FINAL REPORT:
Implement and Operationalize a Customized Meteorological Framework in Bagmati-Adhwara and Kosi Basins in Bihar State

Contact: Thomas Hopson
Research scientist
National Center for Atmospheric Research
hopson@ucar.edu

For the attention of:

Dr. Deepak Singh, Program Leader,
The World Bank
The Hindustan Times House (Press Block) 18-20
Kasturba Gandhi Marg New Delhi - 110001
Email: spriya1@worldbank.org

Company Information:

Name	National Center for Atmospheric Research
Registered Office Address	1850 Table Mesa Dr Boulder, CO 80305 USA
Corporate Office Address	University Corporation for Atmospheric Research P.O. Box 3000 Boulder, CO 80307-3000 USA

Acknowledgements

We are most thankful for the support and constructive suggestions provided throughout this project by Deepak Singh, Senior Disaster Risk Management Specialist, and Dr. Satya Priya, Senior Water Resource Management Specialist, World Bank, New Delhi, India.

This work was made possible through funding support provided by the **World Bank Group**, through the award:
Title: Implement and Operationalize a Customized Meteorological Framework in Bagmati-Adhwara and Kosi Basins in Bihar State
Contract ID: 7183269

Table of Contents

Acknowledgements 3

List of Figures 6

List of Tables 10

Abbreviations Used 11

Overview of Report 13

Introduction 15
Background 15
Objective 15
GIS Delineation 16

Rainfall Data Components 17
Precipitation Estimates 17
Thorpex Tigge Multi-model Forecasts 18
US NCEP and Canada CMC Forecasts 20
NMME Seasonal forecasts 21

Rainfall Forecasting Methods 22
Bias Correction -- Quantile-to-Quantile Mapping 22
Calibration and Multi-modeling -- Quantile Regression 23

Rainfall Forecast Products and Evaluation 24
Bias and Anomaly Correlation Comparisons 24
Multi-model Benefits 28
Web-based Archive Comparisons 37

Riverflow Data Components 40
River Gauge Measurements 40
Satellite River "Width" Measurement 40
Quality Control 41

River Forecasting System 42

Riverflow Forecasting Methods 45
Rating Curve Estimation 45
Stage-Discharge Relationship Flood Forecasting 47
Quantile Regression – individual ensemble and multi-model forecasting 52

Dashboard Framework and Data Access 53
Current Condition Overview 53
Displays of Ensemble Numerical Weather Prediction and Satellite Precipitation 54
Displays of Water Level Forecasts and Observations 59
Review of Past Severe Events 61
Amazon Web Services Access 61

Direct Data Access (FTP) 61

Appendix A -- Quality Control Details 63

River gauge quality control 63

Satellite-deived river width quality control 68

Appendix B -- Quantile-to-Quantile Mapping 71

Appendix C -- Quantile Regression 74

Appendix D -- Verification Metrics 76

Verification and Evaluation Metrics 76

The Brier Score 76

The Brier Skill Score 77

The Rank Histogram 77

References 79

List of Figures

Figure 1: The satellite rainfall products we used for this project. 17
Figure 2: Green dots show the locations of the 12 NWP centers that provide data to the (nominally 3-day delayed) near-real-time THORPEX-TIGGE data archive. This project will primarily utilize forecasts from the following centers: CMA, CMC, ECMWF, UKMO, NCEP, MeteoFrance, JMA, and CPTEC. 19
Figure 3: In our displays, we provide forecasts from the top four centers in this list. But in our multi-model rainfall and riverflow forecasts, we utilize data from the top eight centers. 20
Figure 4: Multi-model, multi-center ensemble seasonal rainfall forecast data available for Bihar. 21
Figure 5: omparisons of the rainfall amounts (mm/day) for August average rainfall from 2011 to 2015 for four of our weather forecast centers 0 to 4 day lead-time forecasts over the Bagmati, Kosi, and Adhwara basins, noting the similarities, but also distinctive differences in both the amounts and locations. 25
Figure 6: biases (subtracting the forecasts from the observations, in mm/day) of the values seen in the figure above. 26
Figure 7: Anomaly correlation skill (the correlation of both the observations and of the forecasts, both differenced from "climatology") in the four centers' forecasts. 27
Figure 8: correlations of the Canada (CMC), ECMWF, NCEP (US), and UKMet office forecasts when averaged over the region's catchments, as a function of lead-time and averaging window (24hr and 5day). Note that the UKMet office foreasts only extend out 7 days. 28
Figure 9: time-series of quality-controlled stream gauge heights, with arrows pointing to the gauge locations. Above these two locations we catchment-integrate rainfall fields, and calibrate these values. 29
Figure 10: 5-day lead-time calibrated time-series of the catchment-integrated ensemble rainfall forecasts for our five centers (colored lines), along with the observed (satellite and rain gauge combined rainfall; black line), along with multi-model forecasts for two centers (CMC and NCEP) and for all 5 centers. . 30
Figure 11: Same as the figure above, but for the catchment-integrated rainfall above gauge 029-mgd5ptn. 31
Figure 12: rank histograms (as a verification metric of how "trust-worthy" are the ensemble forecast's range – see appendix) of the 5-day lead-time catchment-integrated rainfall above our two gauging locations compared to observed rainfall. For rank histograms, the "flatter the better" – what we see below is that essentially the histograms are "flat" (except for some statistical randomness). 32
Figure 13: Brier skill scores as a function of lead-time going out 1- to 15-days in advance for each of our 5 centers individually, and combined into multi-models. The blue line is the brier skill score relative to a "persistence" forecast, and the

green lines are relative to "seasonal climatology". Note the overall superiority of the 5-center multi-model for all lead-times. 33
Figure 14: Same as the figure above, but for the catchment-integrated rainfall above gauge 029-mgd5ptn. 34
Figure 15: RMSE skill scores as a function of lead-time going out 1- to 15-days in advance for each of our 5 centers individually, and combined into multi-models. The blue line is the RMSE skill score relative to a "persistence" forecast, and the green lines are relative to "seasonal climatology". Note the overall superiority of the 5-center multi-model for all lead-times. 35
Figure 16: Same as the figure above, but for the catchment-integrated rainfall above gauge 029-mgd5ptn. 36
Figure 17: percentage of usage of each center's forecasts in the calibration process. For both of our basins above the gauging locations (left and center panels), ECMWF is most used, as well as for all the basins in our region (right-most panel). However, the relative contributions from the other centers varies by location. 37
Figure 18: comparisons forecasts (ECMWF in this case) and observations (satellite rainfall) as seen during the July 2017 flooding events from our "archive" web page (http://bihar.rap.ucar.edu/archive/). Shown are comparisons of 24-hr accumulated rainfall from July 9th to July 10th for 5-day lead-time ECMWF ensemble mean forecast compared to satellite estimates (top panel forecasts, bottom panel merged satellite rainfall). This plot generally shows good agreement, but with variations at small spatial-scales. 38
Figure 19: Similar to the figure above, but for 5-day accumulated rainfall starting on July 9th, and with ECMWF forecasts initialized on the same day. 39
Figure 20: Flowchart of the modified CFAB short-term forecasting scheme. Five steps (I-V) are outlined schematically and within each is a brief description of the data used, tasks performed and etc. The arrows show the path of the data through the system. The components of Step I (clear boxes) represent daily inputs into the scheme. At Step II (dotted box), the TIGGE ensemble forecasts are corrected statistically to reduce systematic error. Step III (hatched boxes) represents the hydrologic modeling process, where all inputs are integrated to produce river flow forecasts at our 87 forecasting points. Step IV (checkered box) represents the process by which all hydrologic uncertainties are accounted for in the probabilistic forecasting process. In Step V, the ensemble forecasts are tailored to produce probabilistic information necessary for displays and user needs. Note that all processes below the horizontal dashed line are done independently for each day and each 1- to 16-day forecast lead-time. 43
Figure 21: conceptual idea of using upstream stages (stage 1 and stage 2) to forecast a downstream location based on a 5-day flood wave travel time from the most upstream to downstream locations. 48
Figure 22: depending on the number of upstream gauges, multiple forecasts can be estimated, one for each gage. 49

Figure 23: the error of each upstream gauge (an example of which is shown here from our website provided) forecast is available directly from the derived "rating curves". 50
Figure 24: the error from the derived rating curves of upstream gauges is then combined in the manner shown in this figure, to generate an "optimal" combined forecast. 51
Figure 25: results of our combined upstream stage forecast for one forecasting site. The blue is the observer, red is the forecast, and the grey represents to the forecast period. 51
Figure 26: Screenshot of current conditions occurring in the region, as seen from our main home page for the project. 54
Figure 27: Overview of the processing and display of catchment-integrated real-time satellite precipitation and TIGGE ensemble precipitation forecasts over Bihar catchments. 55
Figure 28: Step 1 – Select a model or Satellite product 56
Figure 29: Step 2 – Select the forecast initialization time. 57
Figure 30: Step 3 – Select the 24 hour or 5 day accumulation date. 57
Figure 31: Result - An interactive web map is displayed with zoom and pan functionality. 58
Figure 32: By clicking on a watershed a popup window appears with more detailed information about the model ensemble statistics. 59
Figure 33: When zoomed in the sub-catchments are displayed for a more detailed view. 59
Figure 34: displays provide visualize of the geographic extent of the river basins, observations, ensemble forecasts, and quality control analyses at that location. Further above is the live display, and immediately above is a selection of two gauge forecasts over Bihar. 61
Figure 35: River level data from station 001-MBDGHY. The dot color identifies the data as: (blue) likely valid river levels, (red) very likely erroneous data, (green) questionable data, and (cyan) data observed at the exact same time. 65
Figure 36: River level data from station 016-MGD3VNS, showing examples of erroneous extremely high and low river gauge values. 68
Figure 37: Time series of M/C ratio for a location on the lower Ganges River. Pixel colors indicate its QC flag. 70
Figure 38: The quantile-to-quantile (q-to-q) correction system. Both modeled and observed precipitation are binned into quantiles. The model precipitation is mapped onto the observed precipitation fields by setting respective modeled quantiles to observed quantiles. In the figure, the forecast precipitation (p_{fcst}) of the 80th quantile is set to observed precipitation (p_{adj}) 80th quantile. The method ensures that the forecasts produce the same number of "no rain" events as the observations. 73
Figure 39: Previous day's temperature (persistence) used as a forecast of 24hr temperature, Salt Lake City airport data, 1979 to 2001. Red line is the fit for the central tendency (mean) using standard linear linear regression; middle black

line is the fit of the median (0.5 quantile); upper black 0.9 quantile; lower black 0.1 quantile. Notice the similarity but noticeable divergence of the median and mean fits for larger temperatures. Notice also the heteroscedastic behavior of the persistence fitting, which is seen by the convergence of the 0.1 and 0.9 quantile lines for higher temperatures..... 75

Figure 40: Troubled rank histograms: panel on the left shows an ensemble forecast system whose ensembles are too dispersed (the observation falls too often in the middle of the distribution), while the panel on the right would be for a too narrow ensemble forecasting system. For a perfect ensemble forecasting system, rank histograms would be "flat" (i.e. the observation falling into each bin the same number of times)..... 78

List of Tables

Table 1: QC score values, and their meaning. 41
Table 4: QC score values, and their meaning. 64

Abbreviations Used

Abbreviation/Acronym	Expanded Form
App	Self-contained software application
BUET	Bangladesh University of Engineering and Technology
CFAB	Climate Forecast Applications for Bangladesh
CMA	China Meteorological Administration
CMC	Canada Meteorological Centre
CMORPH	CPC MORPHing technique
CPC	NOAA Climate Predition Center
CNES	Centre National d'Etudes Spatiales
CPTEC	Centro de Previsao de Tempo e Estudos Climaticos
CSV	Comma separated values
CWC	Indian Central Water Commission
DEM	Digital Elevation Map
DFO	Dartmouth Flood Observatory
ET	Evapo-transporation
ECMWF	European Centre for Medium-Range Weather Forecasts
Emissivity	Emission of electromagnetic radiation in the microwave-frequency
GBM	Ganges-Brahmaputra-Meghna basin
GDR-D	Geophysical Data Records, version D
GIS	Geographical Information System
GSMaP	Global Satellite Mapping of Precipitation
GTS	Global Telecommunication System
JAXA	Japan Aerospace Exploration Agency
JRC	European Union Joint Research Council
Leaflet	Open-source JavaScript library for interactive maps
MeteoFrance	French Meteorological Service
NASA	U.S. National Aeronautics and Space Administration
NSE	Nash-Sutcliffe Efficiency
NCEP	U.S. National Center for Environmental Prediction
NETCDF	Network common data format
NOAA	U.S. National Oceanic and Atmospheric Administration
NWP	Numerical weather prediction

OSTA	Office of Space and Terrestrial Applications
P	Precipitation
Persistence	An estimate of the weather "tomorrow" being the same as "today"
Plotly	Online analytics and data visualization tool
Q	River discharge
QC	Quality control
QR	Quantile regression
Q2Q	Quantile-to-quantile mapping
R	R statistical computing language
RMSE	Root mean square error
S	River stage height
SS	Skill score
SARAL	Satellite with ARgos and ALtika
Shiny	Web application framework for R
Stg	River stage height
SWOT	Surface Water and Ocean Topography Mission
THORPEX	The Observing system Research and Predictability Experiment
TIGGE	THORPEX Interactive Grand Global Ensemble
Time of Concentration	time needed for water to flow from the most remote point in a watershed to the watershed outlet
TRMM	Tropical Rainfall Measuring Mission
UKMET	U.K. Meteorological Office
UMD	University of Maryland
WRF	Weather Research and Forecasting model

Overview of Report

This project and report covers the development of a customized rainfall and riverflow observing and forecasting framework for the Bagmati-Adhwara and Kosi Basins in Bihar State. The overall objective was to equip the Flood Management Improvement Support Center (FMISC) under the Water Resources Department, Govt. of Bihar with a robust 24X7 feed of customized weather ensembles of satellite based rainfall estimates and short-to-medium range rainfall forecasts, along with and modules that allow FMISC to utilize these data directly into operational FMISC flood forecast models developed for the Bagmati-Adhwara (B-A) and Kosi basins. In addition, NCAR has provided "experimental" seasonal mult-model rainfall products also for Bihar. Web-based dashboards have been developed so that the ensembles are viewable and downloadable for hydrologic model ingest. In addition, operational riverflow forecasts going out 1- to 16-day lead-times in advance have also been produced corresponding to river height gauging locations currently in operation in these basins. These riverflow forecasts utilize the same rainfall data sets as are being provided directly to the FMISC, in addition to employing a set of customized hydrologic modeling and quality control tools to generate optimized ensemble riverflow forecasts for these locations. In addition, much of this system is running in parallel on the Amazon Cloud Services (AWS), and this AWS account is designed for shareable access between NCAR and the FMISC such that FMISC staff can directly access and modify all the technologies and data sets created under this consultancy.

In the context of this objective, the structure of this report is as follows: we first present the rainfall observing and forecasting data sets covering the key outputs of this project that are useful for the implementation of effective flood forecasting warning systems over the Bagmati-Adhwara, and Kosi Basins. Within the main body of this report, we first provide a forecasting system overview to build context for materials presented in the rest of the report. This is followed by a review and evaluation of our data sources useful for flood forecasting. These products include satellite-derived rainfall estimates, ensemble weather forecasts from a range of global forecasting centers, in situ river stage measurements and their quality control, and satellite emissivity signals correlated to changes in river width (or simply, "river width measurements"). We also provide visualizations and links to access many of these data in real-time, with many of these quality controlled for user ease. We then discuss forecasting methodologies we've utilized and introduced in this project, along with a discussion on the validation protocol we utilize for establishing operational reliability and limiting conditions on the forecast utility and skill. This is followed by a discussion on the skill of the forecasting products themselves, both catchment-integrated precipitation and river discharge forecasts. We then conclude the main body of this report by presenting the visualization and information sharing tools we've developed to communicate our operational forecasts. In the appendices, we cover many of these topics in more technical detail.

Introduction

Background

The Flood Management Information System (FMIS) Centre in Water Resources Department (WRD) of Government of Bihar (GoB) aims to generate and disseminate timely and customized information to move from disaster response to improved disaster preparedness and to effectively support flood control and management in the flood-prone areas of the State. Conceptual flood models are being developed in Bagmati-Adhwara (B-A) basin and Kosi basin to enhance the traditional system of stage-level warnings based on gauge-to-gauge correlation in order to transform it to flood forecasting services for effective flood management strategy and planning. Part of the requirements for these conceptual flood models to provide future forecasts of river flow are inputs of both point and basin-wide rainfall estimates and forecasts. Satellite-based ensemble rainfall estimates from multiple satellite missions and sensors may provide a credible alternative in regions of poor rainfall network and to spatially better represent the rainfall patterns in the basin. Improved rainfall forecast capacity can also be provided with short-to-medium range ensemble weather forecasts using deterministic and probabilistic approach and downscaled and re-gridded for the basins. The goal is to build capability and capacity at the FMISC center in Patna so that these rainfall estimates and short (1-3 days) and medium (4-16 days) range weather ensembles forecast products can be integrated in the forecast models in B-A and Kosi basins in Bihar State.

Objective

The objective is to equip the Flood Management Improvement Support Center (FMISC) under the Water Resources Department, Govt. of Bihar with a robust 24X7 feed of customized weather ensembles of satellite based rainfall estimates and short-to-medium range rainfall forecasts, along with tools and modules that will allow FMISC to utilize these data directly into operational FMISC flood forecast models developed for the Bagmati-Adhwara (B-A) and Kosi basins. Web-based dashboards are available to FMISC so that the ensembles will also be viewable and downloadable for hydrologic model ingest. Training is also available to ensure FMISC has the understanding required to not only run these stand-alone modules, but also to maintain and modify them as needed.

To move beyond strictly gauge-to-gauge correlation river height forecasting in the B-A and Kosi basins requires the utilization of rainfall observations; and at longer lead-times, rainfall forecasts. In addition, consideration of integrating and blending data from multiple data sources increases accuracy (in both observations and forecasts), while providing needed back-up redundancy required of operational systems. Additionally, provision of forecasting products in the form of "ensembles" provides a range of possibilities the rainfall can have. This consultancy provides the

FMISC and other partners multi-model, optimized rainfall products for their hydrologic models using state of the art technologies. We note that recent research into multi-modeling rainfall forecasts, in particular, has shown that additional 1- to 3-days in forecasting skill can be gained over a single product alone

GIS Delineation

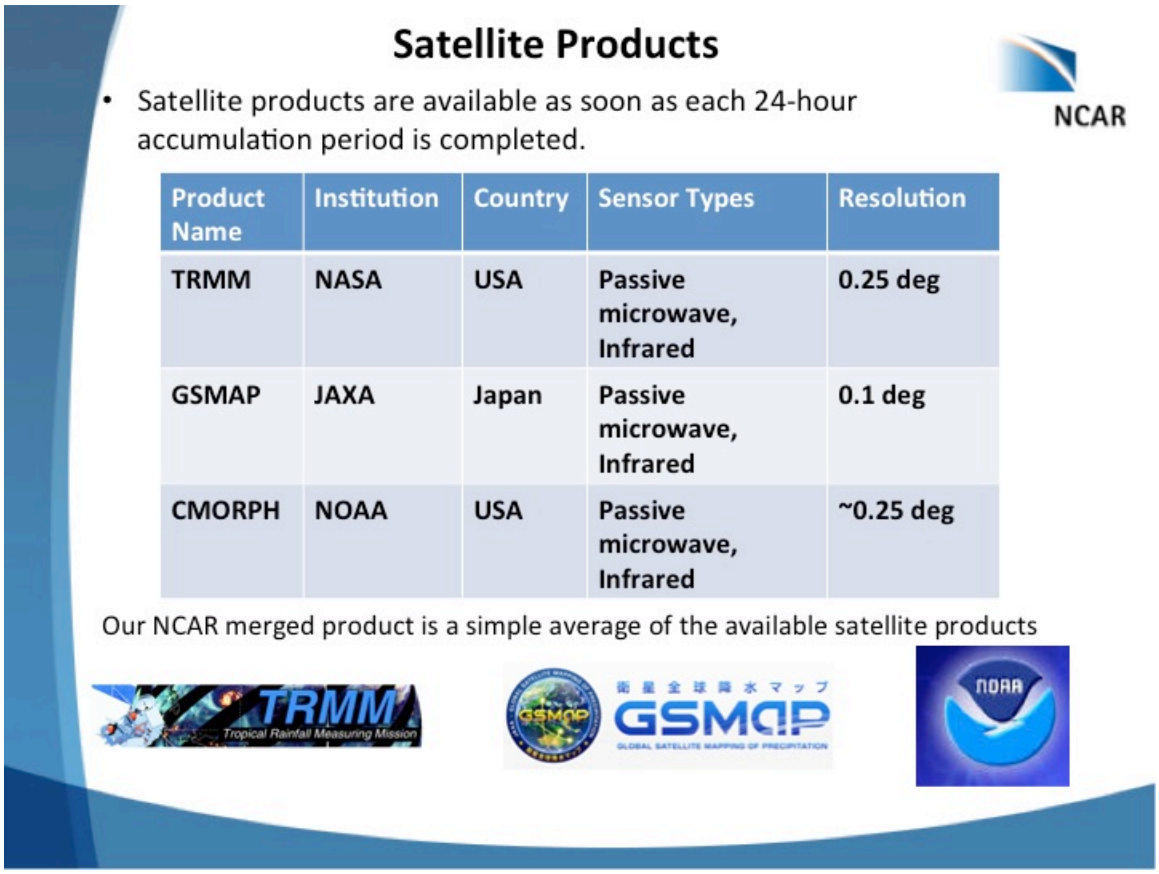
The FMISC has provided catchment shape files for this consultancy, and we have also delineated subbasins at different spatial scales in our domain. In addition, the catchment area above each of each river height gage in our doman (a necessary step for the application of the discharge forecasting models to be independently employed at each location) have been delineated using the SRTM *HydroSHEDS* data set (90m horizontal resolution).

Rainfall Data Components

In this section, we discuss the rainfall products we use in this consultancy, starting with our best estimates of observed rainfall, followed by "medium-range" ensemble rainfall forecasting, and multi-model, multi-center seasonal ensemble rainfall forecasts.

Precipitation Estimates

As our best estimate of "truth" over our domain, we use a merged blending of satellite rainfall estimates from three different global centers as shown below.



Product Name	Institution	Country	Sensor Types	Resolution
TRMM	NASA	USA	Passive microwave, Infrared	0.25 deg
GSMAP	JAXA	Japan	Passive microwave, Infrared	0.1 deg
CMORPH	NOAA	USA	Passive microwave, Infrared	~0.25 deg

Figure 1: The satellite rainfall products we used for this project.

The NASA TRMM and JAXA GSMaP satellite precipitation product (merged to a common 0.1X0.1 degree resolution, accumulations every 3hrs) satellite products are also blended with available (daily, sparse) rain gauge estimates over the same domain for our riverflow forecasts. The rain gauge data are daily 0.5X0.5 gridded

rain gauge analyses that used rain gauges reporting to the WMO GTS network (Xie et al. 1996).

Thorpex Tigge Multi-model Forecasts

Ensemble forecasting has emerged as the practical way of estimating the forecast uncertainty and making probabilistic forecasts. Based on multiple perturbed initial conditions, ensemble approach samples the errors in the initial conditions to estimate the forecast uncertainty (spread in member forecasts). The skill of the ensemble forecast shows marked improvement over the deterministic forecast when comparing the ensemble mean to deterministic forecast after a short lead time.

