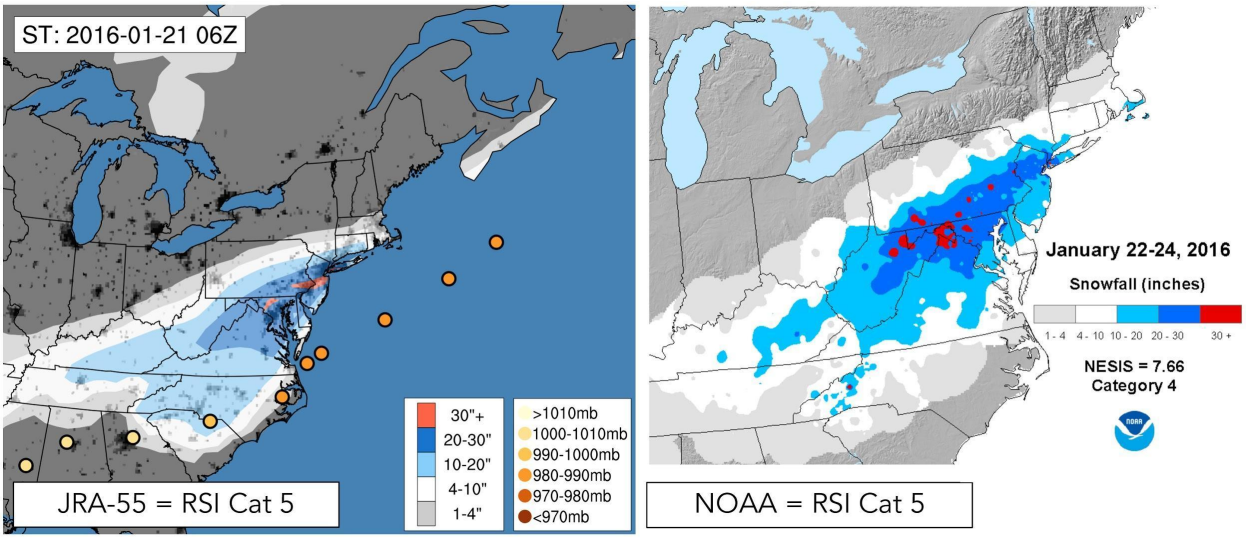


Figure 7: Example of an ETC tracked in reanalysis compared to observations
The panel on the left shows storm-total snowfall (color contours) from a tracked storm (track denoted by colored dots) in JRA-55. The panel on the right shows the same but hand-contoured by NOAA from station observations. The storm on the left was tracked without the need for manual intervention by using software developed in this project.

In terms of model evaluation in projecting future storms, one climate model - the Community Earth System Model (CESM's) Large ENSemble (LENS) with 35 ensemble members was evaluated. It was found that the model can credibly capture the precipitation impact from these events, although it produces a higher number of "low-end" storms. It is unclear whether this is a model bias or resulting from observational uncertainty, however. It was also found that CESM LENS can not only reproduce the spatial patterns of coastal ETCs, but can also reproduce from common genesis locations, which implies dynamic credibility (Figure 8).

Through analysis of variable-resolution CESM model, it was found that the required minimum resolution for adequately simulating ETCs is coarser than TCs, although this doesn't consider other benefits of resolution such as topography. Overall, at the regional scale, the model seems to be able to capture coastal extratropical cyclones patterns credibly. This is consistent with previous work that demonstrates the minimum resolution needed to simulate and track cyclones is ~1deg.

These results imply that the current generation of global ESMs (and therefore child regional models) can generally be trusted for large-scale ETC patterns. This is obviously dependent on other large-scale biases and may vary model-to-model but there is not a structural deficiency arising from coarse grid spacing and storm resolvability like there is for TCs. Models generally project a great deal of uncertainty in precipitation associated with these events. Even reanalysis data which are tightly constrained by observations vary significantly in their representation of total ETC-related precipitation. Therefore, the error bars on constraining even regional precipitation associated with storms remains large. Additionally, resolution improves mesoscale features embedded within ETCs, so hyper-local impacts (i.e., basin-scale and finer) will likely benefit from higher model resolution, even if the large-scale meteorological patterns do not change.