One source of our multi-center ensemble rainfall forecasting data comes from the THORPEX-TIGGE data set. THORPEX stands for **Th**e **O**bserving system **R**esearch and **P**redictability **Ex**periment, while TIGGE stands for the **T**HORPEX **I**nteractive **G**rand **G**lobal **E**nsemble. THORPEX is an international research programme established in 2003 by the World Meteorological Organization to accelerate improvements in the utility and accuracy of weather forecasts up to two weeks ahead. It is part of the World Weather Research Programme and is a key component of the WMO Natural Disaster Reduction and Mitigation Programme. TIGGE is a key component of THORPEX, and is a World Weather Research Programme designed to accelerate the improvements in the accuracy of 1-day to 2 week high-impact weather forecasts for the benefit of humanity by providing access to global ensemble weather forecasts from eight NWP centers around the globe (see figure below).

Unique Datasets/Software Created Thorpex-Tigge

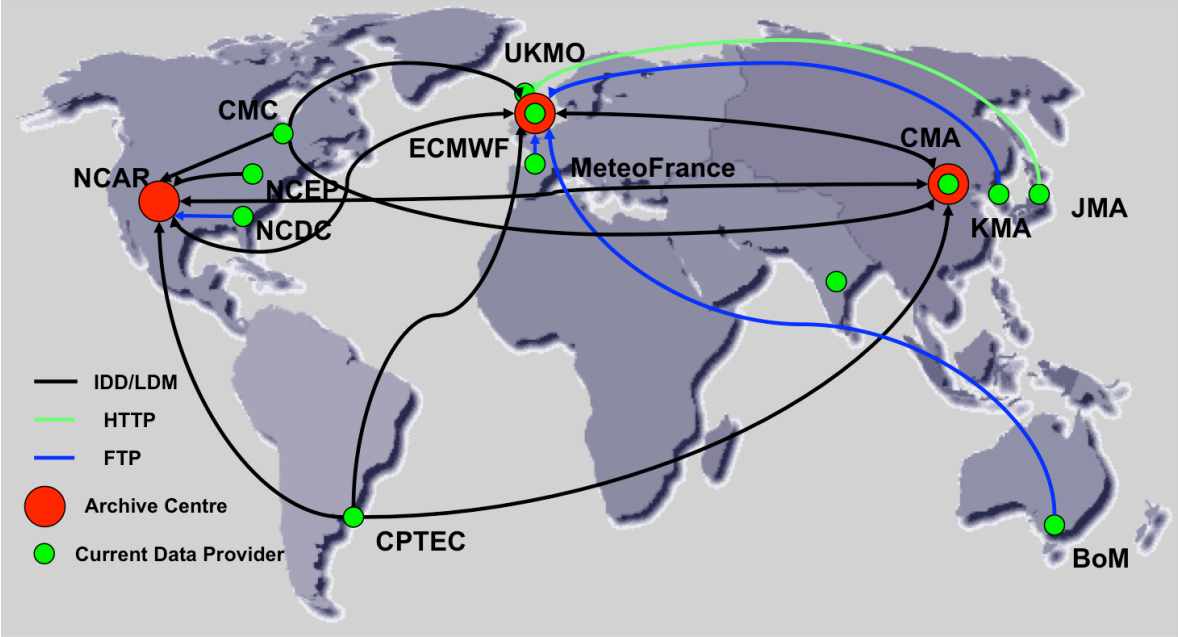


Figure 2: Green dots show the locations of the 12 NWP centers that provide data to the (nominally 3-day delayed) near-real-time THORPEX-TIGGE data archive. This project will primarily utilize forecasts from the following centers: CMA, CMC, ECMWF, UKMO, NCEP, MeteoFrance, JMA, and CPTEC

Forecasts

- Forecasts are on 2 day delay from TIGGE (The International Grand Global Ensemble).

Forecast Center	Country / Region	# of Ensemble Members	Forecast Out to:	Currently on Display
ECMWF	Europe	50	15 days	Yes
UKMO	UK	11	7 days	Yes
EC	Canada	20	16 days	Yes
NCEP	USA	20	16 days	< Dec 2015
CMA	China	14	15 days	No
CPTEC	Brazil	14	15 days	No
MeteoFrance	France	34	4.5 days	No
JMA	Japan	26	11 days	No
BoM	Australia	32	10 days	No
KMA	Korea	23	10.5 days	No

Figure 3: In our displays, we provide forecasts from the top four centers in this list. But in our multi-model rainfall and riverflow forecasts, we utilize data from the top eight centers.

We will utilize data from eight of the contributing NWP data centers for this project, which provide forecasts every day and for every 6hrs ahead out to 16 days in-advance, but are delayed two days before their release. All told, these data constitute over 300 ensemble members being generated each day. In our displays, we provide forecasts from the top four centers in the list of the figure above. But in our multi-model rainfall and riverflow forecasts, we utilize data from the top eight centers.

US NCEP and Canada CMC Forecasts

To increase redundancy and speed up forecast data availability, we also retrieve directly multi-model ensemble rainfall data from the National Center for Environmental Prediction (NCEP), which also included data from the Canadian Meteorological Center (CMC) of the Environment Canada (EC). We retrieve data from the Global Ensemble Forecasting System (GEFS) and CMC forecasts at roughly T190L28 resolution (70 km and 28 levels) with 20 members implemented globally for forecasting up to 16 days, and are available in near-real-time.

NMME Seasonal forecasts

As a "pilot" study, we also provide multi-model, multi-center ensemble seasonal rainfall forecast data the North American Multi-model Ensemble (NMME) as shown below.

NMME Seasonal Hindcasts and Forecasts

Acronym	Centre	Model	Hindcast period	Ensemble size	Lead times (months)
NCEP	NOAA-NCEP	NCEP-CSFv2	1982-2010	24	0.5-9.5
GFDL1	NOAA-GFDL	GFDL-CM2.1-aer04	1982-2010	10	0.5-11.5
GFDL2	NOAA-GFDL	GFDL-CM2.5-FLOR-A06	1982-2010	12	0.5-11.5
GFDL3	NOAA-GFDL	GFDL-CM2.5-FLOR-B01	1982-2010	12	0.5-11.5
NCAR	NCAR-UM	COLA-RSMAS-CCSM4	1982-2010	10	0.5-11.5
NASA	NASA	NASA-GMAO-062012	1981-2010	12	0.5-8.5
CMC1	Canadian MC	CMC1-CanCM3	1981-2010	10	0.5-11.5
CMC2	Canadian MC	CMC2-CanCM4	1981-2010	10	0.5-11.5

8 NMME models, 12 monthly initializations, and 0.5-8.5 month lead times during 1998-2010 are considered

Figure 4: Multi-model, multi-center ensemble seasonal rainfall forecast data available for Bihar.

Rainfall Forecasting Methods

In this section, we discuss the methods we employ to generate calibrated and optimally merged, multi-center ensemble rainfall forecasts catchment-integrated over our region for FMISC use. We start with a discussion on our bias correction algorithm, which removes biases across the whole probability distribution function (PDF), followed by the calibration method we employ, that allows for calibrating individual ensemble forecasts, along with the merging of multi-center, ensemble forecast, to generate a final set of optimal ensembles for hydrologic applications. The appendices discuss aspects of these techniques in more detail.

Bias Correction -- Quantile-to-Quantile Mapping

In this subsection, we discuss a somewhat similar approach to model fitting as described by QR below, called "quantile-to-quantile mapping (q-to-q), but with a different twist such that this approach is less stringent on model fitting when the regressand (observed values) may be error-prone themselves. More technical details of this approach are discussed in **Error! Reference source not found.** below.

Generally, all practical rainfall or hydrologic models require some form of calibration arising from reasons such as simplified parameterization of rainfall processes and incomplete knowledge of watershed properties and the necessary parameterization of transport at unresolved scales. In the case of riverflow forecasts (discussed further below), through the hydrologic model parameter calibration process, some of the additional errors in the watershed inputs (rainfall, in particular) can also be implicitly reduced (e.g. runoff errors arising from an over-bias in rainfall forcing can be implicitly minimized by increasing the parameterization of evapotranspiration). However, such bias reduction through calibration does not occur if there are relative errors between more than one input of the same physical quantity. In particular, using both weather forecasts and observationally-based estimates of precipitation concurrently to generate unbiased discharge forecasts, requires that these two data sources maintain their statistical similarity such that there are no relative biases between their statistical moments (mean, variance, and skewness, in particular). Note that quantile regression (discussed below) could also provide the required bias reduction on a forecast-day-by-forecast-day bases. However, because we feel that the merged satellite precipitation product (i.e. "observations") inherently have their own error, we hypothesize it is more optimal to only utilize a more "global" statistical adjustment of the TIGGE precipitation forecasts to require them to only be drawn from the same probability distribution function as the observations. Quantile-to-quantile mapping is designed to exactly do this.

Calibration and Multi-modeling -- Quantile Regression

In this subsection, we briefly discuss a statistical forecasting and correction approach we introduced into this project that we rely on heavily in this project, called "quantile regression" (QR). In particular, we use this approach in both our ensemble rainfall calibration for ensembles from individual centers and multiple centers; and similarly, we apply QR in a similar way for our riverflow foecasts (discussed further below). We defer more technical details of this approach to below. Quantile regression has been little used in the atmospheric and hydrologic communities. We introduced this approach for this project, relaying on it heavily to generate ensemble forecasts and forecast error corrections. Similar to simple linear regression which minimizes the square error of a model to-be-fit, QR instead minimizes the absolute error, but done for each "quantile" of the distribution one is interested in. The result is that multiple models are generated, one for each quantile, and each provides a reliable probability of what values the rainfall (or river flow) will exceed during the forecasting period (such that "flat histograms" of the forecasts are produced – described further in an appendix below). As we apply it for our forecasting application, we ensure that the resultant ensemble rainfall forecasts (or discharge forecasts) are no worse than the climatological average flow after the model fitting process.

Rainfall Forecast Products and Evaluation

In this section, we provide an evaluation of our rainfall products we are generating for this project, starting with a look at the biases of the forecasts, the inherent skill of each forecasting center's ensemble forecasts. We then follow up with this by showing the benefits of combining all weather centers' ensemble forecasts together to produce mutli-model ensemble rainfall forecasts for hydrologic applications. We conclude by providing a web-based display to compare past historical rainfall events both observed and forecasted.

Bias and Anomaly Correlation Comparisons

In this subsection, we provide an evaluation of our rainfall products we are generating for this project, starting with a look at the biases of the forecasts, and then the inherent skill of each forecasting center's ensemble forecasts.

Below in the figure we see comparisons of the rainfall amounts for August average rainfall from 2011 to 2015 for four of our weather forecast centers 0 to 4 day lead-time forecasts over the Bagmati, Kosi, and Adhwara basins, noting the similarities, but also distinctive differences in both the amounts and locations.

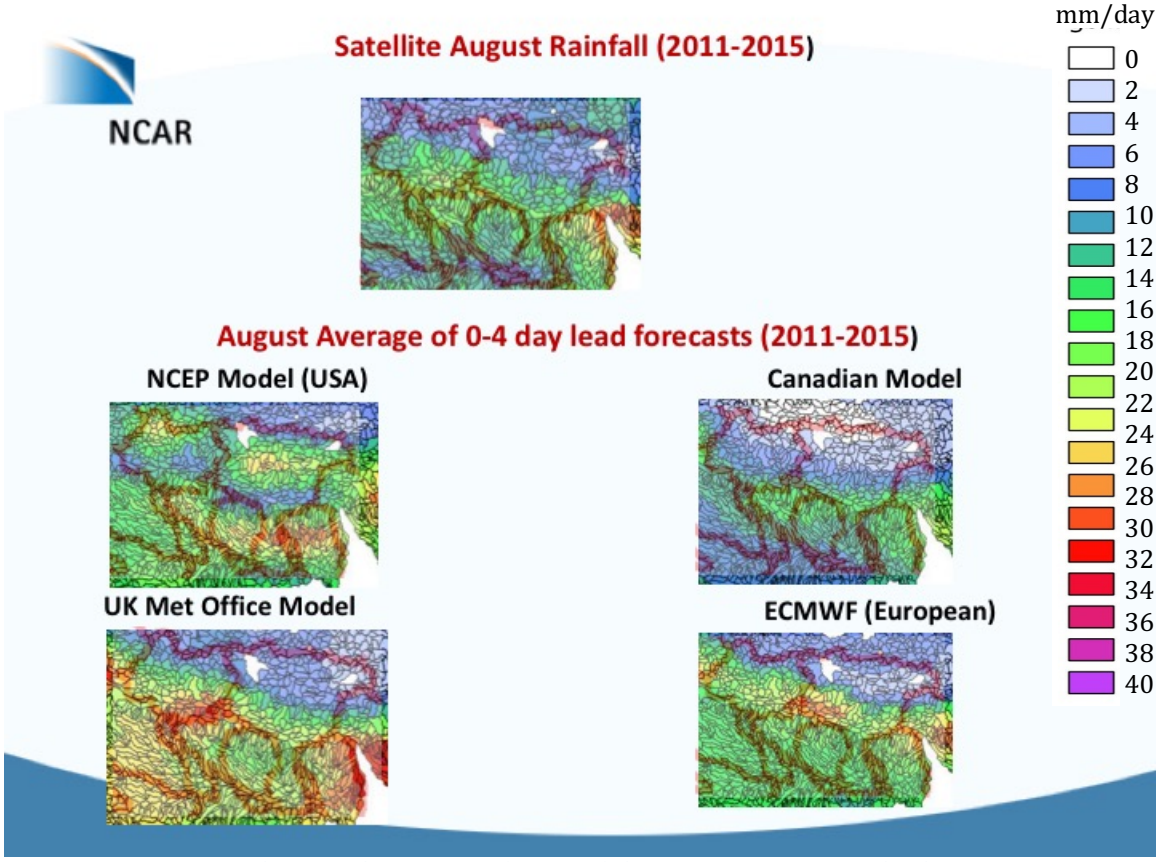


Figure 5: omparisons of the rainfall amounts (mm/day) for August average rainfall from 2011 to 2015 for four of our weather forecast centers 0 to 4 day lead-time forecasts over the Bagmati, Kosi, and Adhwara basins, noting the similarities, but also distinctive differences in both the amounts and locations.

The similarities and differences in the fields become more pronounced when, looking at their direct differences (biases: subtracting the forecasts from the observations), as seen in the figure below. If one were to choose a forecast model based only on these biases, one might chooseLooking at these differences in the biases, one might choose for the Bagmati (first to last) the ECMWF, Canada, UKMet, and then the NCEP models. For the Adhwara, one would choose Canada, ECMWF, UKMet, and then the NCEP. For the Kosi: Canada, UKMet, ECMWF, then NCEP. Note the change in ranking over such relatively short distances!

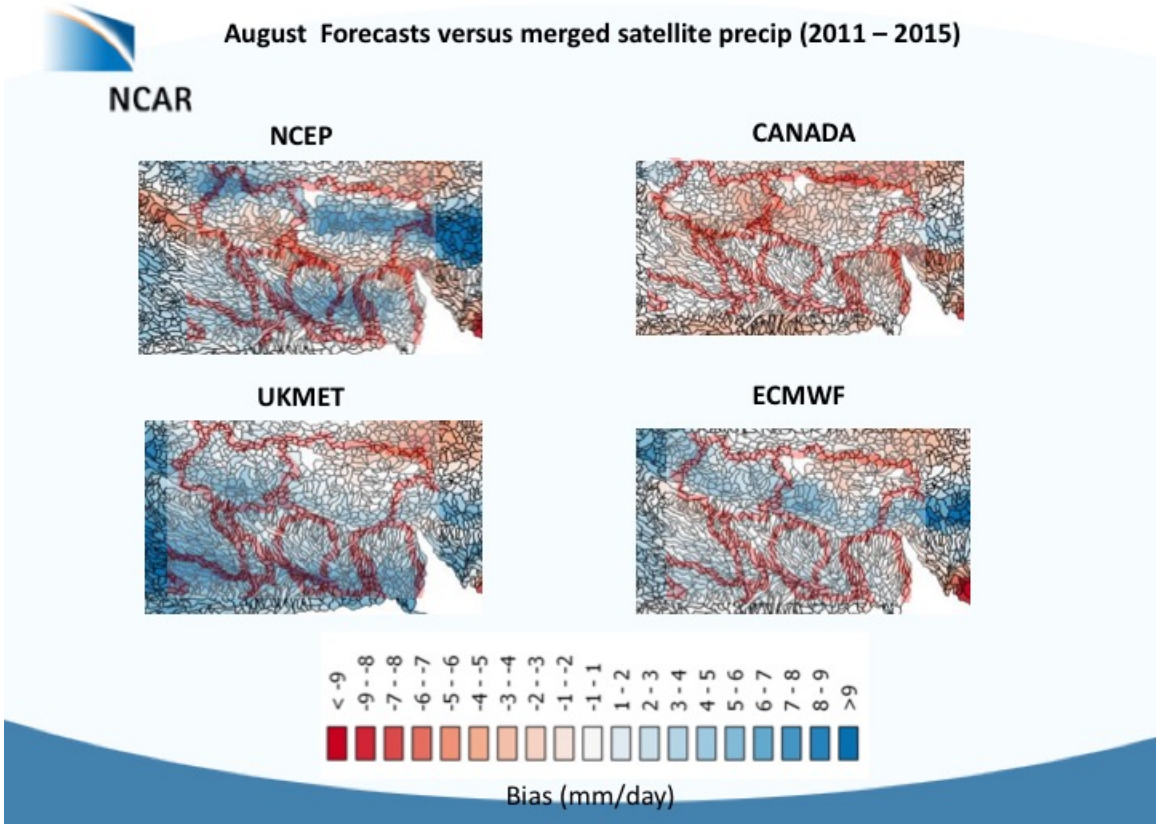


Figure 6: biases (subtracting the forecasts from the observations, in mm/day) of the values seen in the figure above.

However, now let's examine the skill in the forecasts themselves, here defined as their anomaly correlation (the correlation of both the observations and of the forecasts, both differenced from "climatology", as seen below in the figure. Now, the ranking of the models one might choose for the Bagmati would be: ECMWF, NCEP-UKMet (both equal), and then Canada. For the Adhwara, one would choose: ECMWF, NCEP, then Canada-UKMet (both equal). For the Kosi: ECMWF-NCEP-UKMet (all equal), the Canada. Comparing these rankings from those of the biases, we see very different result. If one were to choose the value of the forecasts based solely on the biases as an indicator of the forecast (correlation) skill, one would be misled! What this points out is the importance of being clear on metric one most cares about for your application before choosing a "best" model.

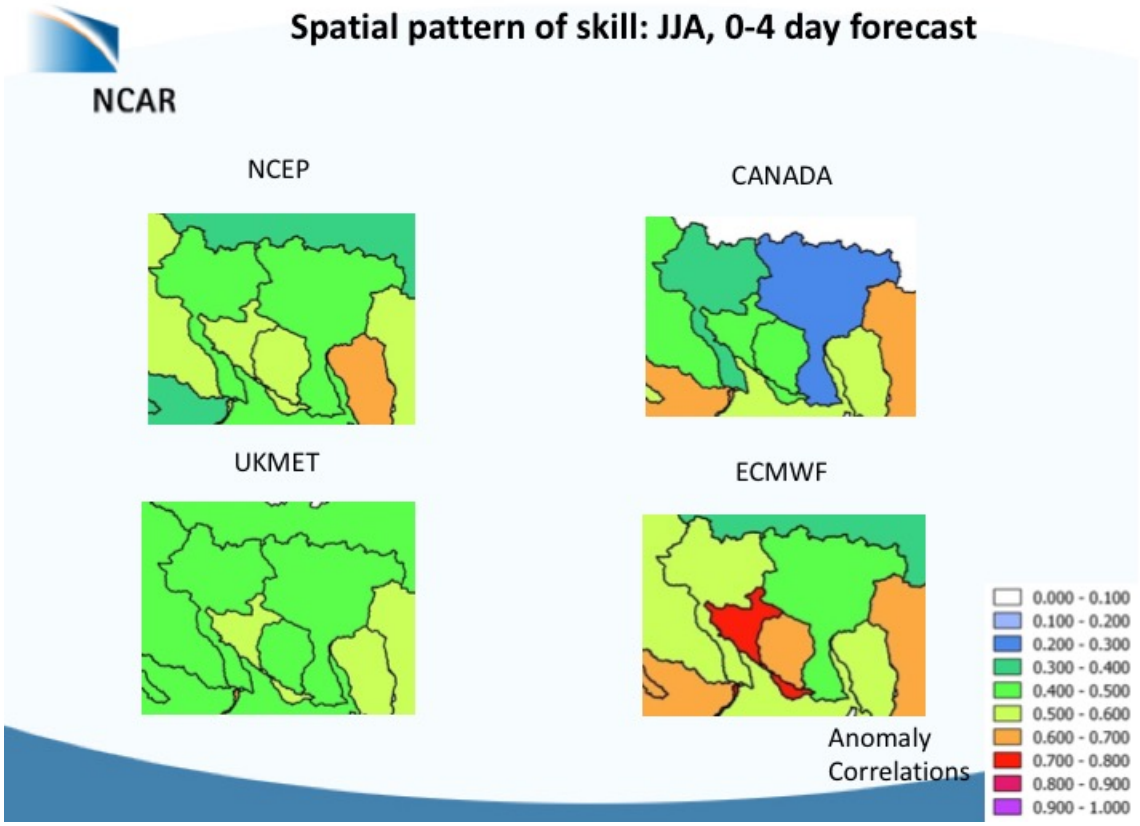


Figure 7: Anomaly correlation skill (the correlation of both the observations and of the forecasts, both differenced from "climatology") in the four centers' forecasts.

What is interesting to in the figure above, just as with the biases, there is a significant variation in forecast skill with location, both across different models, and within the same model. However, averaging the correlation skill across the region's catchments (seen in the figure below), we see that for both the 24hr-average and 5-day average forecasts, ECMWF is superior overall, followed by generally equal rankings for the Canada, NCEP, and UKMet models, and these rankings persist for the first 10days of 24hr average forecasts. For the 5day average forecasts, we see the same result as for the 24hr-averaged forecast for , 0-4day lead-time window; however, for the 5-9day and 10-15day lead-time windows, we see an interesting shifting of the rankings

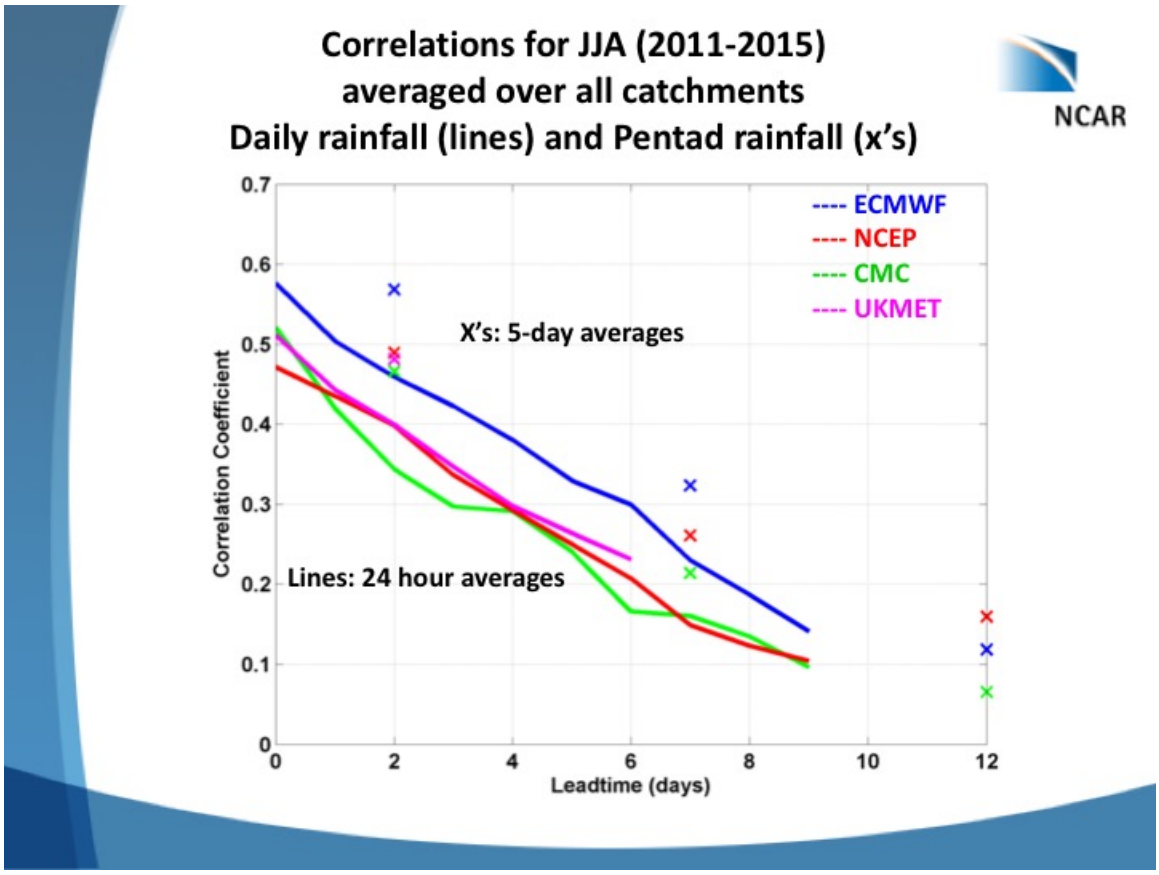


Figure 8: correlations of the Canada (CMC), ECMWF, NCEP (US), and UKMet office forecasts when averaged over the region's catchments, as a function of lead-time and averaging window (24hr and 5day). Note that the UKMet office foreasts only extend out 7 days.

Multi-model Benefits

In this subsection, we show the benefits of combining all weather centers' ensemble forecasts together to produce multi-model ensemble rainfall forecasts for hydrologic applications, using the catchment-integrated rainfall estimates upstream of our river height gauging locations available over the region. What we find is that our calibration routine provides output ensemble forecasts that are statistically "correct" and that have "trust" in their likelihoods. Also, overall, ECMWF does the best job, but a combination of other centers can match ECMWF's performance. Also, the full multi-model ensemble forecast has roughly an additional 2-days of increased skill over just using the best center's (ECMWF) forecasts.

In this application, we use quantile regression (QR; as discussed above and below in the Appendix) to calibrate ensemble forecasts from each of our individual weather forecasting centers, and a combined multi-model forecasts (in the figures below, we

show forecasts from just five of our eight centers: CMA, CMC, CPTECH, ECMWF, NCEP; with each center having a different number of ensembles). Our output in each case is a resultant 9-member ensemble forecast. Also in this application, the regressors we use for QR are the ensemble mean forecast from each center, and the ranked ensemble member corresponding closest to the ensemble quantile we are calibrating for. To assess the skill of the multi-model forecasts, we use a number of verification metrics – see the appendix below for details.

Below in the figure we show time-series of quality-controlled stream gauge heights, with arrows pointing to the gauge locations: Bagmati (station 007-mgd4ptn) gauging station, and a Lower Ganges (station 029-mgd5ptn) gauging station. Above these two locations we catchment-integrate rainfall fields, and calibrate these values as shown in the rest of this section.

.

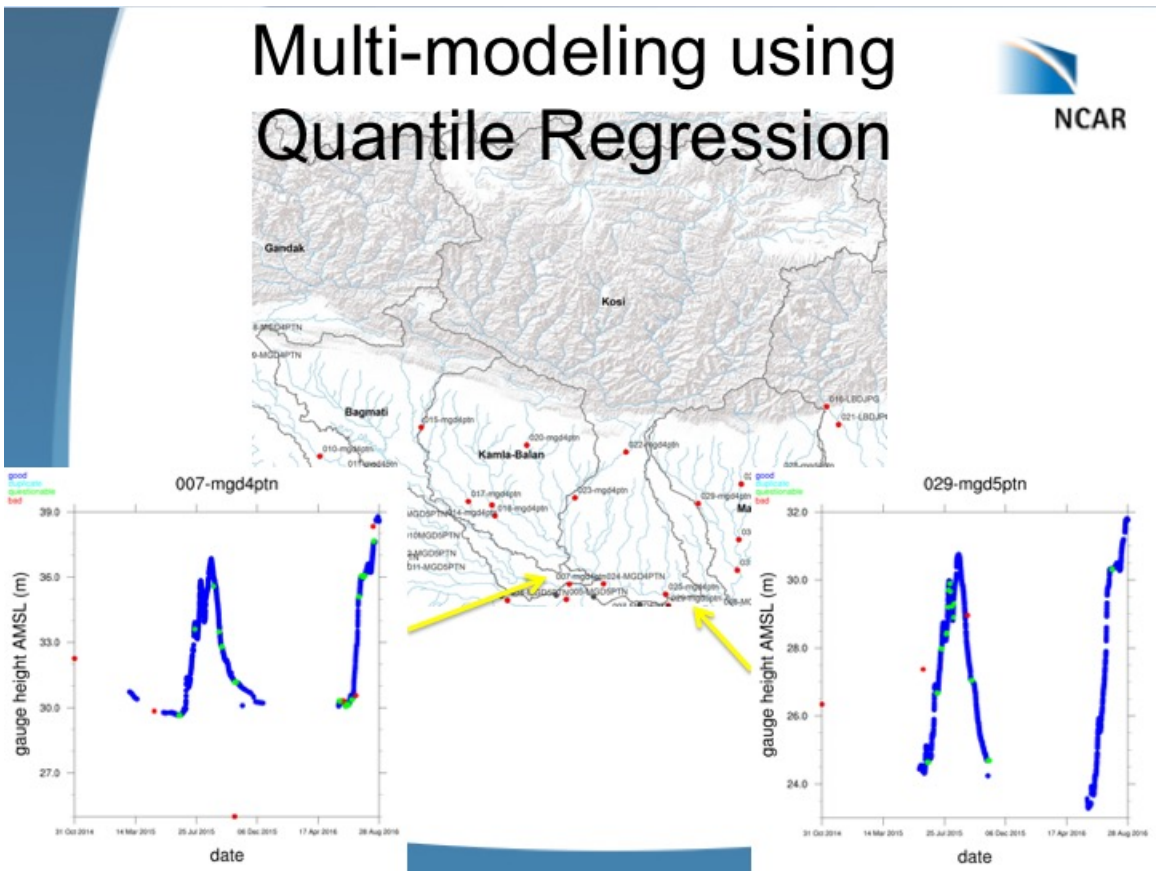


Figure 9: time-series of quality-controlled stream gauge heights, with arrows pointing to the gauge locations. Above these two locations we catchment-integrate rainfall fields, and calibrate these values.

Below we show 5-day lead-time calibrated time-series of the catchment-integrated ensemble rainfall forecasts for our five centers (colored lines), along with the observed (satellite and rain gauge combined rainfall; black line), along with multi-model forecasts for two centers (CMC and NCEP) and for all 5 centers.

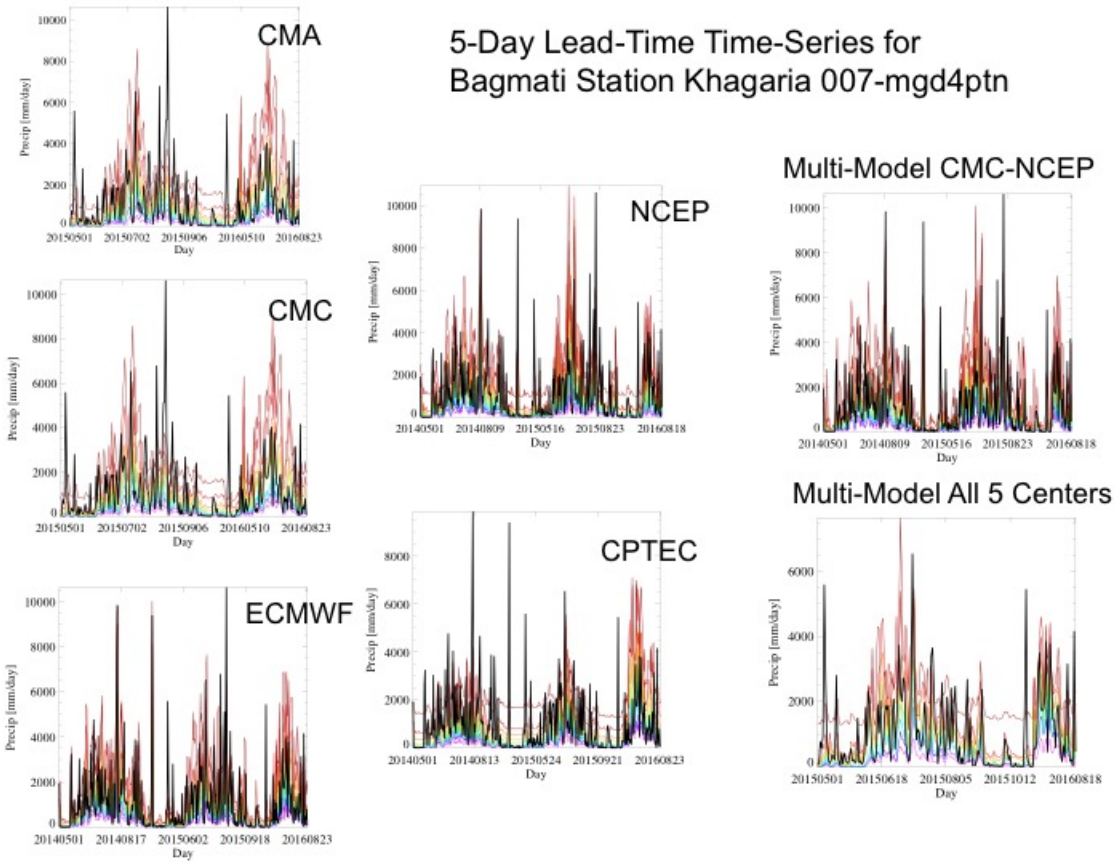


Figure 10: 5-day lead-time calibrated time-series of the catchment-integrated ensemble rainfall forecasts for our five centers (colored lines), along with the observed (satellite and rain gauge combined rainfall; black line), along with multi-model forecasts for two centers (CMC and NCEP) and for all 5 centers.

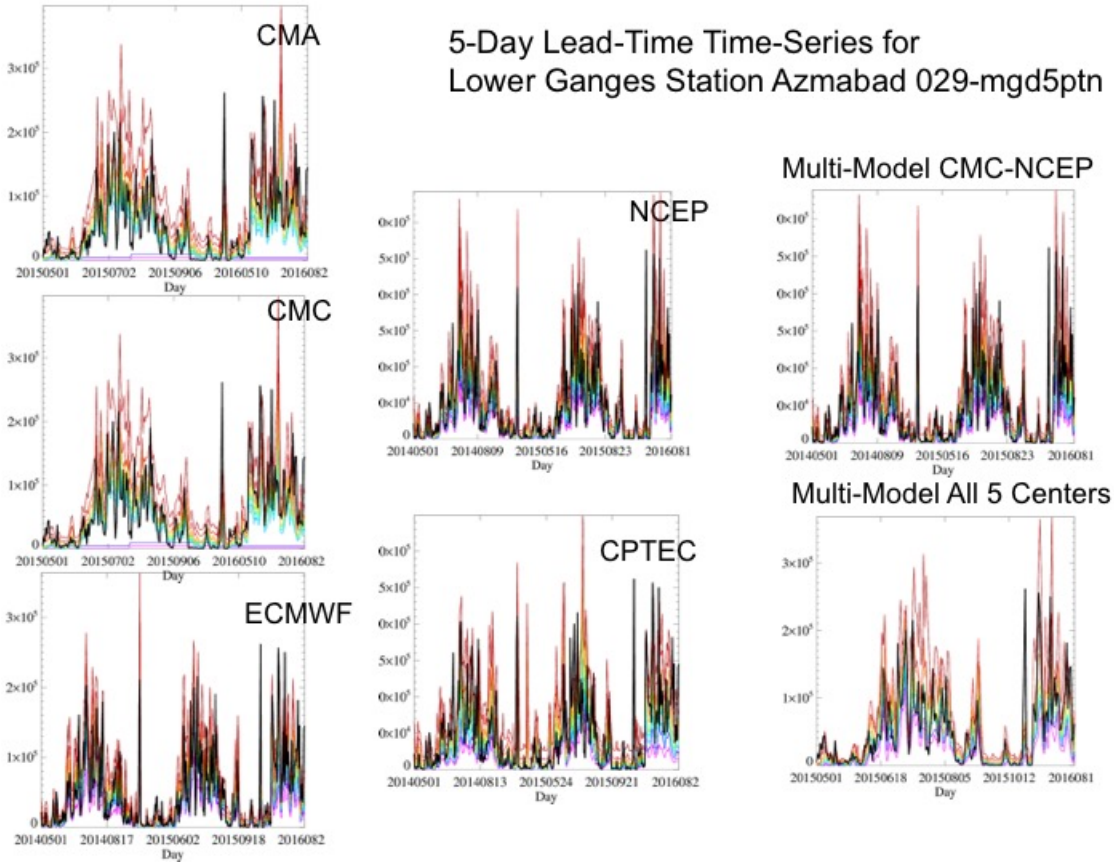


Figure 11: Same as the figure above, but for the catchment-integrated rainfall above gauge 029-mgd5ptn.

To assess the skill of the multi-model forecasts, we use a number of verification metrics – see the appendix below for details. Below, we show rank histograms (as a verification metric of how "trust-worthy" are the ensemble forecast's range – see appendix) of the 5-day lead-time catchment-integrated rainfall above our two gauging locations compared to observed rainfall. For rank histograms, the "flatter the better" – what we see below is that essentially the histograms are "flat" (except for some statistical randomness).

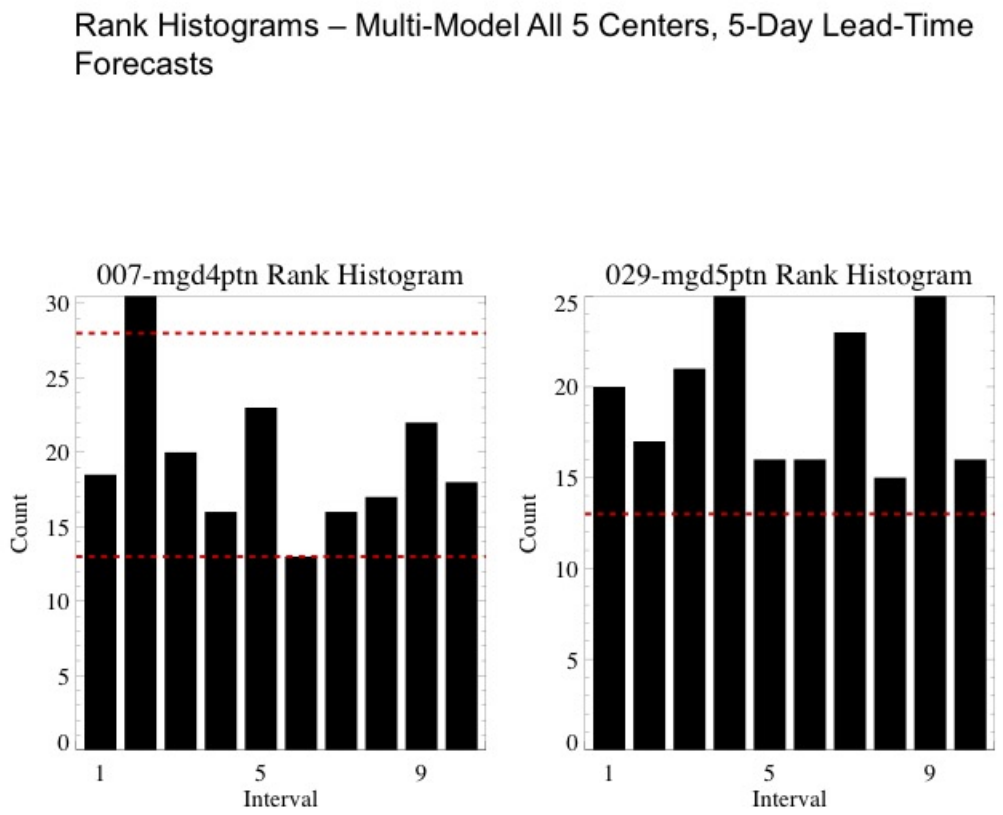


Figure 12: rank histograms (as a verification metric of how "trust-worthy" are the ensemble forecast's range – see appendix) of the 5-day lead-time catchment-integrated rainfall above our two gauging locations compared to observed rainfall. For rank histograms, the "flatter the better" – what we see below is that essentially the histograms are "flat" (except for some statistical randomness).

To further assess the skill of the multi-model forecasts, we compare the same data (5-day lead-time catchment-integrated rainfall above our two gauging locations compared to observed rainfall) using the Brier skill score metric (again, see the appendix below for details).

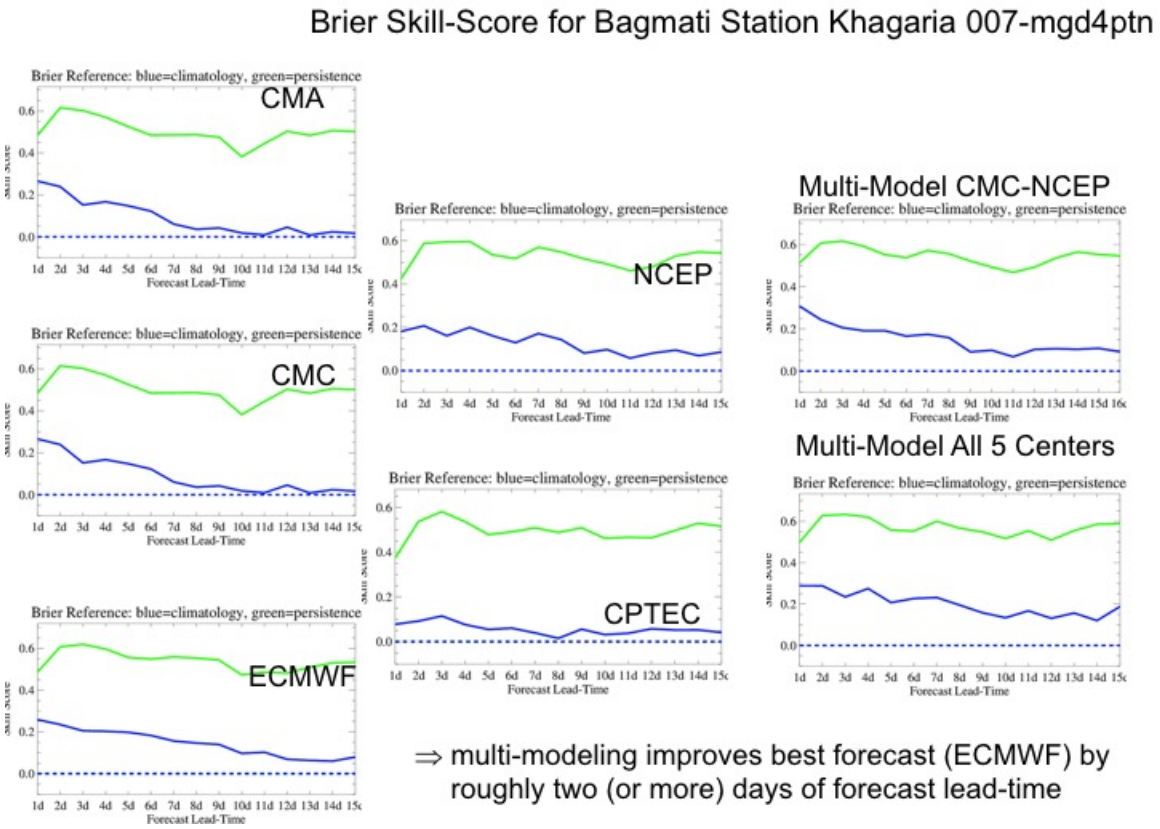


Figure 13: Brier skill scores as a function of lead-time going out 1- to 15-days in advance for each of our 5 centers individually, and combined into multi-models. The blue line is the brier skill score relative to a "persistence" forecast, and the green lines are relative to "seasonal climatology". Note the overall superiority of the 5-center multi-model for all lead-times.

What we see in both figures above (for gauge 007-mgd4ptn) and below (for gauge 029-mgd5ptn) is that generally, ECMWF is the best forecast for all lead-times, but that the combination of CMC and NCEP can strongly compete (and possibly out-perform) the single best model (ECMWF), while the 5-center multi-model generally provides the best possible forecast for all lead-times.

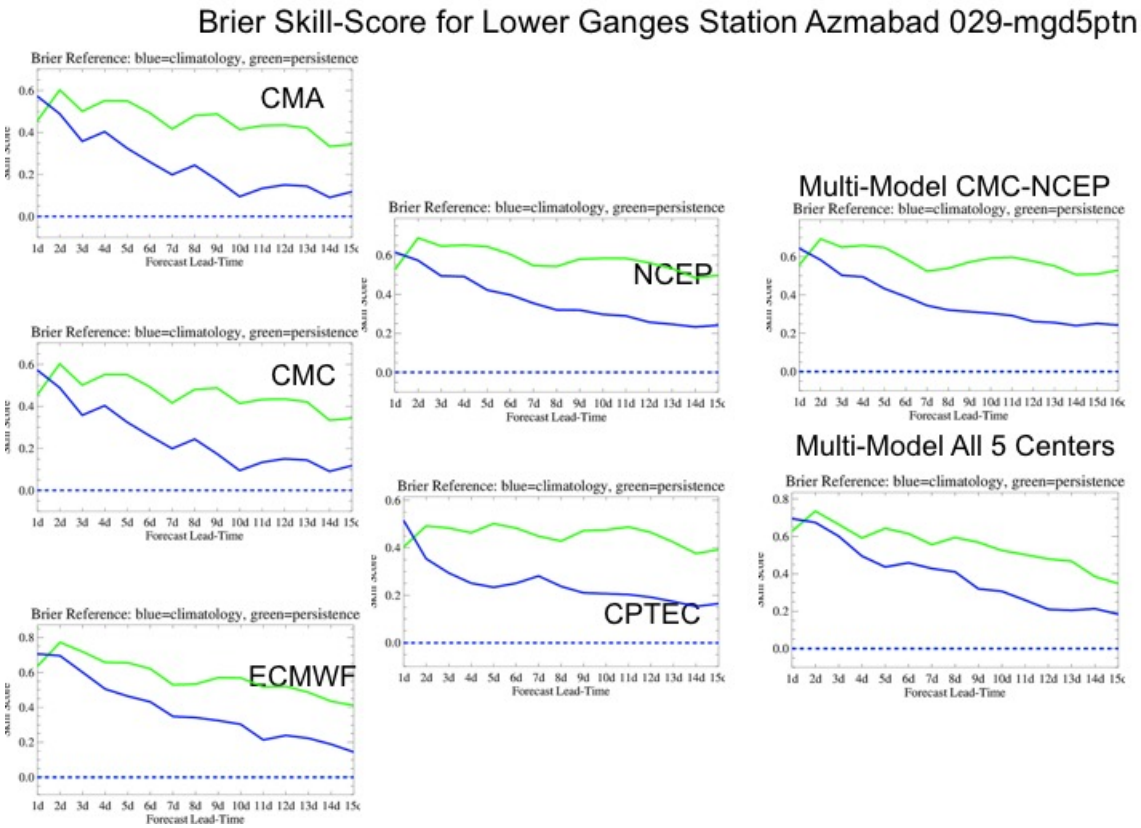


Figure 14: Same as the figure above, but for the catchment-integrated rainfall above gauge 029-mgd5ptn.