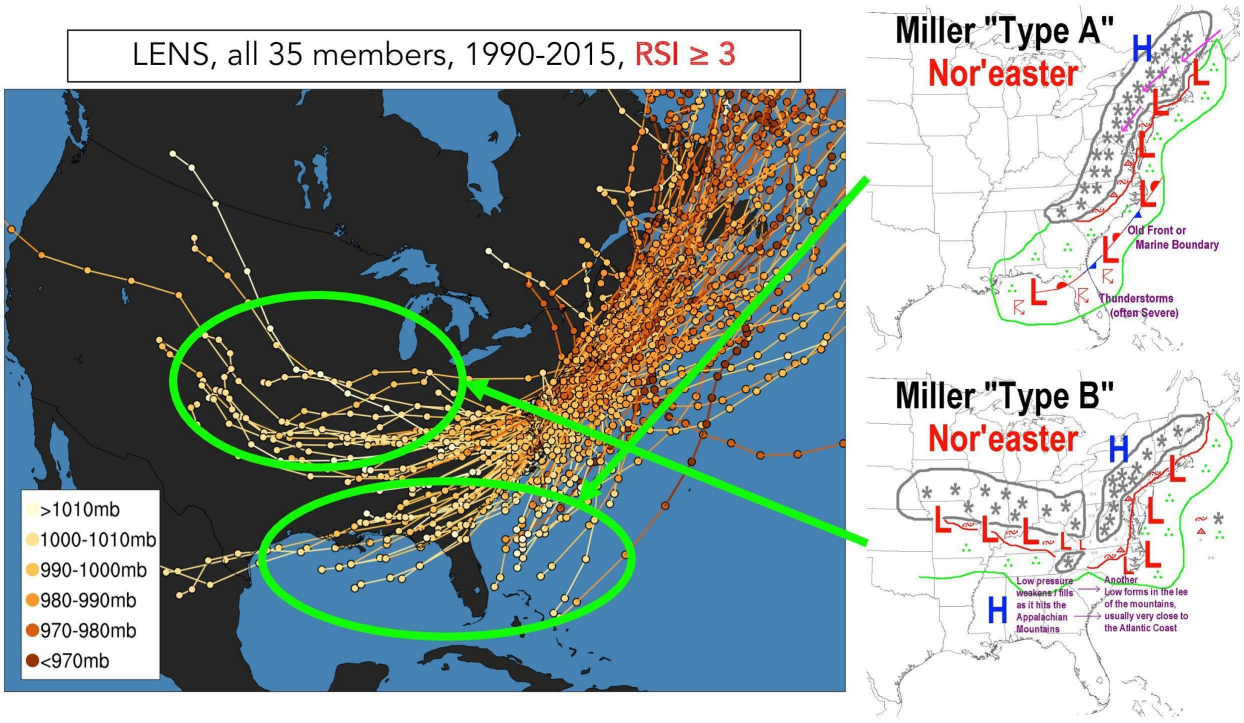


Figure 8: Trajectories of all major snowstorms (RSI>=3) in the present-day CESM ensemble. 6-hourly storm centers are denoted by dots and color-shaded by intensity. All ensemble members are included. Storm formation locations are compared to two common types of ETCs that impact the northeastern United States on right. Storm formation diagrams on right, courtesy of NOAA.

4.3.3. Discussions and Conclusions

With proper tools, the current class of earth system models should fundamentally capture ETCs at the regional scale. More nuanced hyperlocal precipitation impacts would benefit from higher resolution and/or downscaling. The resolution threshold is lesser for ETCs than TCs, mesoscale convection, etc. High-resolution models produce more accurate fine-scale structure although it is unclear how credible these scales are. However, global CESM produces slightly too many weak ETCs/storms but is reasonably well constrained observationally for more intense events. Precipitation remains tricky, and there is a lack of agreement even in reanalysis.

The statistical sample size for rare storm-level events (i.e., storms that only occur once every 25-50 years) is a limitation. Single realizations of future climate run drastically under-sample the potential for these storms. Therefore, further work focusing on more ensembles (either intra-model or intermodal) would be beneficial. Further work can also target compound or sequential storms. These storm-level metrics can also be used to tag phenomena such as rain-on-snow events to ETC passage or other aspects of the atmospheric circulation, which can give a more holistic view of winter/spring hydrology in the northeastern United States.

Acknowledgements and way forward

We are deeply grateful to all of Project Hyperion's water managers and scientists who patiently participated in the many back-and-forth engagements that form the basis of this report. We are also thankful to Bruce Riordan who co-led the engagements, Paul Ullrich for his agile leadership of the project, and Smitha Buddhavarapu for her careful review and edits of this report. Hyperion's successor project "HyperFACETS" is currently underway (2019-present) and will expand on the project's research activities and further work on creating broadly applicable tools for co-producing actionable climate science.

This work was supported by the Office of Science, Office of Biological and Environmental Research, Climate and Environmental Science Division, of the U.S. Department of Energy under contract DE-AC02-05CH11231 as part of the Hyperion Project, An Integrated Evaluation of the Simulated Hydroclimate System of the Continental US (award DE-SC0016605).

Appendix 1

Table A1: List of metrics and summary of scientific activities pursued by Hyperion project scientists