To further assess the skill of the multi-model forecasts, we compare in the two figures below the same data (5-day lead-time catchment-integrated rainfall above our two gauging locations compared to observed rainfall) using the RMSE skill score metric (again, see the appendix below for details), similar to what was shown in the two figures above for the Brier skill score.

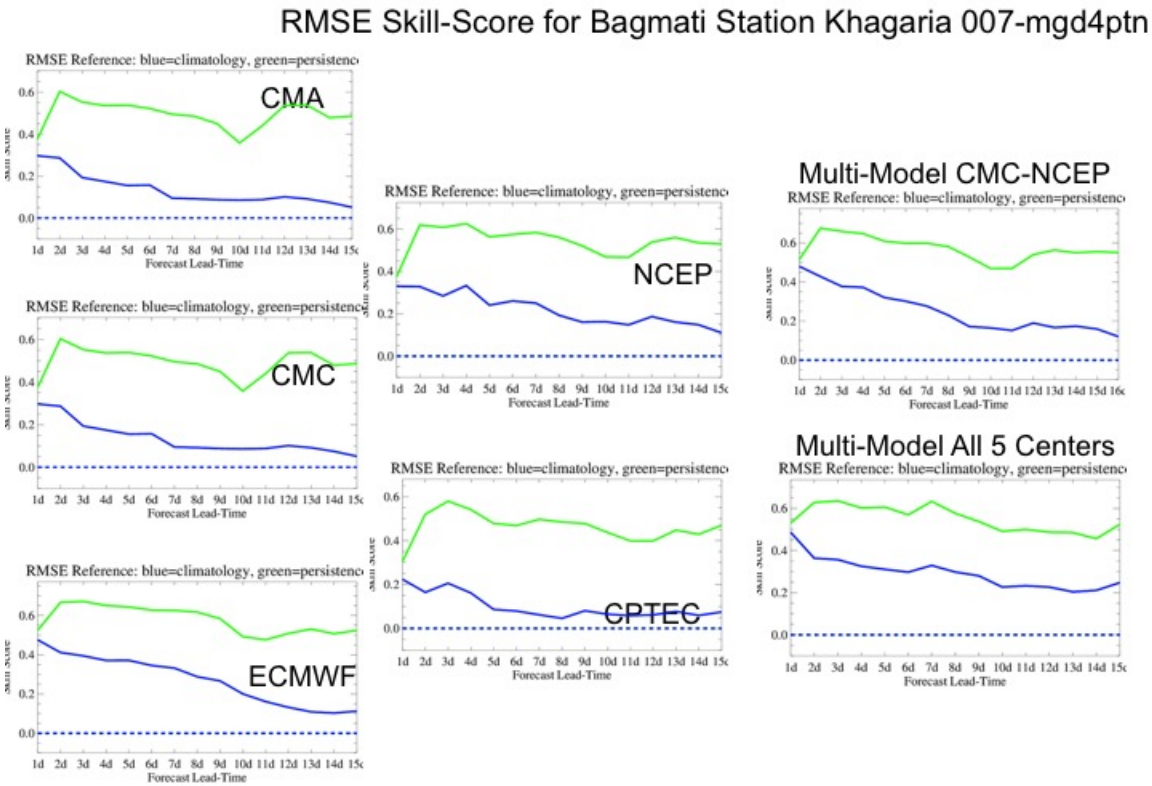


Figure 15: RMSE skill scores as a function of lead-time going out 1- to 15-days in advance for each of our 5 centers individually, and combined into multi-models. The blue line is the RMSE skill score relative to a "persistence" forecast, and the green lines are relative to "seasonal climatology". Note the overall superiority of the 5-center multi-model for all lead-times.

For the RMSE skill score (as we did for the Brier skill score earlier) is that we see in both figures above (for gauge 007-mgd4ptn) and below (for gauge 029-mgd5ptn) is that generally, ECMWF is the best forecast for all lead-times, but that the combination of CMC and NCEP can strongly compete (and possibly out-perform) the single best model (ECMWF), while the 5-center multi-model generally provides the best possible forecast for all lead-times.

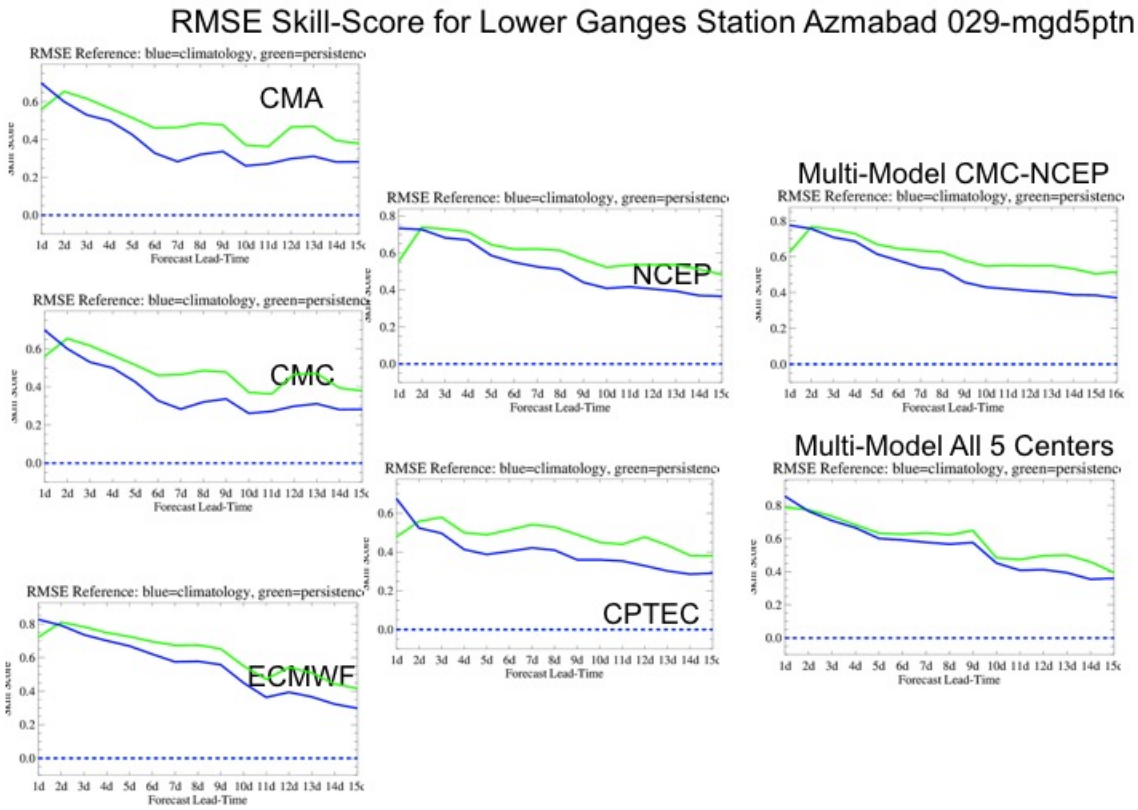


Figure 16: Same as the figure above, but for the catchment-integrated rainfall above gauge 029-mgd5ptn.

Indirectly we can also assess the skill of each weather center's forecasts by analyzing the percentage of usage of each center's forecasts in the calibration process. As seen in the figure below, we see that for both of our basins above the gauging locations, ECMWF is most used, as well as for all the basins in our region (right-most panel). However, the relative contributions from the other centers varies by location.

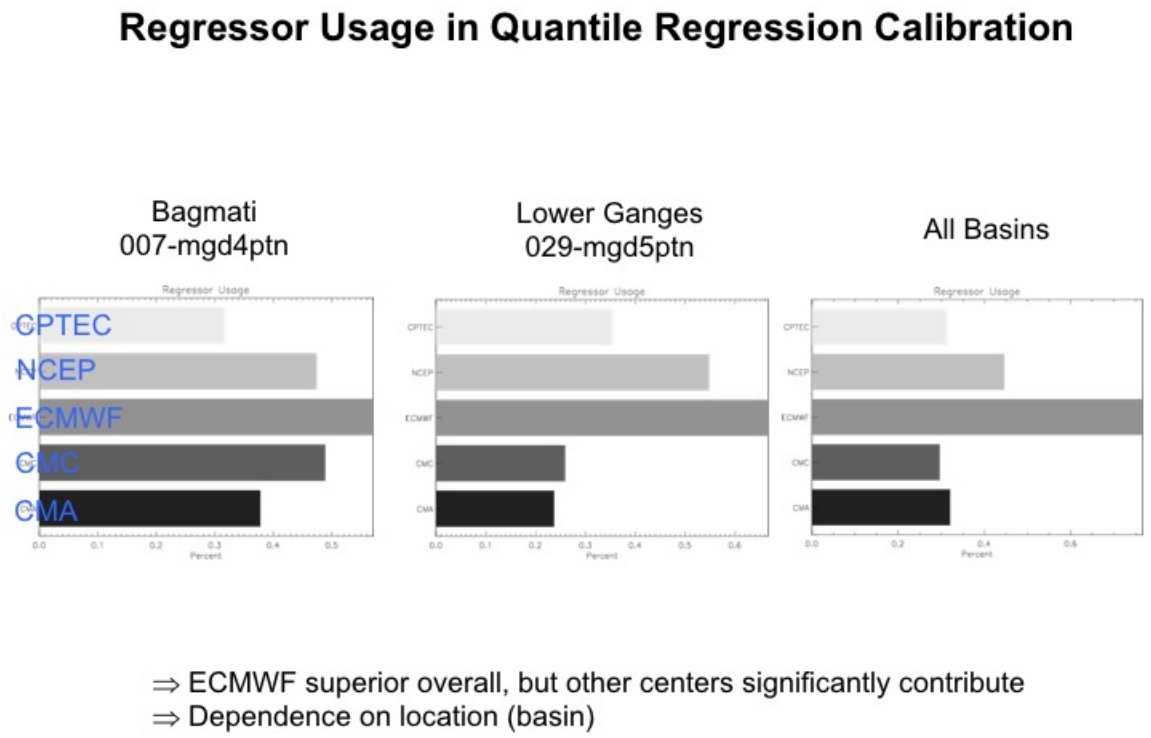


Figure 17: percentage of usage of each center's forecasts in the calibration process. For both of our basins above the gauging locations (left and center panels), ECMWF is most used, as well as for all the basins in our region (right-most panel). However, the relative contributions from the other centers varies by location.

In summary, we find that over our region that ECMWF generally outperforms other centers after postprocessing for a variety of verification metrics. However, the combination of NCEP and CMC (Canada) can reach similar combined skill to ECMWF (e.g. for our two example basins). Multi-modeling roughly gains two days of forecast lead-time as a rule-of-thumb. Finally, we find in general, that the center with the best forecast skill is strongly location/catchment-dependent

Web-based Archive Comparisons

Here in this subsection, we conclude by providing a web-based display to compare past historical rainfall events both observed and forecasted as a further graphical way to see skill assessed for each weather center. Note that we reserve a more detailed discussion further below in the section, "Dashboard Framework and Data Access".

Under the "Rain Archives" tab of our project's web site (http://bihar.rap.ucar.edu/archive/) are comparisons with the July and August 2017 severe rain and flooding events. Seen below we show comparisons of 24-hr accumulated rainfall from July 9th to July 10th for 5-day lead-time ECMWF ensemble mean forecast compared to satellite estimates (top panel forecasts, bottom panel merged satellite rainfall). This plot generally shows good agreement, but with variations at small spatial-scales.

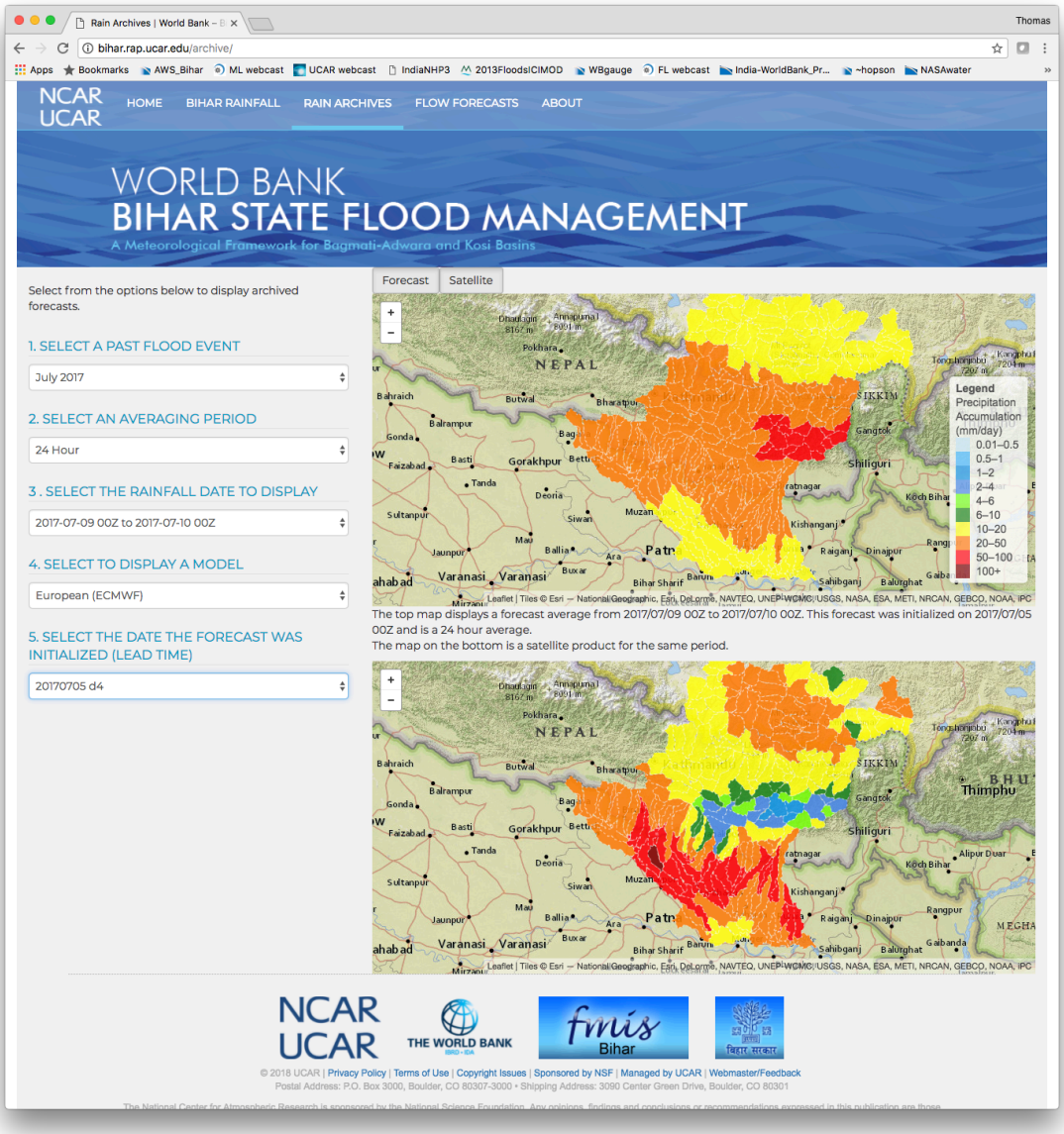


Figure 18: comparisons forecasts (ECMWF in this case) and observations (satellite rainfall) as seen during the July 2017 flooding events from our "archive" web page (http://bihar.rap.ucar.edu/archive/). Shown are comparisons of 24-hr accumulated rainfall from July 9th to July 10th for 5-day lead-time ECMWF ensemble mean forecast compared to satellite estimates (top panel forecasts, bottom panel

merged satellite rainfall). This plot generally shows good agreement, but with variations at small spatial-scales.

Comparing the above figure with the below, which is the same plot but for 5-day accumulated rainfall from July 9th to July 14th for ECMWF forecasts that was initialized on July 9th , and satellite rainfall accumulated over the same 5 days – we still see general agreement, but perhaps a bit better skill, although with some displacement discrepancies in the fields.

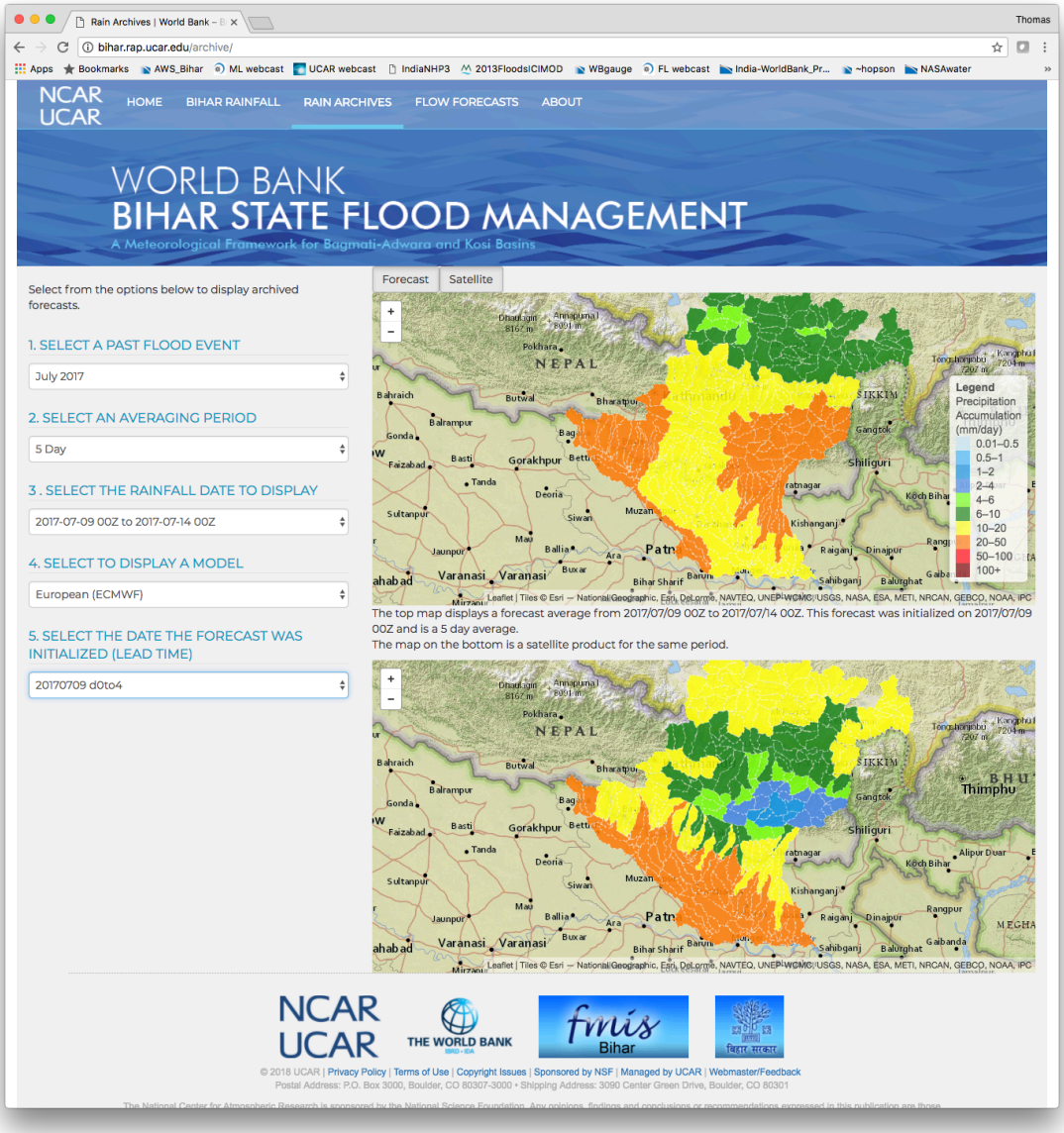


Figure 19: Similar to the figure above, but for 5-day accumulated rainfall starting on July 9th, and with ECMWF forecasts initialized on the same day.

Riverflow Data Components

In this section, we discuss the observational data that are used in the riverflow forecasting system in operations for Bihar and the FMISC.

River Gauge Measurements

Real-time river and reservoir stage data are currently collected from the Flood Forecast Monitoring (FFM) Directorate of the Central Water Commission (CWC) from www.india-water.gov.in/ffs. Data for each station of interest are downloaded every hour from the web services using the site's custom protocol. The stage level, precipitation, and trend are stored in a MYSQL geospatial relational database for easy access by scientific and display processes. Those records are retrieved for use in initializing the forecast model, generating forecast plots, and populating the web display. While data are currently only retrieved for stations in the Ganges and Brahmaputra basins, the process could easily be scaled to include all stations around the country. Current records extend from February 28th of 2015 to the present.

Several services have been created to retrieve subsets of the data in the database and return that data in a specific format. The web display's service is of particular utility because it returns the data prepared in a standard GeoJSON format. Web services such as this are the fundamental building blocks of data interoperability. While this particular service does not support the full Web Feature Service (WFS) specification, adding such capability would immediately allow it to be used with a wide array of WFS clients: model initialization, data visualization, and system monitors.

Satellite River "Width" Measurement

The Dartmouth Flood Observatory (DFO), in partnership with the Joint Research Council, is providing multi-site estimates of river width for both the Ganges and Brahmaputra (see JRC-Ispra, http://www.gdacs.org/floodmerge/ and DFO, http://www.dartmouth.edu/~floods/). These estimates are publically-available at http://old.gdacs.org/flooddetection/. Actual measurements are of upstream relative changes in emissivity due to changes in river width (microwave imagery visible through cloud cover). This signal is then effectively "advected" to downstream forecasting locations (see Stage-Discharge Relationship Flood Forecasting subsection below).

.

Quality Control

In this subsection we present our automated quality control method implemented for river stage, discharge, and satellite-derived "river width" time series data. For operational flood forecasting over a large domain with multiple stage reading sites, automated quality control (QC) of measurements is essential to providing accurate forecasts, without spurious values being ingested and corrupting the forecasting algorithms. We have been collecting stage measurements and archiving them into a MYSQL database. However, there are appreciable errors found in almost every one of these sites. To see examples, go to the "Flow Forecasts" tab found under our project website, http://bihar.rap.ucar.edu/flow/ (discussed further below). Clicking on a visible station (dots), shows the original time series of record for this station in the lower part of the screen, with the quality-controlled time series shown at right. Notice that in most of these QC figure there are red points flagged as "bad", turquiose points flagged as "duplicates", and green points flagged as "questionable" (which is given a QC number varying between 0-100). Please see the appendix below for more details on our quality-control procedures. Immediately below, we list the questionable data features we identify and how they are flagged.

QC score	Meaning
100	Good
0-99	questionable rate of change or possible stuck sensor
-1	Missing
-2	Isolated
-3	too high/low
-4	extreme rate of change
-5	stuck sensor
-11 -12	Duplicate missing neighbor (satellite QC only)

Table 1: QC score values, and their meaning.

River Forecasting System

The system developed and implemented operationally for this project is restructured such that multiple forecasting methods can be operated in parallel (and combined into a "multi-model" at later stages of the system). A summary of the data sources used in the forecasting system is as follows: in situ river stage measurements; gridded satellite and raingauge estimates from 3 sources, ensemble weather forecasts from 8 weather centers; and upstream river stage measurements, upstream river "width" esimates. Multi-modeling (combination approaches) these forecasts still allowing the separate pieces to be combined to optimize forecast skill.

As shown in the figure below we discuss the overall forecasting structure used in this project and the subsequent forecasting skill analysis. Shown in this figure below is the flow chart of the components used in this project. This flow chart is similar to that used in the CFAB project (Hopson and Webster 2010; Webster et al 2010), with ensemble weather forecasts, satellite precipitation estimates, and river flow measurements used as "I. Initial Data Input" at the top. However, in this application, we introduced forecasts from 8 centers (TIGGE forecasts, now going out to 16-day lead-times -- previously 10days for the CFAB system), an additional satellite precipitation estimate (JAXA), and new upstream river flow information provided by Central Water Commision upstream gages, and DFO-JRC upstream river widths. These new upstream data sets required development of new algorithms (shown schematically in the "Stage-Discharge Relationship" box under "III. Hydrologic Modeling", and discussed below in a lower section). As a result, we are producing independent ensemble river discharge forecasts based on these different approaches.

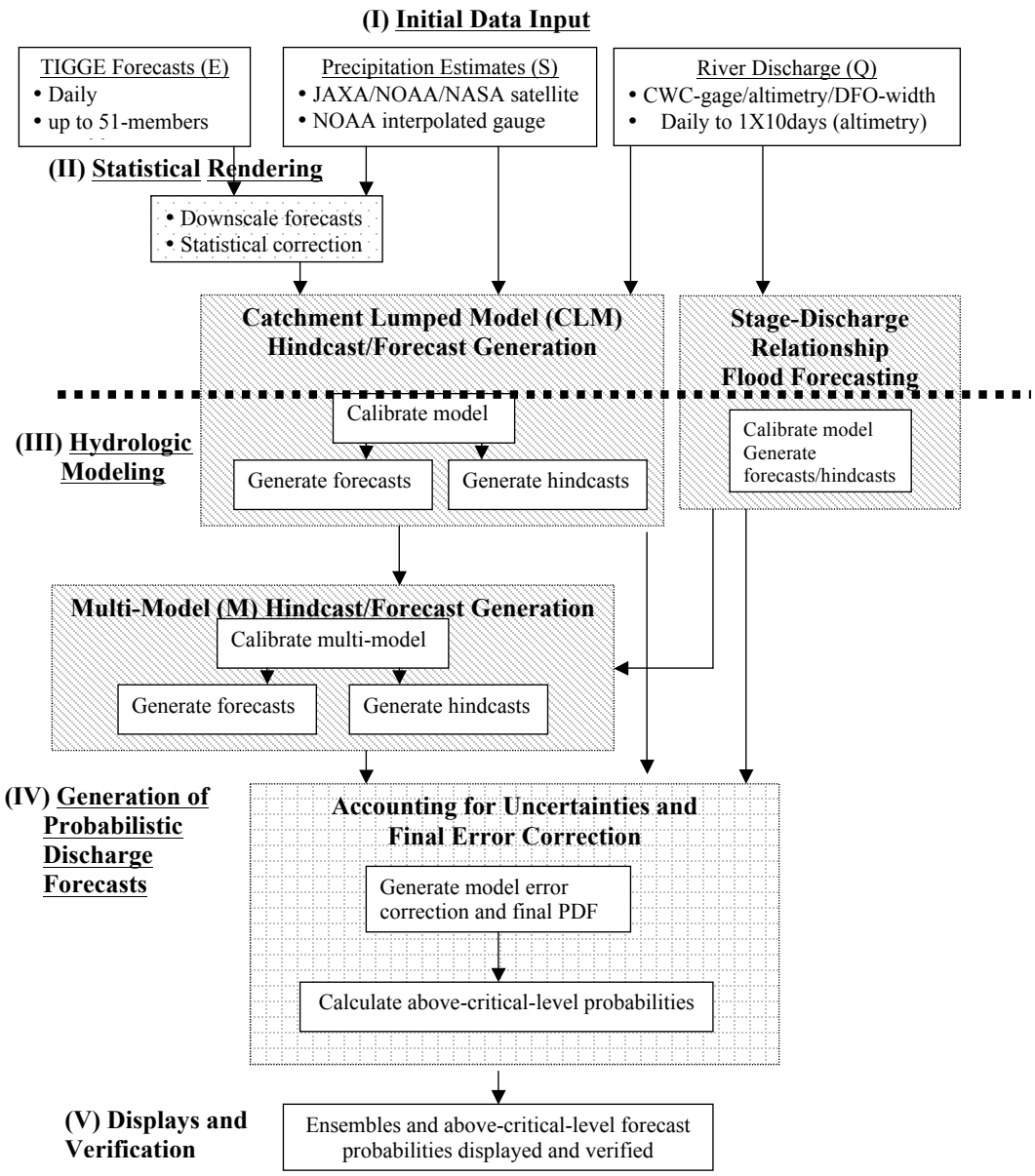


Figure 20: Flowchart of the modified CFAB short-term forecasting scheme. Five steps (I-V) are outlined schematically and within each is a brief description of the data used, tasks performed and etc. The arrows show the path of the data through the system. The components of Step I (clear boxes) represent daily inputs into the scheme. At Step II (dotted box), the TIGGE ensemble forecasts are corrected statistically to reduce systematic error. Step III (hatched boxes) represents the hydrologic modeling

process, where all inputs are integrated to produce river flow forecasts at our 87 forecasting points. Step IV (checkered box) represents the process by which all hydrologic uncertainties are accounted for in the probabilistic forecasting process. In Step V, the ensemble forecasts are tailored to produce probabilistic information necessary for displays and user needs. Note that all processes below the horizontal dashed line are done independently for each day and each 1- to 16-day forecast lead-time.

Riverflow Forecasting Methods

In this section we discuss the variety of approaches we developed and employed in the generation of flood forecasts at locations within Bihar. We discuss each of these individually as discussed below, starting with the method we developed to estimate rating curves from the stage data we have been collecting for gauging locations within Bihar. We anticipate that the FMISC has rating curves at many of these locations already, and believe access to these by this consultant could lead to increases in the skill of the river flow forecasts that we are providing.

Rating Curve Estimation

In many parts of the globe, one may have access to approximate river height (stage) measurements via remote sensing or other means, but without access to the corresponding river flow. In this section we discuss the method and present the results of a method to approximate rating curves (estimates of river flow from stage measurements) where there are no discharge measurement available to us, using discharge measurements from other locations in the river network (upstream or downstream). We applied our approach to the Bihar in situ Indian Central Water Commission stage gaging sites (the same locations we are providing operational river discharge forecasts for) as part of this project. Later in the subsection, we also present a "twist" on the traditional rating curve relationship to also derive a "rating curve" for the relationship of upstream stage to downstream discharge (as an additional function of time lag). These relationships can be directly used for the forecasting problem, with the optimal lag anticipating flood wave travel times, and the "scatter" (error) about the rating curve fit also provides a metric for the degree of predictability the relationship holds, as well as a weighting metric for combining with other upstream gages.

We have designed an approach for estimating rating curves for each of the relevant river gaging sites, converting the stage ([m]) into discharge ([m^3/s]), where there is no (or at least, we have no access to) in situ river flow observations. Similarly, the approach has also derived a "rating curve" of downstream river flow (at a future lagged time) from upstream stage measurements, which is something very useful for hydrologic forecasting purposes. For our data, we used available stage measurements for locations along the river networks in Bihar, along with downstream discharge measurements at Hardinge Bridge (Ganges) in Bangladesh to estimate hydrologic rating curves at each gaging location upstream. The underlying process to derive the "rating curves" at the same location where the stage measurements were collected is discussed first below.

Stage measurements of the location we want to derive discharge values at ($Stg_{upstream}$) were first quality controlled as described in the previous section. For

each location, this cleaned stage data were matched with corresponding discharge values at the respective downstream location ($Q_{downstream}$) where we have discharge data available. Stage and discharge data were matched by date; stage measurements +/-12 hours were considered, and if there were multiple stage observations within the window, the closest to 00:00GMT was selected, as this is nominally the time of the discharge observations.

Once the pairing was done, the downstream discharge values were "normalized" by the ratio

$$\frac{\overline{(P-ET)}_{upstream}}{\overline{(P-ET)}_{downstream}}$$

which is the spatial and temporal average of the precipitation minus evapotranspiration of the upstream catchment area above $Stg_{upstream}$, divided by the same quantity but averaged over the whole catchment area above the downstream discharge gage location ($Q_{downstream}$). Precipitation "P" was derived from our merged satellite precipitation estimates, and evapotranspiration "ET" was derived from the average over all the "analysis" fields of our 8 THORPEX-TIGGE weather forecast products. The temporal averaging period was the entire 2015 monsoon season.

To reflect the travel time between the upstream stage measurement location and the downstream discharge gauge, data were matched at an optimal lag. Two methods of determining this optimal lag were employed: in the first, stage measurements were correlated with discharge at lags of 0-30 days, and the lag that produced the maximum correlation value was selected. In the second, rating curves were fit on stage and discharge at multiple lags, and the lag producing the best fit, as determined by Nash-Sutcliffe efficiency (NSE) was selected. These lags were compared to a calculated travel time determined by distance between the stage measurement location and the discharge gauge, as well as channel slope. The traditional rating curve form was used for stage S_i converted to discharge Q_i,

$$Q_i = A\,(S_i - S_0)^n$$

with the parameters A, S_0, and n determined by nonlinear least squares optimization. (Multiple-part rating curves were also considered; however, visual inspection of the one-part fitting for a variety of sites didn't show clear enough distinction between the fit to low and high flows to justify introducing the added complexity.)

Starting at the most downstream gages, the process was then extended gage-by-gage upstream: using newly-estimated rating curves and thus discharge values at sites just upstream of Hardinge Bridge (Ganges), these sites and their discharge values were then used to estimate rating curves at sites further upstream, and so on. At each location, the optimal-fit rating curve (as determined by the largest NSE) was

selected by comparing with previously derived rating curves that used discharge from locations further downstream.

The underlying process to derive the "rating curves" at downstream (i.e. "forecasting") locations was done identically to what was just discussed, except the downstream discharge values ($Q_{downstream}$) were not first "normalized" before the optimal fitting was carried out.

Quality of Rating Curve Fit Filtering

Rating curve estimates were fitted for the stage gaging locations as discussed above. However, this set needed to be further reduced down for forecasting purposes, given that the (unconstrained) fitting of our rating curve functions converged to unphysical exponents. The traditional rating curve form was used for stage S_i converted to discharge Q_i,

$$Q_i = A\,(\,S_i - S_0\,)^n$$

with the parameters A, S_0, and n determined by nonlinear least squares optimization. However, the exponent "n" found for many of our gauging locations was either significantly less than a value of one, or significantly greater than a value of 5. As such, these gauging locations were (temporarily) removed from our list of forecasting sites (but can very easily be added back in if additional rated stage-Q valued could be provided)..

Stage-Discharge Relationship Flood Forecasting

This approach builds from the "in situ and forecasting rating curves" we discussed above, that derived a "rating curve" of downstream river flow (at a future "optimal" lagged time) from upstream stage measurements; which is something very useful for hydrologic forecasting purposes. Using these rating curves with observed stage (whether from CWC stage gage or river "width" measurements), forecasts of downstream discharge can be produced at 0 day (current condition) to 16 day forecasts (or beyond).

The figures and figure captions immediately below show conceptually the idea behind these forecasts and how they are combined, where we are capitalizing on travel-time delays from upstream to downstream locations.

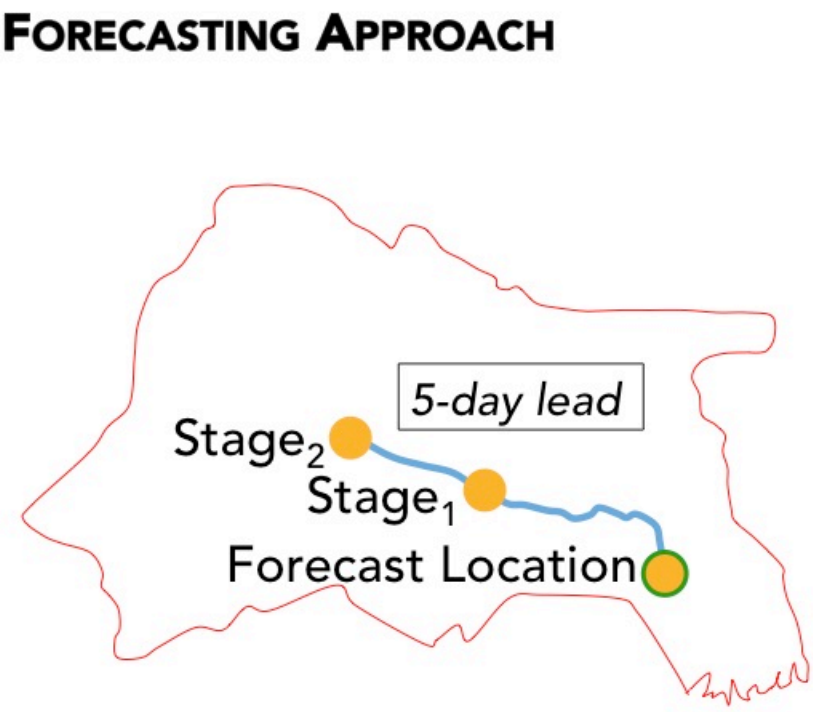


Figure 21: conceptual idea of using upstream stages (stage 1 and stage 2) to forecast a downstream location based on a 5-day flood wave travel time from the most upstream to downstream locations.

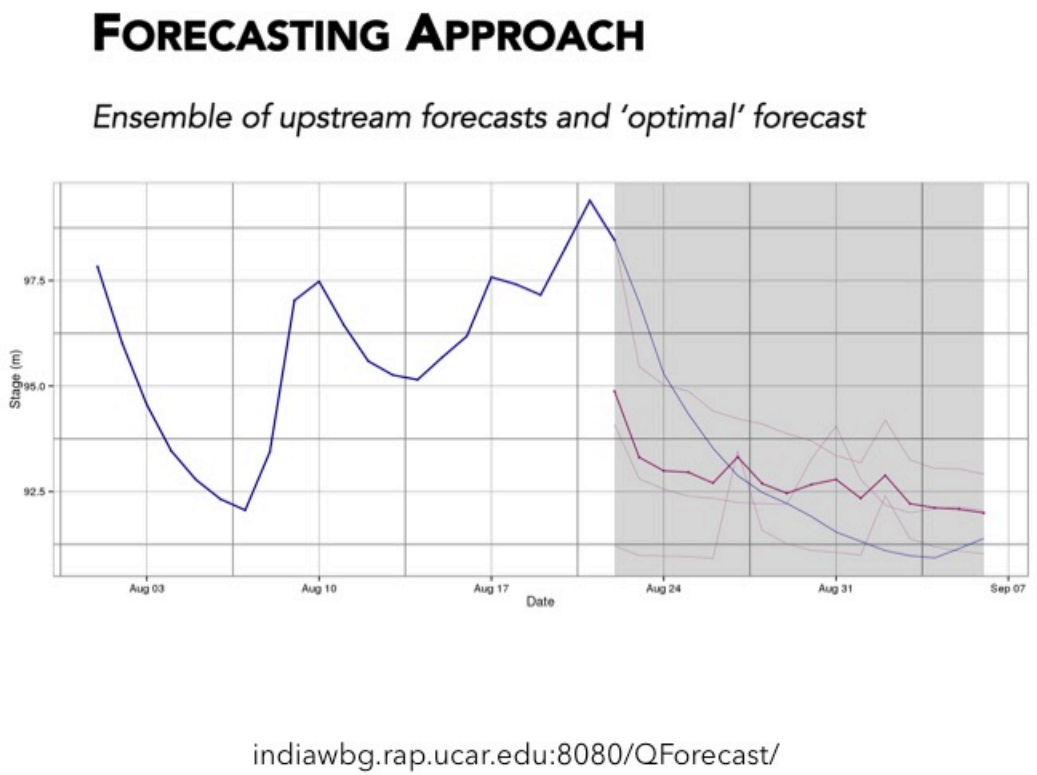


Figure 22: depending on the number of upstream gauges, multiple forecasts can be estimated, one for each gage.

FORECASTING APPROACH

Forecast info table

	Fcst. Date	Fcst. Lead (days)	Upstr. Sta.	Lag	Date	Upstr. Stage (m)	Weight	Single Sta. Fcst. (m)	Single Sta. NSE	Combined Sta. Fcst. (m)	Combined Sta. NSE	Obs. Stage (m)
1	2015-08-24	2	007-LYDAGRA	2	2015-08-22	391.30	0.25	90.97	0.255	92.99	0.573	95.29
2	2015-08-24	2	009-LYDAGRA	2	2015-08-22	112.91	0.42	95.02	0.611	92.99	0.573	95.29
3	2015-08-24	2	011-LYDAGRA	2	2015-08-22	198.23	0.33	92.56	0.470	92.99	0.573	95.29

indiawbg.rap.ucar.edu:8080/QForecast/

Figure 23: the error of each upstream gauge (an example of which is shown here from our website provided) forecast is available directly from the derived "rating curves".

FORECASTING APPROACH

$$Q_{fcst} = \frac{\sum_1^n C_i (Stage_i + a_i)^{n_i} \cdot \frac{1}{\varepsilon_i}}{\sum_1^n \frac{1}{\varepsilon_i}}$$

Figure 24: the error from the derived rating curves of upstream gauges is then combined in the manner shown in this figure, to generate an "optimal" combined forecast.

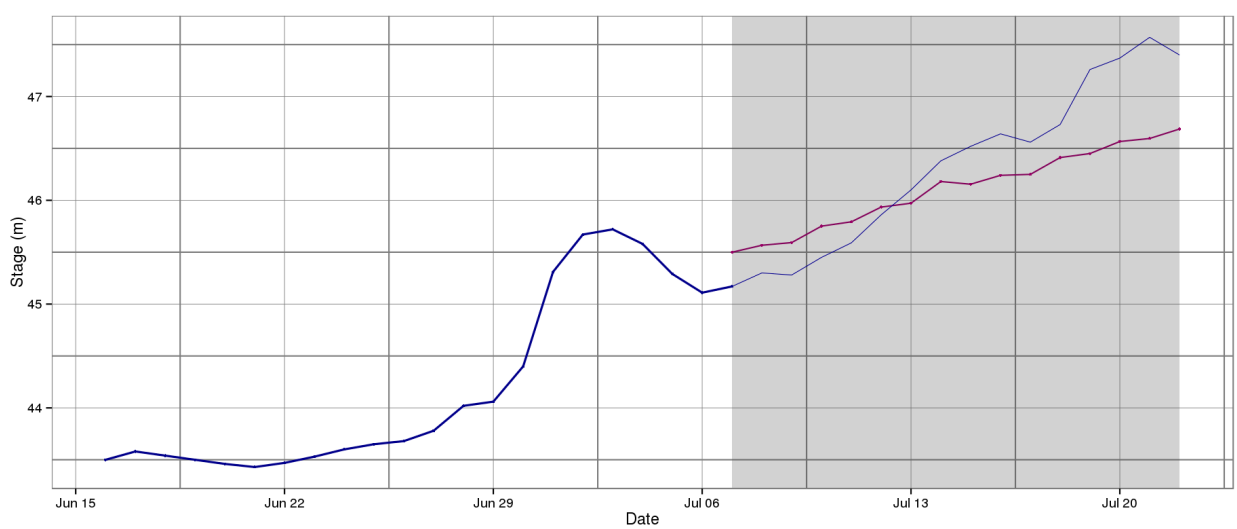


Figure 25: results of our combined upstream stage forecast for one forecasting site. The blue is the observer, red is the forecast, and the grey represents to the forecast period.

Quantile Regression – individual ensemble and multi-model forecasting

In this subsection, we briefly discuss a statistical forecasting and correction approach we introduced into this project that we rely on heavily in this project, called "quantile regression" (QR). In particular, we use this approach in both "III. CLM Generation", and "IV Accounting for Uncertainties and Final Error Correction"in the figure above showing the riverflow forecasting scheme (as well as used in the precipitation forecast calibration for individual centers and in their multi-modeling). We defer more technical details of this approach to the appendix below. Quantile regression has been little used in the atmospheric and hydrologic communities. We introduced this approach for this project, relaying on it heavily to generate ensemble forecasts and forecast error corrections. Similar to simple linear regression which minimizes the square error of a model to-be-fit, QR instead minimizes the absolute error, but done for each "quantile" of the distribution one is interested in. The result is that multiple models are generated, one for each quantile, and each provides a reliable probability of what values the river flow will exceedance on a daily basis (such that "flat histograms" of the forecasts are produced – described further in an appendix below). As we apply it for our forecasting application, we ensure that the resultant ensemble discharge forecasts are no worse than the climatological average flow after the model fitting process.

Dashboard Framework and Data Access

In this section we discuss the customized (and further customizable) dashboard framework we have developed for the FMISC. We provide an overview of displays of current conditions occurring over Bihar; another page devoted to THORPEX TIGGE ensemble weather prediction and satellite-based precipitation products; another page devoted to comparing past severe rainfall events over Bihar; and a page devoted to flood forecasts and observations, again, over Bihar. We provide an overview of the process to access this information on this site, with the site itself found at: http://bihar.rap.ucar.edu/. We also discuss the forecast products we provide directly for download from FTP access.

Current Condition Overview

We are providing a view of current conditions in watershed from the main tab of our website, both in terms of rainfall and in river flow, as seen in the screenshot figure below. Gauging locations are shown, along with their level of alert given by their color. By clicking on a particular gauging location, the upstream basin area above the gauge is highlighted, showing also the recent rainfall that has fallen in that basin. The click also pops up a subwindow showing the river stage forecasts for that gauge, as seen below.