S No.	Science Activity	Lead Scientists	Description
1.	Precipitation IDF curves: Model skill and future projections	Abhishekh Srivastava and Richard Grotjahn	This work analyzed annual maximum precipitation (AMP) in historical simulations of VR-CESM and NA-CORDEX models. Models have considerable biases with respect to the station-based AMP. A new approach is adopted wherein models are selected based upon their historical performances. The historical and future (RCP8.5) simulation data of selected models are then bias-corrected and pooled to estimate non-stationary changes in the IDF estimates of 24-hr precipitation. The analysis suggests that almost all stations over the Susquehanna will observe significant increases in 24-hr precipitation for 2-100 year return periods. **Related Papers:** ● Srivastava, A., R. Grotjahn, and P.A. Ullrich (2019) "**A unified approach to evaluating precipitation frequency estimates with uncertainty quantification: Application to Florida and California watersheds**" *J. Hydrology* 578, pp. 124095, doi: 10.1016/j.jhydrol.2019.124095.
2.	Precipitation metrics (other than IDF): Skill evaluation	Abhishekh Srivastava and Richard Grotjahn	Precipitation indices (ETCCDI) namely annual mean P, SDII, CDD, CWD, Rx1day, Rx5day, R10mm, R20mm and Fr95T have been analyzed in the historical NA-CORDEX models. The two performance criteria (1) Taylor diagram and (2) interannual variability skill score (IVSS) are used to estimate model performance against station-based data. The results are summarized in the form of a "heat map" (also called stop-light diagram or portrait diagram). The analysis indicates that models have moderate skills in simulating both the observed spatial and temporal patterns of ETCCDI indices. **Related Papers:** ● Srivastava et al., (2019) **"Quantifying changes in 24-hr precipitation extremes by pooling NA-CORDEX models: Susquehanna watershed and Florida peninsula"**. *Water Resources Research. (*Under review).
3.	Streamflow Modeling: Skill evaluation	Chaopeng Shen and Wen-Ping Tsai	Historical streamflow simulations of the hydrological model CLM-PAWS were developed using NLDAS and VR-CESM forcings. CLM-PAWS demonstrated good performance for the mid-low streamflow

S No.	Science Activity	Lead Scientists	Description
			simulation. Further work on the hydrological model is underway to assess streamflow simulations for different datasets (such as CORDEX) as well as for future projections. **Related Papers:** ● Shen, C. (2018) "**A trans-disciplinary review of deep learning research and its relevance for water resources scientists**" *Water Resour. Res.*, 54 (11), pp. 8558-8593, doi: 10.1029/2018WR022643. ● Shen, C.P., E. Laloy, A. Elshorbagy, A. Albert, J. Bales, F.-J. Chang, S. Ganguly, K.-L. Hsu, D. Kifer, Z. Fang, K. Fang, D. Li, X. Li, and W.-P. Tsai (2018) "**HESS Opinions: Incubating deep-learning-powered hydrologic science advances as a community**", *Hydrol. Earth Syst. Sci.*, 22, pp. 5639-5656, doi: 10.5194/hess-22-5639-2018. ● Tsai, W-P., X.Y. Ji, K. Fang, C.P. Shen (2019) "**Revealing causal controls of storage-streamflow relationships with a data-centric Bayesian framework combining machine learning and process-based modeling**" *Water Resour. Res.* (Under Review) **AGU 2019 abstract:** https://agu.confex.com/agu/fm19/meetingapp.cgi/Paper/609941
4.	Snow Water Equivalent (SWE) triangle: Model skill and future projections	Alan Rhoades	A multi-metric framework is developed to assess agreements and disagreements of the annual snow season in spatially continuous snow water equivalent (SWE) estimates derived from reanalyses, regional and variable-resolution climate model simulations and a statistical downscaling approach. The more ephemeral snowpack of the northeast created issues for the direct use of the SWE triangle framework (developed for mountainous western U.S.), however, reanalysis and model SWE estimates more closely matched, particularly compared with California and Colorado. Under a high-emissions scenario, an ensemble of regional climate model simulations (i.e., NA-CORDEX) projects a dramatic decline in peak SWE volume and snow season length (mainly due to a reduction in accumulation season length of ~30 days) across the six Susquehanna River sub-basins. Figures A1, A2, A3 and A4 in the Appendix present some of the results from this study.
5.	Extra Tropical	Colin Zarzycki	This project developed metrics and software

S No.	Science Activity	Lead Scientists	Description
	Cyclones(ETCs) and Tropical cyclones: Model skill and sensitivity to variable resolution domain	and Kevin Reed	packages in order to evaluate coastal storms in gridded climate data, and evaluated the performance of these metrics in reanalyses, which allows a direct comparison to observations. Metrics were iteratively improved and showed capability to automatically extract key hydrologic events. The study also evaluated the performance of a current class of Earth System Models (ESMs) at ~1 deg resolution to reproduce ETCs over historical period. Through use of variable-resolution ESMs, the project evaluated the resolution sensitivity and value add associated with finer grid spacings and ETCs. **Related Papers:** ● P. A. Ullrich and **C. M. Zarzycki** (2017), TempestExtremes: A framework for scale-insensitive pointwise feature tracking on unstructured grids, *Geosci. Model Dev.*, **10**, 1069-1090, doi:10.5194/gmd-10-1069-2017. (Github repository) ● **C. M. Zarzycki** (2018), Projecting changes in societally-impactful northeastern U.S. snowstorms. *Geophys. Res. Lett.*, **45**, 12067–12075, doi:10.1029/2018GL079820.
6.	Variable resolution model: Future simulations	Colin Zarzycki	This study applied CESM Large Ensemble model runs to evaluate the projected changes in snowfall and total precipitation associated with ETCs over the northeastern United States during a mid-century and end-of-century period. **Related Papers:** ● **C. M. Zarzycki** (2018), Projecting changes in societally-impactful northeastern U.S. snowstorms. *Geophys. Res. Lett.*, **45**, 12067–12075, doi:10.1029/2018GL079820.
7.	Mesoscale Convective Systems (MCS) metrics modeling	Simon Wang and Binod Pokharel	Metrics were developed to track the MCS over the northeast US that covers the Susquehanna River Basin. The NARCCAP Data were also utilized. **Related Papers:** ● Pokharel, Binod, et al. "Climate of the weakly-forced yet high-impact convective storms throughout the Ohio River Valley and Mid-Atlantic United States." *Climate Dynamics* 52.9-10 (2019): 5709-5721. https://link.springer.com/article/10.1007/s00382-018-4472-0 ● Pokharel, Binod, et al. "Diagnosing the Atypical Extreme Precipitation Events Under Weakly