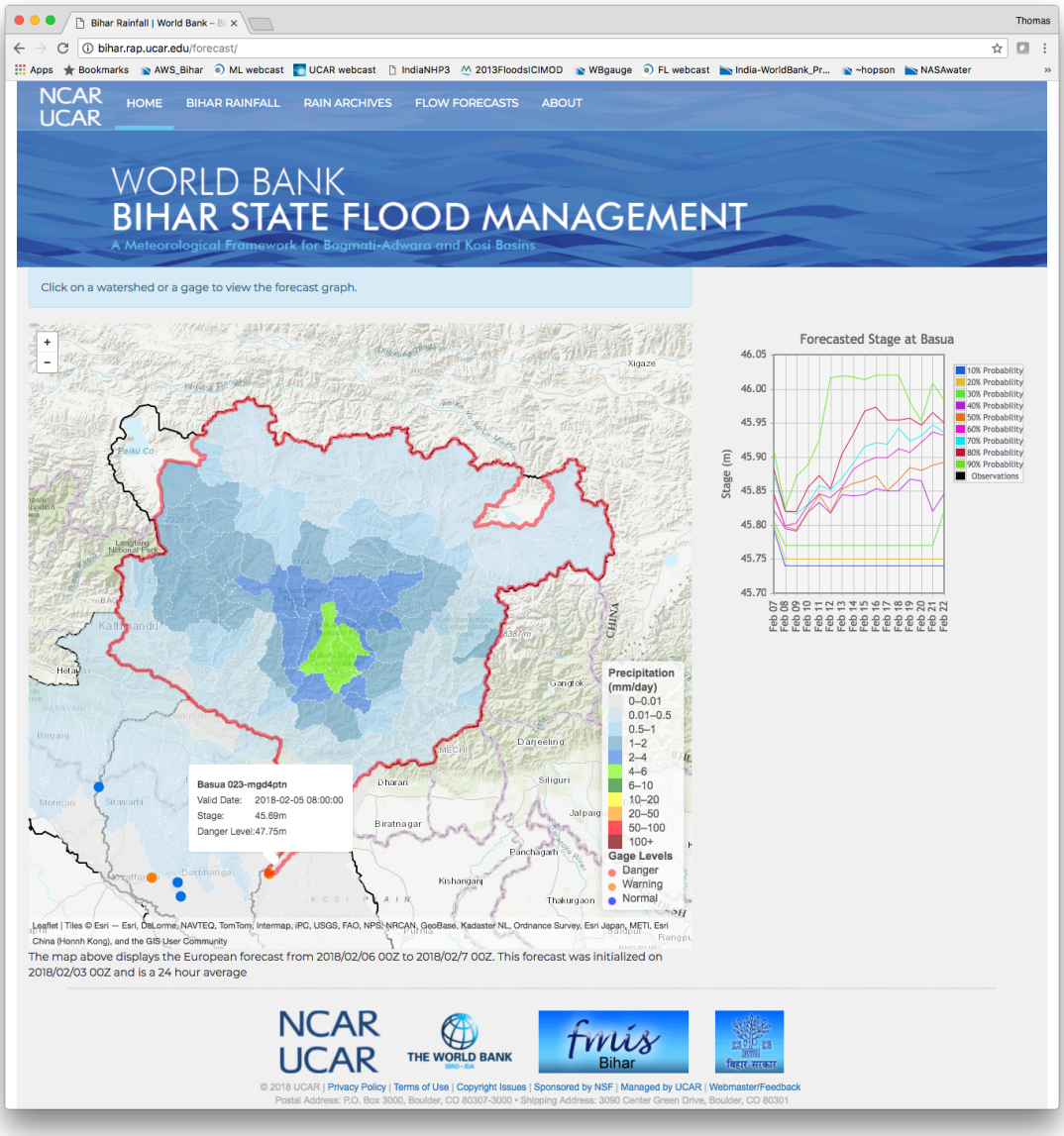


Figure 26: Screenshot of current conditions occurring in the region, as seen from our main home page for the project.

Displays of Ensemble Numerical Weather Prediction and Satellite Precipitation

In this section, we provide an overview of displays of our THORPEX TIGGE ensemble weather prediction and satellite-based precipitation products. We provide an overview of the process to access this information below, with the site itself found at: http://bihar.rap.ucar.edu/rainfall/.

Because of the essential importance rainfall plays in generating runoff, many flooding events can be anticipated just by locating areas of severe rainfall, especially relevant for flash flood events occurring over smaller catchment spatial scales. With this motivation, we are generating an hourly-updated display of rainfall occurring over Bihar (follow the link in our project's web site under "Bihar Rainfall").

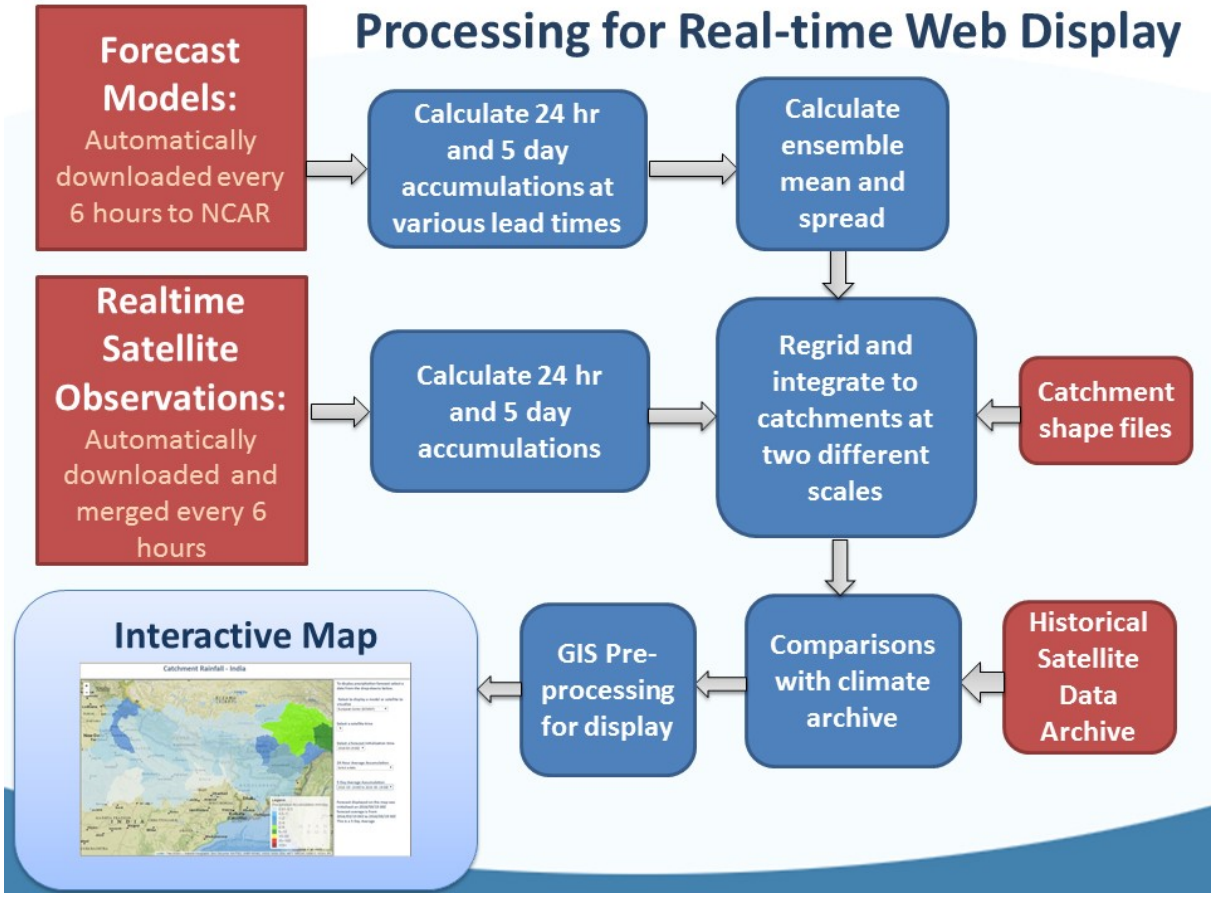


Figure 27: Overview of the processing and display of catchment-integrated real-time satellite precipitation and TIGGE ensemble precipitation forecasts over Bihar catchments.

This interactive web display, shown in outline form in the figure above, gives users an easy way to visualize and compare precipitation accumulation for different weather models and satellite products. The near real time web display (currently) incorporates ensemble precipitation forecasts from four of our eight THORPEX-TIGGE forecast centers and our merged NASA TRMM, NOAA CMORPH, and JAXA GSMaP satellite precipitation product. This portal is an interactive web display that expands Geographic Information Systems (GIS) mapping capabilities to visualize forecast precipitation accumulations averaged over watershed basins. Models and satellite estimates are calculated to 24 hour and 5 day average accumulations. The models that are displayed are the European ECMWF, the UK UKMET, the Canadian

CMC, and the USA NCEP models. Seasonal forecasts of the North American Multi-model Ensemble (NMME) are also provided, calculated at 1month and 3month accumulations, from seven different models: CFSv2, CMC1, CMC2, GFDLFLOR, GFDL, NASA, NCARCCSM4. Each model has a different number of ensemble members. The ensemble average and other statistics are aggregated to a catchment and sub-catchment basin. By default the larger catchments are displayed. When zoomed in the sub-catchment become visible for a more detailed view. Ensemble average per watershed is being displayed on the map and ensemble statistics for the watershed are available through a popup window.

This portal is built using a number of technologies. All forecast and satellite products are exported to text comma separated values file (.csv). This file format allows data to be saved in a table structured format but as a text file. The text file forecasts are produced every 24 hours. Once the forecast and satellite precipitation products are created, a GIS python script is automatically run to convert these data into a spatial GIS-friendly format. The web application is built using a free open-source JavaScript API called Leaflet. This library is ideal for powerful data driven visualizations. This portal takes advantage of the most appropriate tools and technologies to easily display the information from the weather forecast and satellite products. In what follows, we step through the process of displaying the different products found on this site.

Step 1 – Select a model or Satellite product

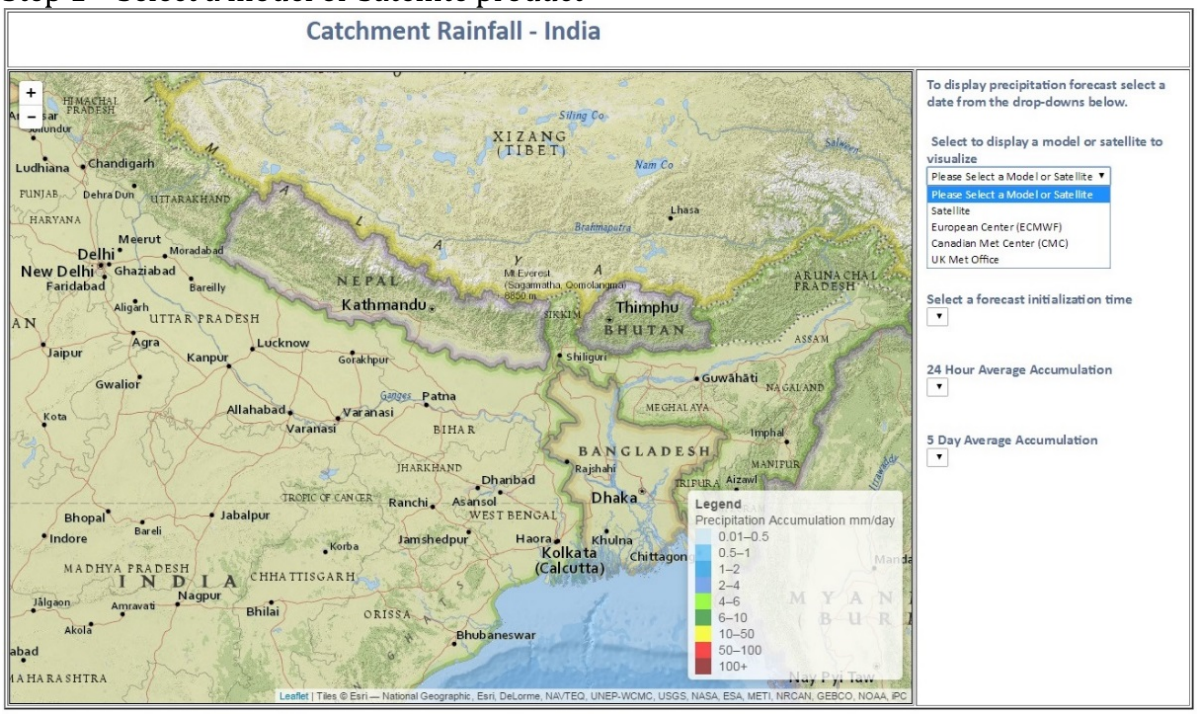


Figure 28: Step 1 – Select a model or Satellite product

Step 2 – Select the forecast initialization time.

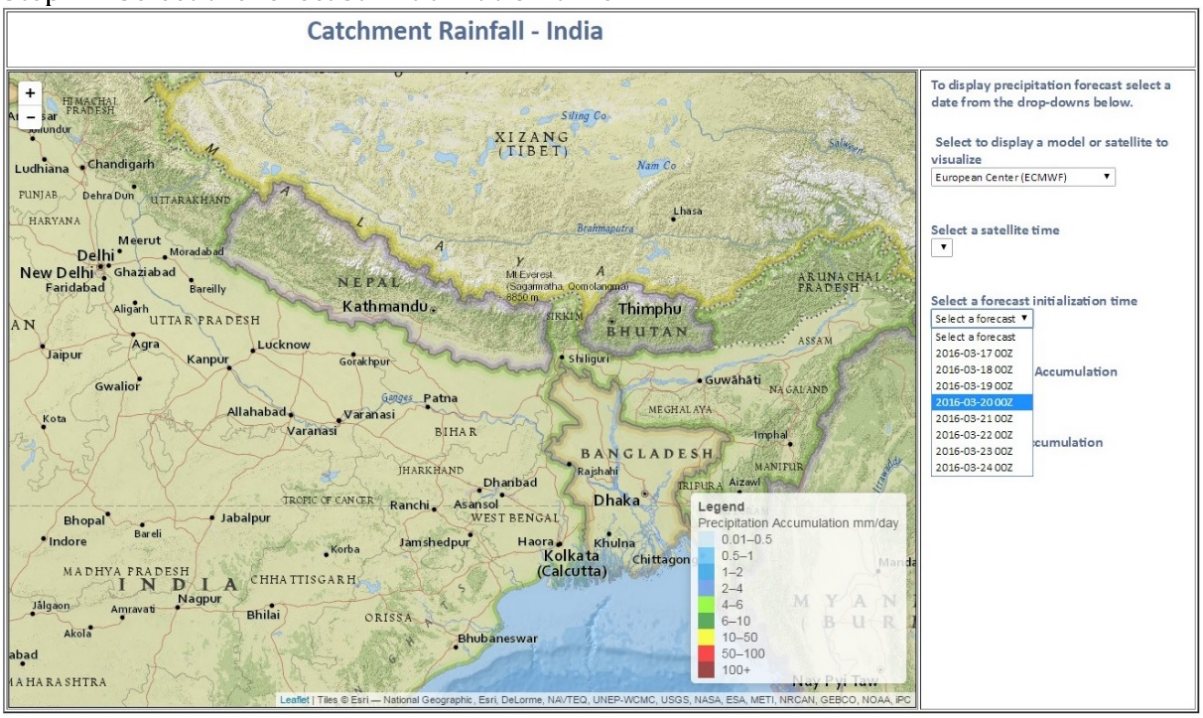


Figure 29: Step 2 – Select the forecast initialization time.

Step 3 – Select the 24 hour or 5 day accumulation date.

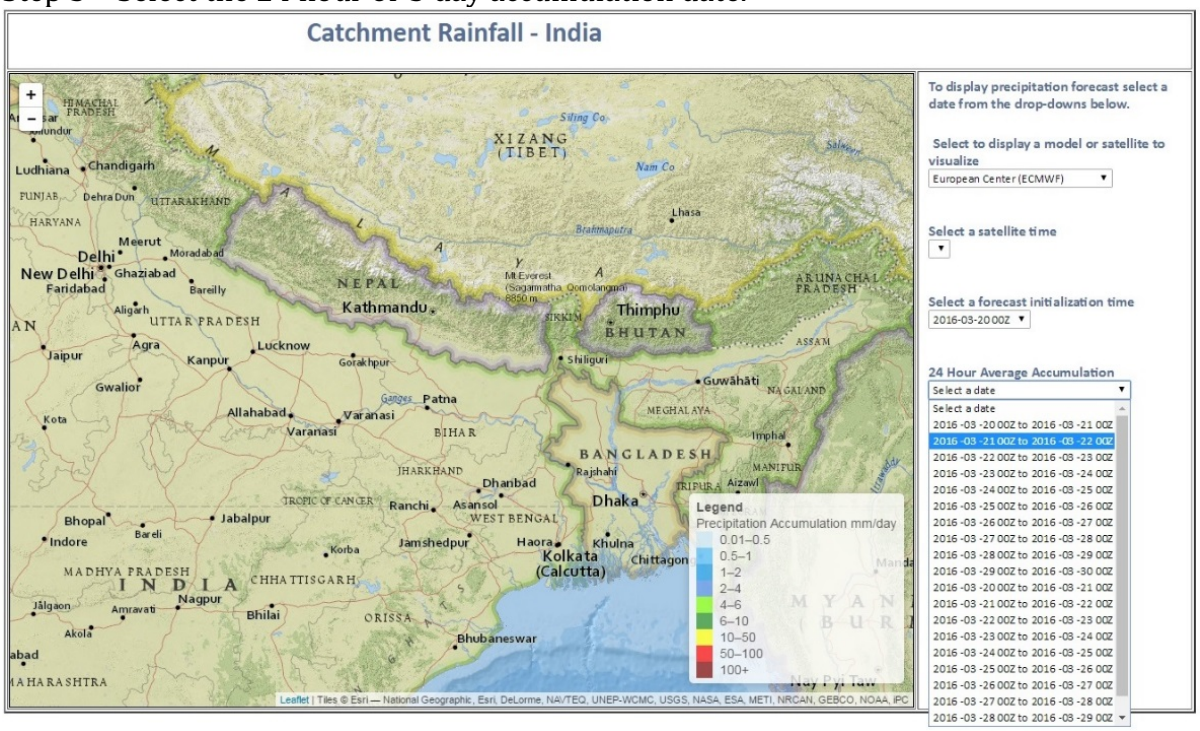


Figure 30: Step 3 – Select the 24 hour or 5 day accumulation date.

Result - An interactive web map is displayed with zoom and pan functionality.

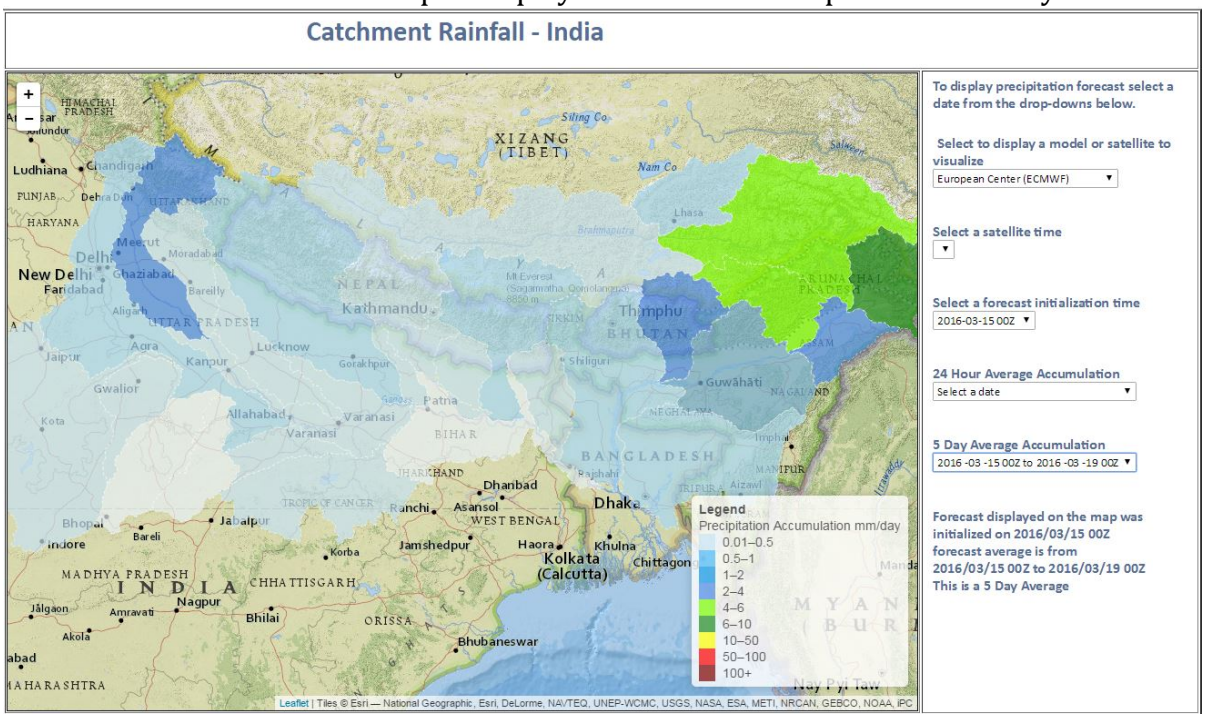


Figure 31: Result - An interactive web map is displayed with zoom and pan functionality.

By clicking on a watershed a popup window appears with more detailed information about the model ensemble statistics.

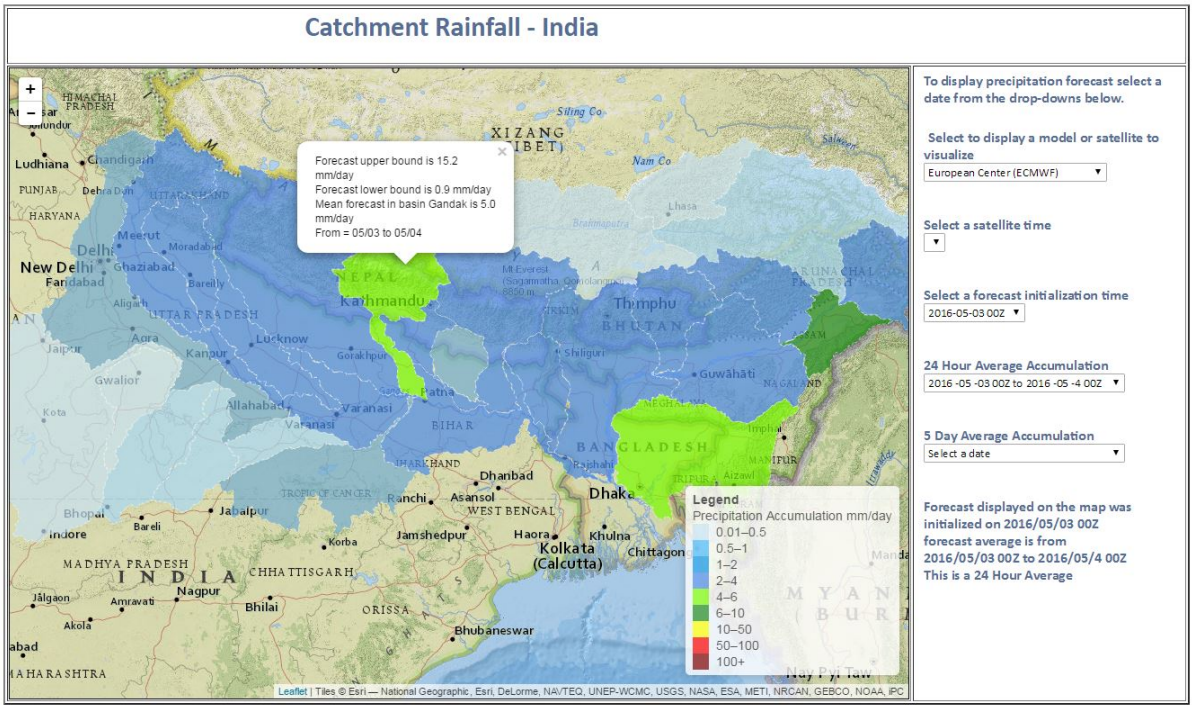


Figure 32: By clicking on a watershed a popup window appears with more detailed information about the model ensemble statistics.

When zoomed in the sub-catchments are displayed for a more detailed view.

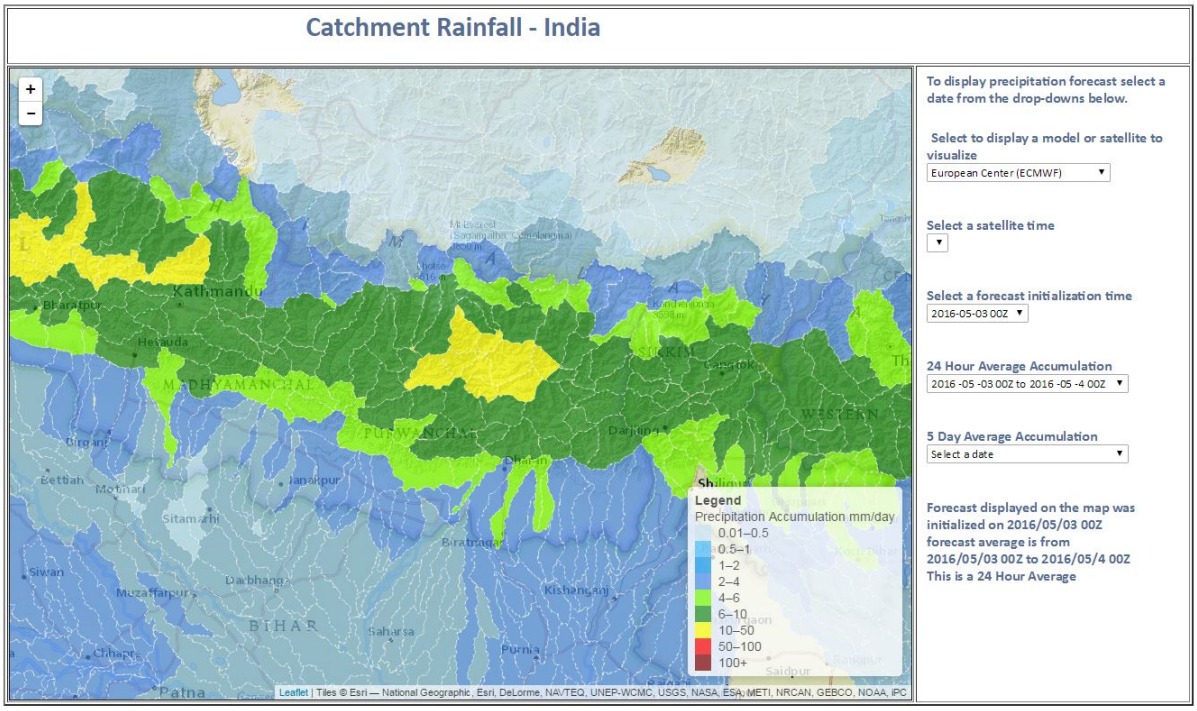
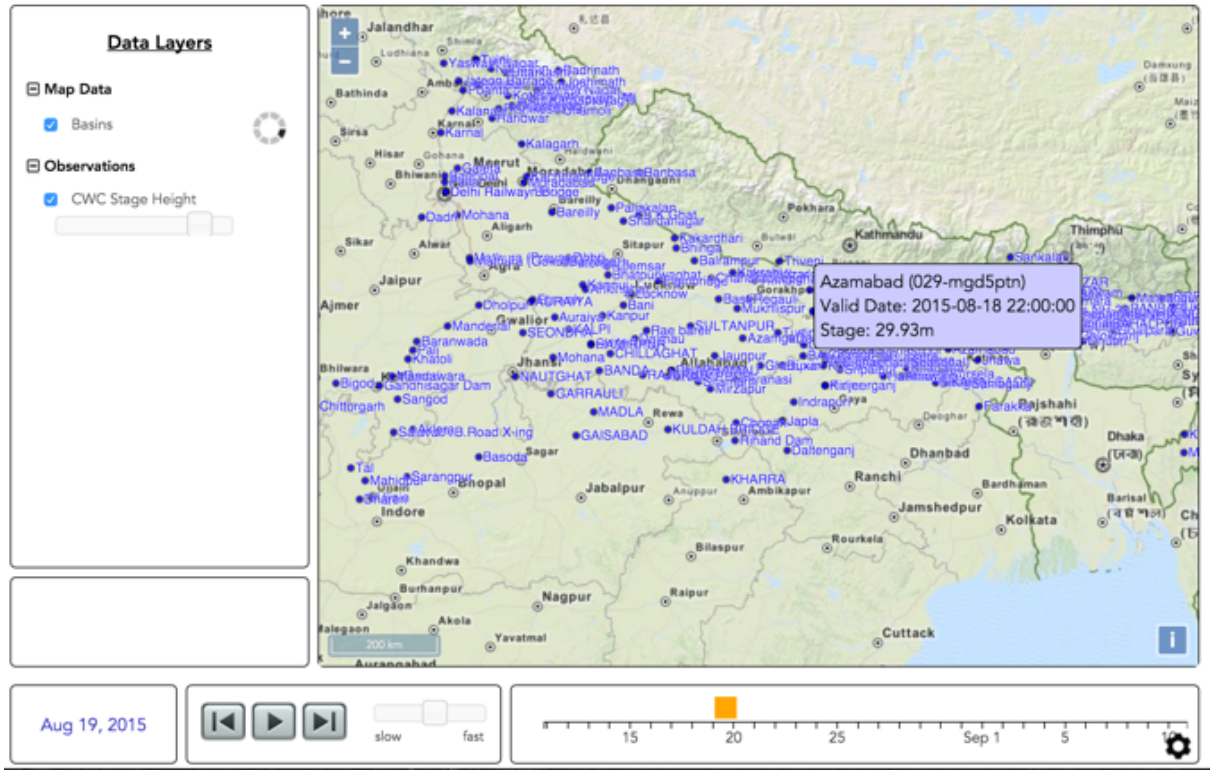
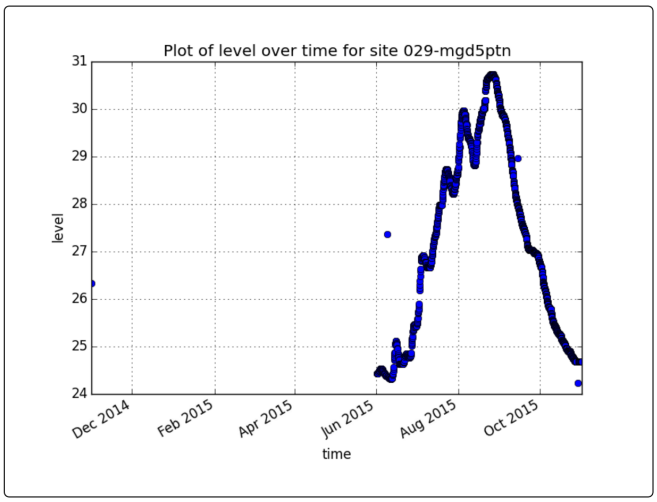


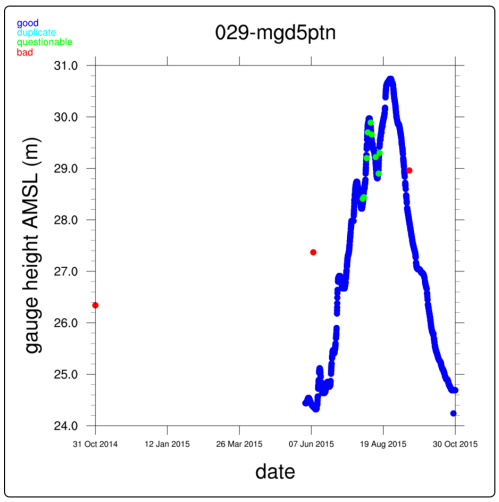
Figure 33: When zoomed in the sub-catchments are displayed for a more detailed view.

Displays of Water Level Forecasts and Observations

In this subsection, we discuss the processes that led to the display of CWC in situ river stage gauging data found at http://bihar.rap.ucar.edu/flow/. This site also provides displays of our operational ensemble river forecasts under a pilot testing link. The web display is map-based in order to visualize the geographic extent of the river basins and to allow gridded model forecasts to be easily displayed. Stations in the display may be selected by clicking them with the mouse. When selected, a station displays its detail information in a pop-up window. Selected stations are also automatically queried to generate plots of observations, ensemble forecasts, and quality control analyses at that location. Examples are shown in the figures immediately below.


Data Layers
Map Data
Basins
Observations
CWC Stage Height
Azamabad (029-mgd5ptn)
Valid Date: 2015-08-18 22:00:00
Stage: 29.93m
Aug 19, 2015
slow
fast
15
20
25
Sep 1
5


Plot of level over time for site 029-mgd5ptn
level
time
Dec 2014
Feb 2015
Apr 2015
Jun 2015
Aug 2015
Oct 2015


good
duplicate
questionable
bad
029-mgd5ptn
gauge height AMSL (m)
date
31 Oct 2014
12 Jan 2015
26 Mar 2015
07 Jun 2015
19 Aug 2015
30 Oct 2015

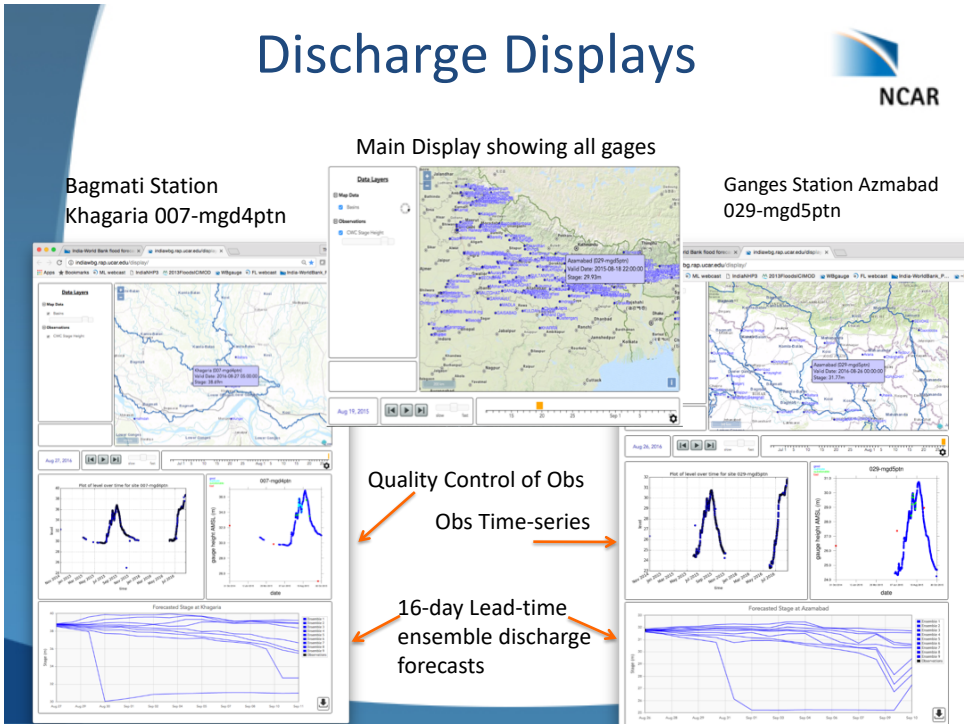


Figure 34: displays provide visualize of the geographic extent of the river basins, observations, ensemble forecasts, and quality control analyses at that location. Further above is the live display, and immediately above is a selection of two gauge forecasts over Bihar.

The display is written in JavaScript, so that it may be viewed in a web browser on a desktop computer or a mobile smartphone.

Review of Past Severe Events

Displays of rainfall comparisons showing forecasted and observed rainfall for past severe events over Bihar were discussed further above in this report, and are found at: http://bihar.rap.ucar.edu/archive/ from our website.

Amazon Web Services Access

Our website and much of our computations are being computed and hosted live on the cloud. This account can be shared with (and given to) the FMISC. This site can be viewed at: http://ec2-35-165-102-63.us-west-2.compute.amazonaws.com/forecast/index.php

Direct Data Access (FTP)

Operational daily updated forecasts are also being provided at the NCAR FTP site in four different formats (ESRI ASCII, netcdf, CSV, and PNG) for use by the FMISC and two other consultants working in the region. This site is found at the following FTP site at: ftp://ftp.rap.ucar.edu/incoming/irap/bihar/

Appendix A -- Quality Control Details

River gauge quality control

Raw river level gauge data can occasionally have errors, such as those shown in **Error! Reference source not found.**. Therefore, it is important to have quality control (QC) procedures in place, to flag bad data, as well as identify suspicious data. We employed several quality control procedures on several months of data from over 300 river level gauges in India. The data from each station were used to develop quality control thresholds for its own data; this assumes that there is a sufficient amount of data (e.g., several months), and that most data at each station are accurate. The quality control procedures were used to identify the following common errors: 1) multiple observations at the exact same time, 2) data points that are exceptionally larger or smaller than nearly all other observations at that station, 3) excessively rapid changes in river level data, and 4) river levels that are indicated to be constant for too long that are likely the result of a stuck gauge. Due to the large amount of data to be QCed, it was important that the procedures be designed to be very computationally efficient. Different QC values are used to identify values that fail each procedure, while a QC value of 100 indicates highly probable good data. Since data are compared with each other for each QC procedure, data that fail one procedure are not considered for comparison in future procedures; likewise, each procedure rejects erroneous data one-by-one, from the most extreme to the least, so that nearest-neighbor procedures do not falsely identify good data. However, the raw data value is always retained, accompanied by its appropriate QC flag value. The details of each procedure are laid out below.

The first procedure determines if the raw data are actual levels and dates, and not seeming nonsense. Any levels and/or dates that are unidentifiable are marked with the QC value of -1 for missing (Table 2).

QC score	Meaning
100	Good

0-99	questionable rate of change or possible stuck sensor
-1	Missing
-2	Isolated
-3	too high/low
-4	extreme rate of change
-5	stuck sensor
-11 -12	Duplicate missing neighbor (satellite QC only)

Table 2: QC score values, and their meaning.

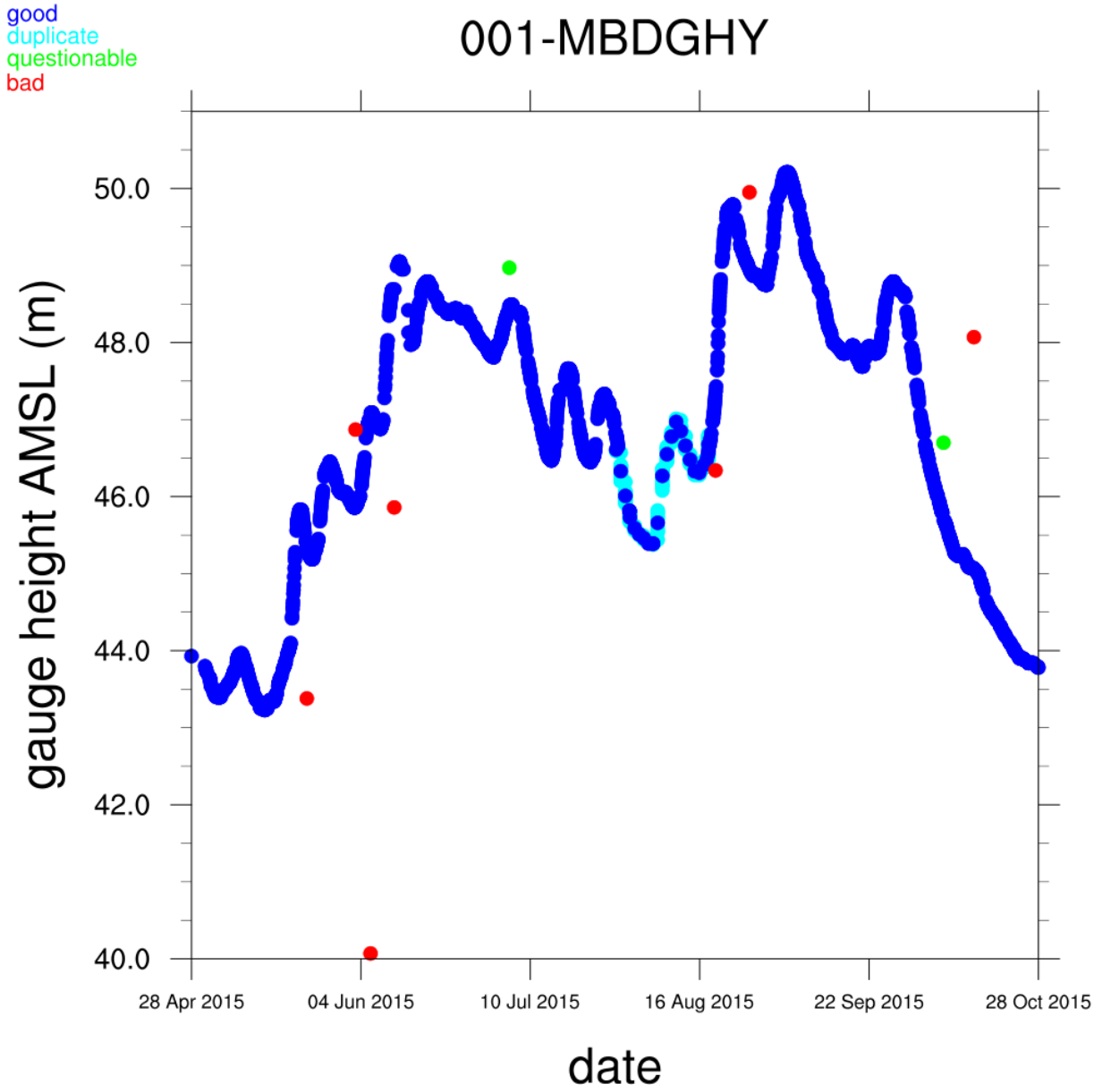


Figure 35: River level data from station 001-MBDGHY. The dot color identifies the data as: (blue) likely valid river levels, (red) very likely erroneous data, (green) questionable data, and (cyan) data observed at the exact same time.

The second procedure identifies all observations at the exact same time. If they all have identical river levels, only one is retained. Of those that share a common time, but are not identical, there are two tests employed. If there are only two different observations at the exact same time, the one with the greater difference with its other nearest neighbor in time is identified with a QC value of -

11. If there are three or more different river levels sharing the same time stamp, the median is calculated, and all observations are marked with a QC value of -11, except for the level closest to the median. Examples of multiple observations identified at the exact same time are indicated by cyan dots on **Error! Reference source not found.**.

The third procedure identifies data points with no other nearby data points in time. If a value has no other valid value within 10 days, it is marked by a QC value of -2.

The fourth procedure identifies gauge levels that are exceptionally lower, or higher, than nearly all other levels. To do this, the 10th and 90th percentiles are calculated. The Xth percentile is defined as the level at which only X% of the river levels are lower than that value. Then, the difference, ΔL_{10th}^{90th}, between the 10th and 90th percentiles is calculated. If a gauge level is greater than $L_{90th} + c_H * \Delta L_{10th}^{90th}$, or lower than $L_{10th} - c_L * \Delta L_{10th}^{90th}$, it is marked with a QC value of -3. The constants c_H and c_L were optimized so that they rejected most excessively high and low values, while not falsely identifying any valid values. These optimized values were found to be $c_H = 4.1$ for and $c_L = 1.3$. Examples of river gauge levels identified by this procedure are shown in Fig. 2.

The fifth procedure identifies data with excessively rapid time rates of change on either side, namely sharp spikes and dips in the data. This is done by calculating the time rate of change, both before and after each data point, and then summing the two together: $\left(\sum \left|\frac{dL}{dT}\right|\right)$. Next, the 95th percentile, is calculated from all these values for each station. If any $\sum \left|\frac{dL}{dT}\right|$ exceeds $c_T * \sum \left|\frac{dL}{dT}\right|_{95th}$, it is marked with a QC value of -4. The value of c_T was optimized at 16 so that it rejected most bad data, yet did not reject hardly any good data. Examples of data identified by this QC procedure are shown by red dots on Fig. 1. There remained, however, additional minor errors that were not flagged by this procedure, such as some of those shown in green dots on Fig. 1. Therefore, another smaller constant, c_B, was calculated to

identify data points with rapid changes on either side; if their value $\sum\left|\frac{dL}{dT}\right|$ exceeded c_B they were given a QC score of 0-99, according to the following equation:

$100 - 100 * \frac{\sum\left|\frac{dL}{dT}\right| - c_B * \sum\left|\frac{dL}{dT}\right|_{95th}}{c_T * \sum\left|\frac{dL}{dT}\right|_{95th} - c_B * \sum\left|\frac{dL}{dT}\right|_{95th}}$. The higher the score, the more probable it is that the data is good, while the lower scores indicate a high probability it is bad. c_B was optimized at 5 to identify nearly all visually bad gauge level spikes. Examples of gauge levels identified by a score of 0-99 are indicated by green dots in Fig. 1.

Lastly, the sixth procedure identifies river gauge levels that have been constant for an unusual length of time, indicative of a possible stuck gauge. Since the temporal resolution of the river level data varies substantially with time and location, a count of successive constant observations (s), of any time length, were used to identify possible stuck gauges. A lower threshold s_B was used to identify a possible stuck gauge, while an upper threshold s_T was used to indicate a nearly certain stuck gauge. All level data constant for greater than s_T consecutive observations are given a QC score of -5, while all data constant between s_B and s_T time steps were given a QC score between 0-99, according to the equation: $100 - 100 * \frac{s - s_B}{s_T - s_B}$. All river level data in a string of constant values are given the same QC score, from the first to the last.

The next steps to implement in this QC system are: 1) collect at least one year of river gauge level data, in order to capture at least one full monsoon cycle; this will allow for re-calculation of better thresholds for each station. 2) Determine a way to reject "spikes" of data that are more than a single point. 3) Identify spikes of bad data with missing temporal neighbors. There are certainly other QC procedures that can also be developed and implemented on these data.

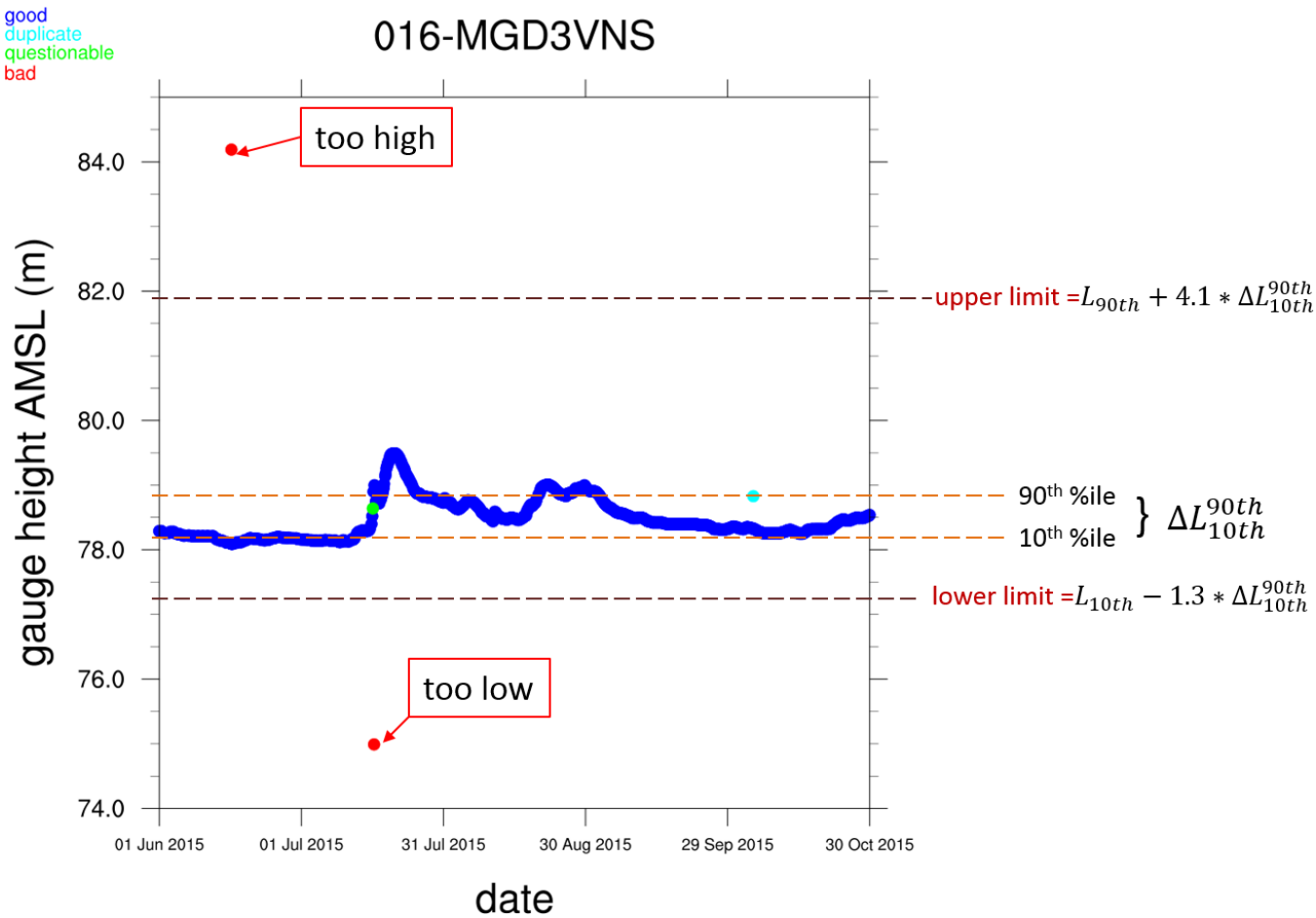


Figure 36: River level data from station 016-MGD3VNS, showing examples of erroneous extremely high and low river gauge values.