S No.	Science Activity	Lead Scientists	Description
			Forced Synoptic Setting: The West Virginia Flood (June 2016) and Beyond." *Climate Prediction S&T Digest* (2018): 8.https://repository.library.noaa.gov/view/noaa/17399/noaa_17399_DS1.pdf#page=16

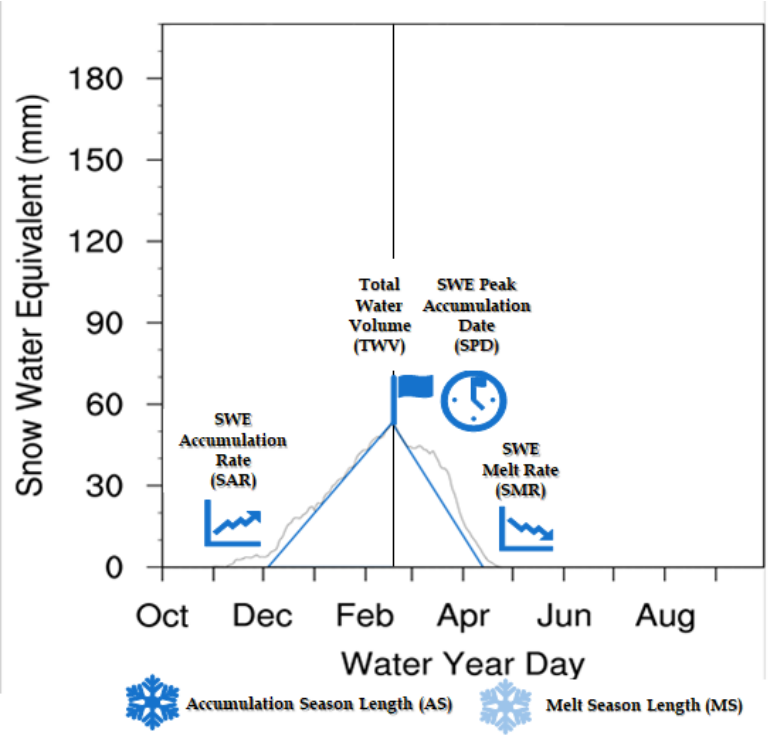


Figure A1: Snow water equivalent (SWE) triangle metrics
Six snow water equivalent (SWE) triangle metrics visually represented and overlaid on the observationally constrained Livneh 2015 dataset historical average snowpack life cycle for the Susquehanna River Basin.

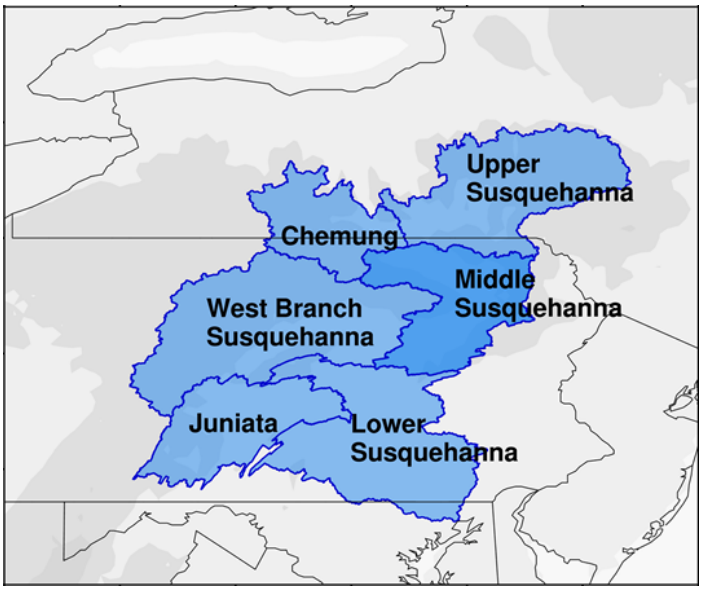


Figure A2: ***The six Susquehanna River sub-basins used to evaluate SWE datasets***

		SAR	TWV	SMR	SPD	AS	MS
L15 (reference)		0.00	0.00	0.00	0.00	0.00	0.00
NLDAS_2_SAC	*Observationally constrained*	-1.00	-0.89	-0.22	-0.49	-0.74	-0.87
NLDAS_2_VIC		-0.35	-0.29	0.13	0.06	0.02	-0.41
ECMWF_ERAINT_CRCM5 at 12km		-0.54	-0.57	-0.20	-0.43	-0.41	0.13
ECMWF_ERAINT_CRCM5 at 25km		-0.56	-0.51	-0.05	-0.03	-0.07	-0.22
ECMWF_ERAINT_CRCM5 at 50km		-0.46	-0.43	-0.18	-0.00	-0.10	0.22
ECMWF_ERAINT_CanRCM4 at 25km	*RCM forced by atmospheric reanalysis*	-1.03	-0.86	-0.34	-0.08	0.08	-0.37
ECMWF_ERAINT_HIRHAM5 at 50km		-1.18	-1.04	-0.23	-0.43	-0.38	-0.94
ECMWF_ERAINT_WRF at 12km		-0.49	-0.52	-0.27	-0.44	-0.38	-0.21
ECMWF_ERAINT_WRF at 25km		0.17	-0.01	0.07	-0.26	0.21	0.28
ECMWF_ERAINT_WRF at 50km		0.16	-0.01	0.19	-0.18	0.24	0.01
CanESM2_CanRCM4 at 50 km		-1.05	-0.94	-0.47	-0.43	-0.70	-0.40
CanESM2_CRCM5 at 50 km		-0.44	-0.48	-0.35	-0.34	-0.33	-0.18
MPI_ESM_CRCM5 at 50 km		-0.21	-0.20	-0.29	0.03	0.52	0.31
MPI_ESM_RegCM4 at 50 km		1.75	1.16	0.85	-0.15	0.40	-0.19
GFDL_WRF at 50 km	*RCM forced by GCM*	0.51	0.24	-0.20	-0.16	0.11	0.24
GFDL_WRF at 25 km		0.43	0.19	-0.15	-0.14	0.32	-0.00
HadGEM2_WRF at 50 km		0.53	0.23	0.29	-0.22	0.20	0.24
HadGEM2_WRF at 25 km		0.47	0.25	-0.04	-0.06	0.44	0.09
VR28_EPAC_141 at 25 km		0.07	0.18	0.02	0.41	0.09	-0.29
VR28_EPAC_94 at 25 km		-0.18	-0.08	0.34	0.36	-0.01	-0.51
VR28_EPAC_47 at 25 km		0.12	0.21	0.14	0.35	0.15	-0.27
VR28_NATL_EXT at 25 km	*Variable-Resolution Global Climate Model*	1.35	1.50	0.35	0.87	0.77	0.06
VR28_NATL_REF at 25 km		0.58	0.70	0.53	0.58	0.61	-0.48
VR28_NATL_WAT at 25 km		1.29	1.37	0.16	0.65	0.62	0.19

Figure A3: Z scores for the SWE triangle metrics across six Susquehanna River sub-basins

The six SWE triangle metrics include snowpack accumulation rate (SAR), total water volume at peak accumulation (TWV), snowpack peak accumulation date (SPD), snowpack melt rate (SMR), the length of the accumulation season (AS), and the length of the melt season (MS). The Z score is computed by using the mean and standard deviation from Livneh, 2015 (L15). Red (blue) indicates positive (negative) Z score bias, and saturation indicates the magnitude of

bias. Similar to previous figures, text color is used to distinguish resolution in (b) and global climate model forcing data set in (c). SWE = snow water equivalent.

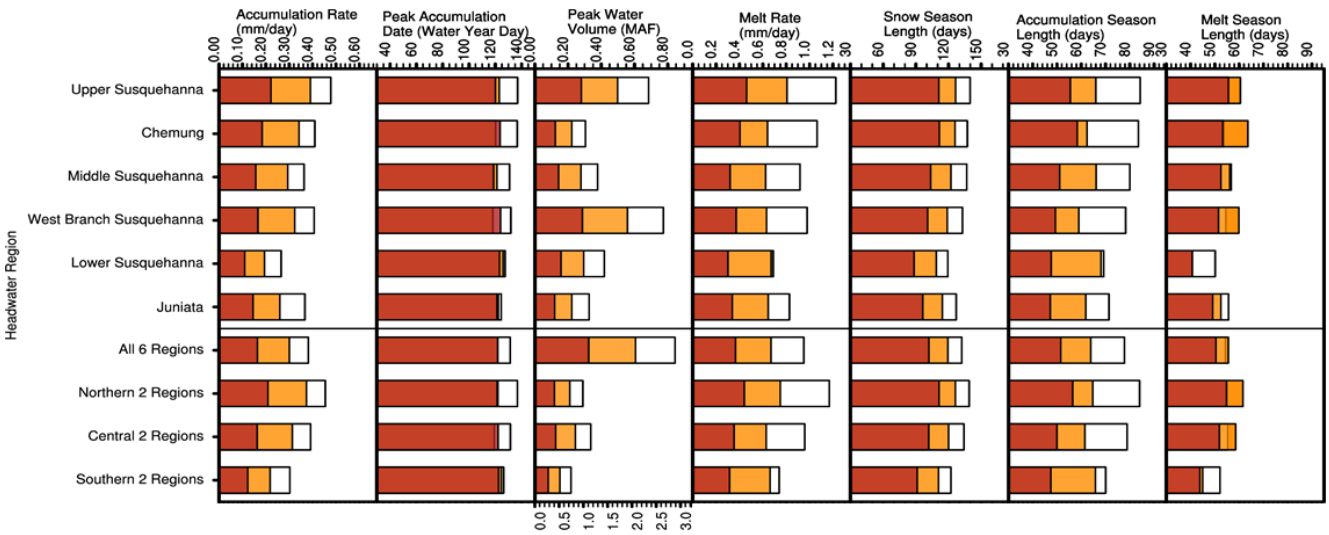


Figure A4: Snow water equivalent triangle metrics projections

SWE metrics for the eight North American Coordinated Regional Climate Downscaling Experiment simulations across the six Susquehanna River sub-basins and the northern, central, and southern aggregate regions. Figure shows future changes in these metrics for mid-century and the end of century with RCP 8.5. Color is used to distinguish 1985–2005 (white), 2039–2059 (orange), and 2079–2099 (red). For Peak Water Volume the top x axis is used for individual regions and the bottom x axis is used for aggregate regions. MAF = million acre-feet.

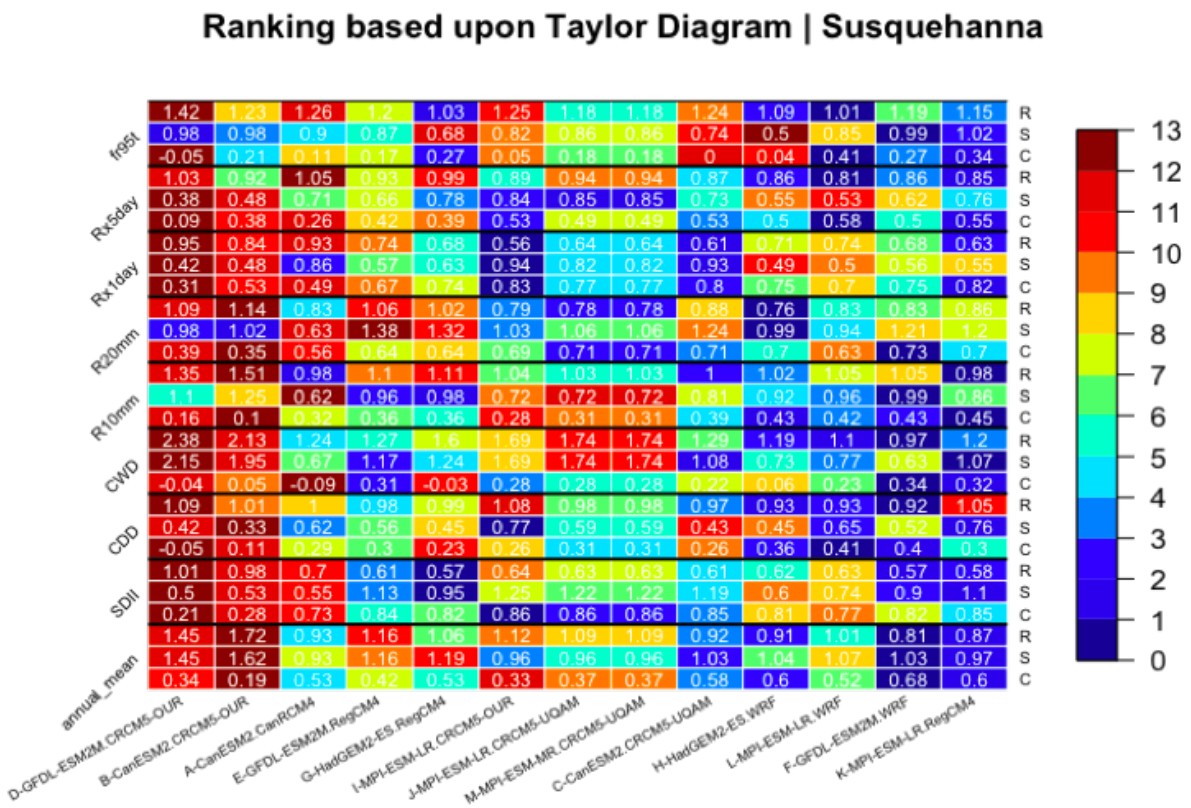


Figure A 5: Model rankings for different NA-CORDEX models and for different precipitation metrics based upon Taylor diagram.

The diagram compares the spatial pattern of the long term means of P-indices in terms of their standard deviation, root mean square error (RMSE) and spatial correlation wrt. observation. Color bar indicates ranking of models between 1 and 13. The letters on the right vertical axis have the following meanings. C: spatial correlation between model and the observation, S: Standard deviation in model divided by the standard deviation in the observation, R: RMSE between model and the observation divided by the standard deviation in the observation. The numbers inside the boxes indicate their actual values. The figure shows that WRF set of models are among the best performers based upon Taylor diagram.

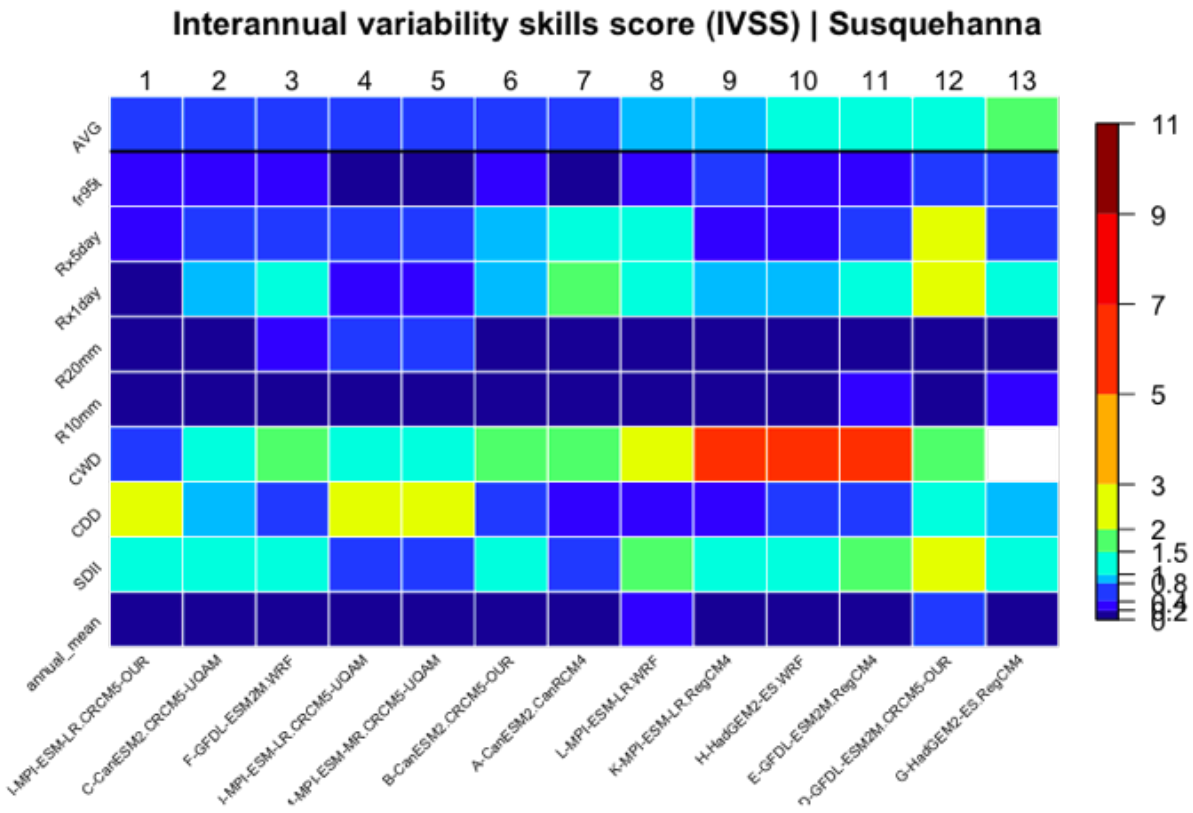


Figure A6: Ranking of NA-CORDEX models based upon IVSS, for different precipitation metrics.
IVSS is a temporal variability skill score based upon the ratio of temporal standard deviation of the indices in model to the standard deviation of the indices in the observation at each station. Color bar indicates IVSS (score); the smaller the score the better is the model performance. The numbers on top indicate model ranking between 1 and 13 based upon IVSS. The figure shows that nearly half of the models have reasonably good (and similar) skills in simulating the temporal variability in the observation.

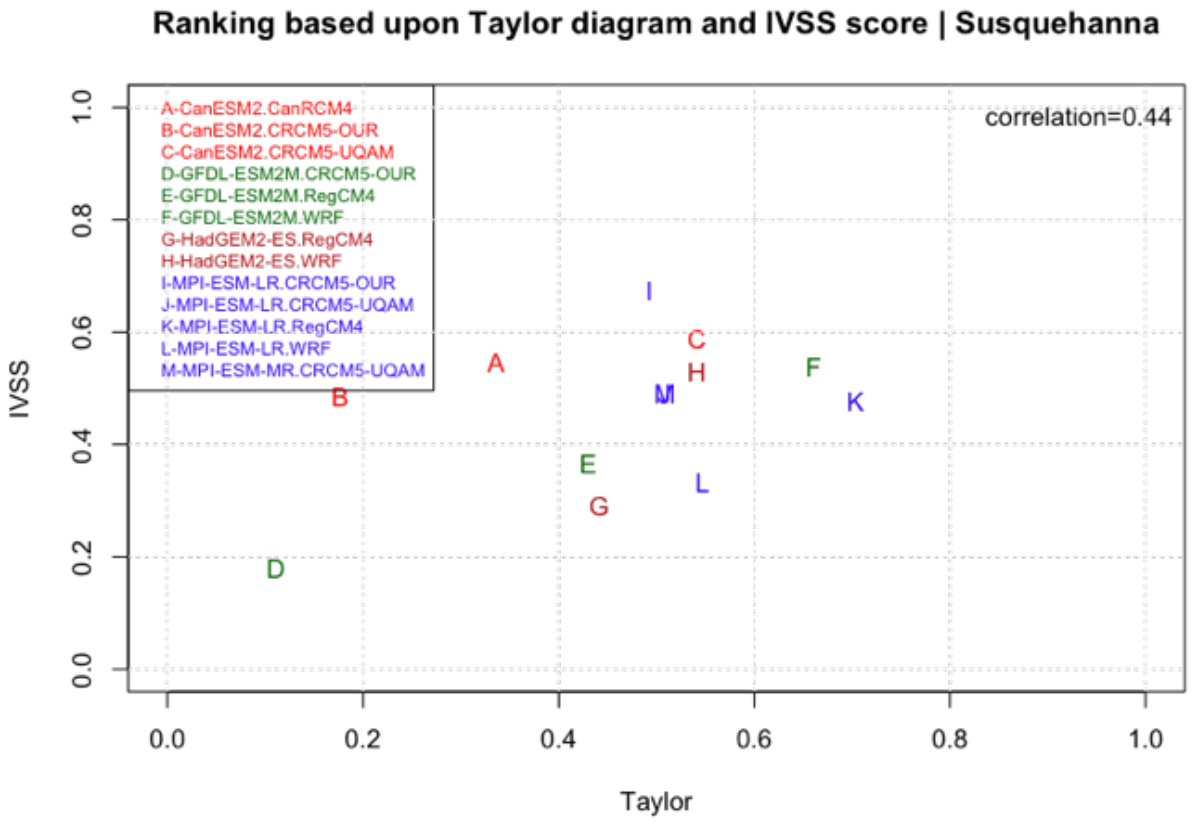


Figure A7: Overall rankings based on both Taylor diagram and IVSS for NA-CORDEX models, for various precipitation metrics

The correlation between model ranking scores from Taylor diagram and IVSS is shown in the upper left corner. The correlation coefficient indicates the degree to which NACORDEX models performing well in spatial simulation of indices also perform well on temporal scale. The figure shows that most of the models have moderate skills in simulating the spatial and temporal patterns of the observed indices- models lie in the center of the scatter diagram.

Overall, Figures A5, A6 and A7 suggest that for the precipitation metrics analyzed, the MPI set of models have the most consistent (moderate) performance. GFDL-ESM2M.CRCM5-OUR (D) performs worst among all models.

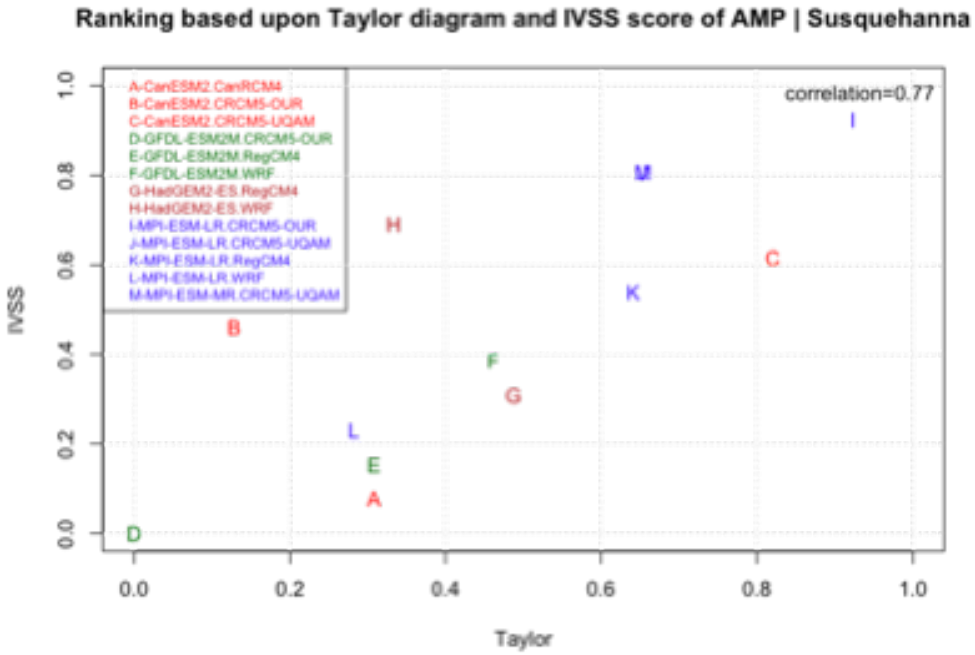


Figure A8: Overall rankings based on both Taylor diagram and IVSS for NA-CORDEX models for annual maximum precipitation

The correlation between model ranking scores from Taylor diagram and IVSS is shown in the upper left corner. The correlation coefficient indicates the degree to which models performing well in spatial simulation of indices also perform well on temporal scale. This figure shows that for the metric of Annual Maximum Precipitation, Models I, M, C,K,F are the best performing ones. In the previous figure, AMP is referred to as Rx1day.

References

[1]Jagannathan, K., A. D. Jones, and I. Ray, The making of a metric: Co-producing decision-relevant climate science. Bull. Amer. Meteor. Soc., doi: https://doi.org/10.1175/BAMS-D-19-0296.1.

[2]Ullrich, P.A. and C.M. Zarzycki (2017) "TempestExtremes v1.0: A framework for scale-insensitive pointwise feature tracking on unstructured grids" Geosci. Model. Dev. 10, pp. 1069-1090, doi: 10.5194/gmd-10-1069-2017.

[3]Squires, Michael F., et al. "The regional snowfall index." Bulletin of the American Meteorological Society 95.12 (2014): 1835-1848.