Satellite-deived river width quality control

Similar quality control procedures were used to identify possible erroneous data from satellite-derived river width products. These river width estimates for very large rivers are obtained by using satellite microwave emissivity values to approximate the percentage of water in a pixel (~10 km x 10 km), M, compared to that of a nearby pixel with no water, C, to create a ratio. This ratio can then be used to estimate river width and subsequently the river height. An example time series of M/C ratio for a location on the lower Ganges River is shown in Fig. 3.

The first QC procedure identifies river widths that are exceptionally lower, or higher, than nearly all other levels in the time series; this is done in the same way as the river gauge height QC procedure, by identifying the 10th and 90th percentiles, and using those to create upper and lower thresholds for the data. If data fall outside these thresholds, they are indicated with a QC flag of -3. The second QC procedure is also a replication of a gauge height procedure, which identifies data with excessively rapid time rates of change on either side; extreme rates of change give a QC value of -4, while questionable rates of change give QC scores of 1-99. Since these QC procedures flag a significant minority of values (e.g., Fig. 3), a third QC flag (-12) is used to identify times with missing or bad data immediately before and/or after, that were not able to be tested using the extreme rate of change QC procedure. The -12 QC flags do not indicate that the data are bad, just that there is lower confidence in their accuracy, since they aren't able to be sufficiently tested.

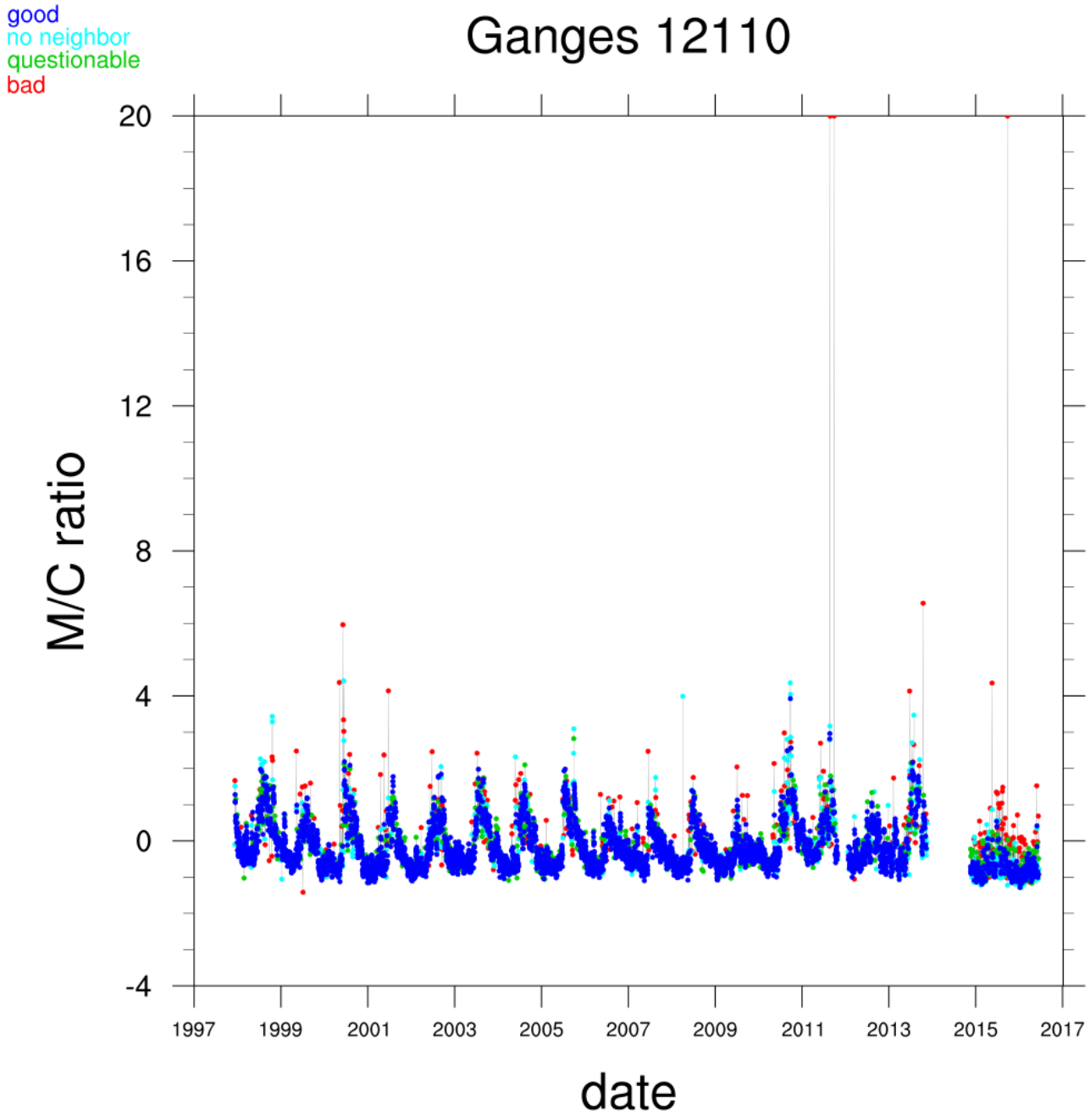


Figure 37: Time series of M/C ratio for a location on the lower Ganges River. Pixel colors indicate its QC flag.

Appendix B -- Quantile-to-Quantile Mapping

Generally, all practical hydrologic models require some form of calibration arising from reasons such as incomplete knowledge of watershed properties and the necessary parameterization of transport at unresolved scales. Through the hydrologic model parameter calibration process, some of the additional errors in the watershed inputs (rainfall, in particular) can also be implicitly reduced (e.g. runoff errors arising from an over-bias in rainfall forcing can be implicitly minimized by increasing the parameterization of evapotranspiration). However, such bias reduction through calibration does not occur if there are relative errors between more than one input of the same physical quantity. In particular, using both weather forecasts and observationally-based estimates of precipitation concurrently to generate unbiased discharge forecasts, requires that these two data sources maintain their statistical similarity such that there are no relative biases between their statistical moments (mean, variance, and skewness, in particular). Note that quantile regression (discussed above) could also provide the required bias reduction on a forecast-day-by-forecast-day bases. However, because we feel that the merged satellite precipitation product (i.e. "observations") inherently have their own error, we hypothesize it is more optimal to only utilize a more "global" statistical adjustment of the TIGGE precipitation forecasts to require them to only be drawn from the same probability distribution function as the observations. Quantile-to-quantile (q-to-q) mapping described next is designed to exactly do this.

To minimize systematic differences in the hydrologic model inputs that can lead to systematic hydrologic forecasting errors (in our case, satellite precipitation estimates [which were used to calibrate the hydrologic models] and the THORPEX-TIGGE ensemble precipitation forecasts which drive the hydrologic model forward in time), a quantile-to-quantile (q-to-q) mapping technique was implemented in this project. As applied here, the q-to-q technique forces each TIGGE ensemble precipitation forecast catchment-averaged over each basin we are forecasting for (input E, Fig 3) to be statistically sampled from the cumulative distribution function (CDF) of the associated observational historical record (i.e. satellite-precipiation estimates) catchment-integrated for the same basin (input S Fig 2). In essence, this converts the numerical weather prediction product from a physical variable to a probability forecast by performing a bias correction at each quantile of the forecast CDF. The manner in which the technique is applied to the TIGGE medium-range forecasts is described below in more detail.

Calculate the monsoon season "climatological" empirical CDF's for both the observations and the forecast model catchment-integrated valuesfor basin *i*. The observational (*o*) climatological CDF for basin *i* and precipitation amount *y*, $C_o^i(y)$, is derived from the merged satellite data (available from 2002 onwards). The forecast (*M*) CDF's at lead-time *f* and basin *i*, $C_M^{i,f}(y)$, are derived from stored forecasts

(available 2014 onwards), which are first catchment-integrated over the same basin, and the CDF computation is done independently for each forecast lead-time *f* (24 hour, 48 hour, etc.). Formally, these CDF's are given by

$$C_o^i(x) = \mathrm{P}\{X \le x\}, \tag{1}$$

and

$$C_M^{i,f}(y) = \mathrm{P}\{Y \le y\}, \tag{2}$$

where P{α} denotes the probability of the event α, and x (y) a specific value of the generic random variable X (Y) representing the observed (forecast) weather variable of interest (here, precipitation only). Using these CDF's, a direct quantile-by-quantile bias correction is made to each basin *i*, lead-time *f*, and precipitation ensemble member *fcst* of the forecast, $y = p_{fcst}^{i,f}$, of the forecast by simply setting (4) equal to (5), and determining the value of $x = p_{adj}$ that satisfies the equation. In this manner, the "observation-space" quantile is matched with the "forecast model-space" quantile as shown in the figure below, and p_{adj} is used in replace of $p_{fcst}^{i,f}$ in the hydrometeorological application.

Note that this technique ensures that the forecasts produce the same climatological number of "no rain" events as observed. It also preserves the spatial and temporal (ranked) covariances of the weather variable forecast fields (as generated by the numerical weather prediction model – although not so important for our application of catchment-integrated values).

If there is no contemporaneous forecast and observational data, we suggest that the q-to-q technique may be the best one can do in bringing forecasts into statistical similarity with the observations. However, although the mapping improves the forecast bias at each quantile by steering the forecasts into having the same frequency of events as the observation climatological record, this process still does not necessarily ensure that the forecasts' skill will be increased. Such skill degradation can (arbitrarily) arise because the technique does not directly improve the forecast model's reliability by neglecting the conditional relationship between observations and forecasts.

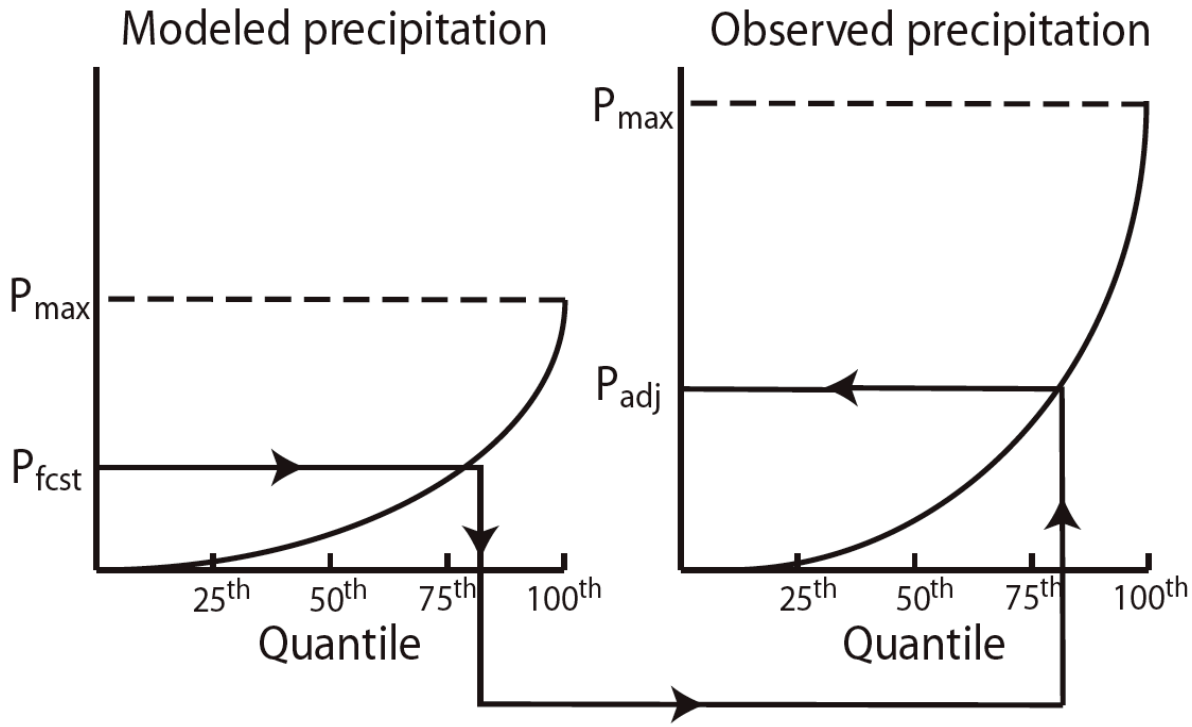


Figure 38: The quantile-to-quantile (q-to-q) correction system. Both modeled and observed precipitation are binned into quantiles. The model precipitation is mapped onto the observed precipitation fields by setting respective modeled quantiles to observed quantiles. In the figure, the forecast precipitation (p_{fcst}) of the 80th quantile is set to observed precipitation (p_{adj}) 80th quantile. The method ensures that the forecasts produce the same number of "no rain" events as the observations.

Appendix C -- Quantile Regression

Quantile regression (QR) has been little used in the atmospheric and hydrologic communities. We introduced this approach for this project, relaying on it heavily to generate ensemble forecasts and forecast error corrections that provide reliable probabilities of what values the river flow will exceedance on a daily basis (such that "flat histograms" of the forecasts are produced – described further in an appendix below).

Introduced in 1978 (Koenker and Bassett, 1978), quantile regression (QR) is an absolute error estimator that can conditionally fit specific quantiles of the regressand distribution (beyond just the mean or median), which does not rely on parametric assumptions of how either the regressand or residuals are distributed. And by virtue of being an ℓ_1-method, the conditional fit is less sensitive to outliers than square error estimators (Koenker and Portnoy 1997). See also Bremnes (2004) for an application of QR to calibrating weather variable output. Specifically, let $\{y_i\}$ represent a set of observations of the regressand y of interest, and $\{\mathbf{x}_i\}$ an associated set of predictor values. Analogous to standard linear regression, a linear function of $\mathbf{x}$ can be used to estimate to a specific quantile q_θ of y

$$q_\theta(\mathbf{x}_i;\boldsymbol{\beta}) = \beta_0 + \sum_{k=1}^{n} \beta_k x_{ik} + r_i \tag{3}$$

with residuals $r_i = y_i - q_\theta(\mathbf{x}_i;\boldsymbol{\beta})$ and $\theta \in (0,1)$. However, instead of minimizing the squared residuals as with standard linear regression, in QR a weighted iterative minimization of $\{r_i\}$ is performed over $\boldsymbol{\beta}$:

$$\min \sum_{i=1}^{n} \rho_\theta(r_i) = \arg\min_{\beta} \sum_{i=1}^{n} \rho_\theta(y_i - q_\theta(\mathbf{x}_i;\boldsymbol{\beta})) \tag{4}$$

with weighting function

$$\rho_\theta(u) = \begin{cases} \theta u & u \geq 0 \\ (\theta - 1)u & u < 0 \end{cases} . \tag{5}$$

In addition to the benefit of being less sensitive to outliers compared to standard linear regression, QR optimally determines the relationship of a regressor set on specific quantiles of the regressor, with no parametric assumptions. This point can be seen in the figure below where, by way of example, we have applied Eqs. (1) – (3) for persistence (previous day's temperature) as a forecast of 24hr July temperature (Salt Lake City airport 1979 to 2001; discussed below) for the 0.1, 0.5 (median), and 0.9 quantiles. The red line is the fit for the central tendency (mean) using standard linear regression; the middle black line is the fit of the median ($q_{0.5}$); upper black line is for $q_{0.9}$; and the lower black for $q_{0.1}$. Notice the similarity but noticeable divergence of the median and mean fits for larger temperatures. Notice also the heteroscedastic behavior of the persistence QR fitting, which is seen by the convergence of the 0.1 and 0.9 quantile lines for higher temperatures. Physically,

such behavior can be justified from higher temperatures being typically associated with high pressure anomalies, which have longer persistence than cold front passages.

Another powerful benefit of QR is that the cost function Eq (3) is precisely associated with the creation of a flat rank histogram verification, which itself is a necessary requirement for a calibrated ensemble forecast. Here, we implicitly define a "calibrated ensemble" as one being equivalent to a random draw from an underlying (but typically unknown) probability distribution function. In addition, although QR constrains the resultant quantile estimators to satisfying this requirement, at the same time it also constrains the estimators to optimal "sharpness" (Wilks 1995; i.e. creating "narrow" forecast PDFs as compared to a purely climatological distribution).

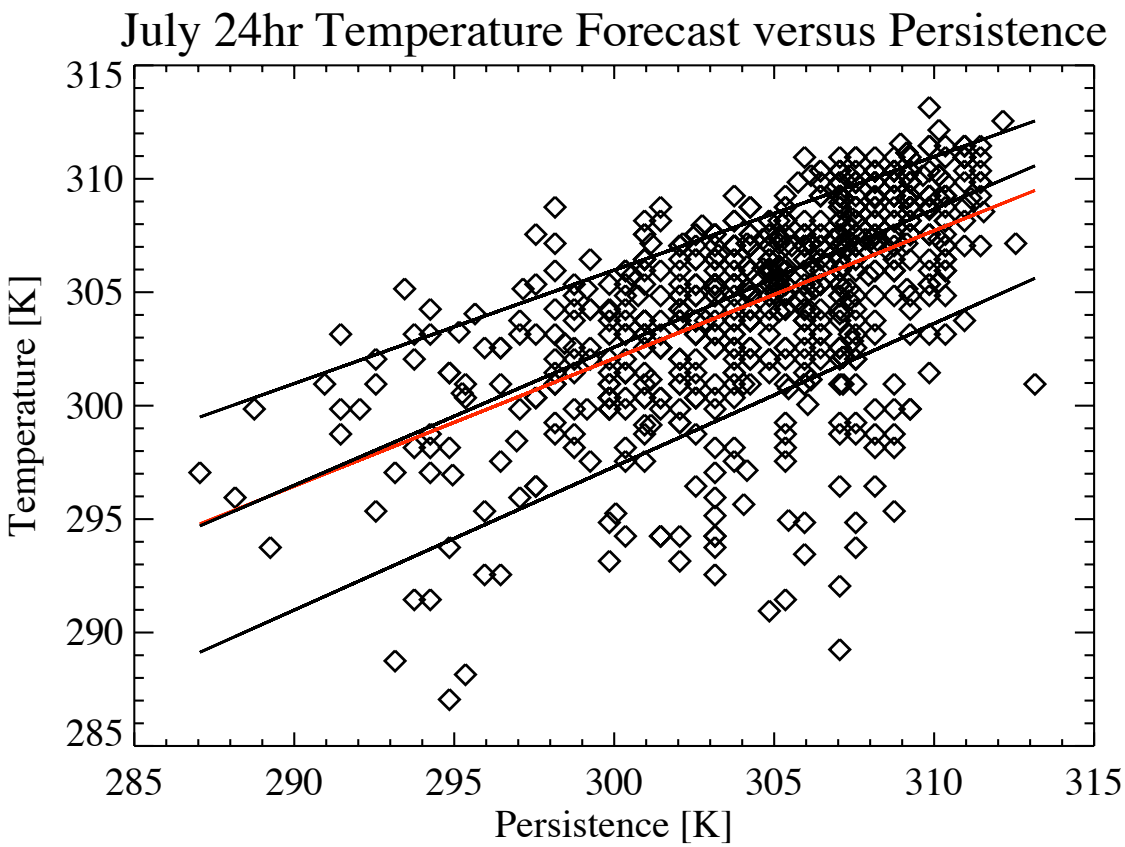


Figure 39: Previous day's temperature (persistence) used as a forecast of 24hr temperature, Salt Lake City airport data, 1979 to 2001. Red line is the fit for the central tendency (mean) using standard linear linear regression; middle black line is the fit of the median (0.5 quantile); upper black 0.9 quantile; lower black 0.1 quantile. Notice the similarity but noticeable divergence of the median and mean fits for larger temperatures. Notice also the heteroscedastic behavior of the persistence fitting, which is seen by the convergence of the 0.1 and 0.9 quantile lines for higher temperatures.

Appendix D -- Verification Metrics

In this appendix we discuss the verification methods applied in this report.

Verification and Evaluation Metrics

In this appendix we briefly discuss the three evaluation metrics we applied in this report to evaluate the rainfall and discharge forecasting systems to assess their skill and calibration: the Brier Score, the Brier Skill Score, and the Rank Histogram.

The Brier Score

The Brier score is a function that measures the accuracy of probabilistic predictions. It is applicable to tasks in which predictions must assign probabilities to a set of mutually exclusive discrete outcomes. The set of possible outcomes can be either binary or categorical in nature, and the probabilities assigned to this set of outcomes must sum to one (where each individual probability is in the range of 0 to 1).

The Brier score can be thought of as a measure of the "skill" of a set of probabilistic predictions. More precisely, across all items $i \in 1...N$ in a set *N* predictions, the Brier score measures the mean squared difference between:

- The predicted probability assigned to the possible outcomes for item *i*
- The actual outcome o_i

Therefore, the lower the Brier score is for a set of predictions, the more skillful the predictions are. Note that the Brier score, in its most common formulation, takes on a value between zero and one, since this is the largest possible difference between a predicted probability (which must be between zero and one) and the actual outcome (which can take on values of only 0 and 1).

The Brier score is appropriate for binary and categorical outcomes that can be structured as true or false.

The most common formulation of the Brier score is

$$BS = \frac{1}{N}\sum_{t=1}^{N}(f_t - o_t)^2$$

In which f_t is the probability that was forecast, o_t the actual outcome of the event at instance t (0 if it does not happen and 1 if it does happen) and N is the number of forecasting instances. In effect, it is the mean squared error of the forecast. This formulation is mostly used for binary events (for example "rain" or "no rain").

For our application, the binary forecasting event we use is whether the discharge will be above or below the upper 75th percentile of flow values (i.e. a moderately

extreme flow event), so we test for how well the forecast system can anticipate high flow values.

The Brier Skill Score

Forecast Skill (or skill score) is a generic term referring to the accuracy and/or degree of association of prediction to an observation or estimate of the actual value of the predictand (i.e., what is being predicted). The term 'forecast skill' can be used both quantitatively and qualitatively. In the former case, skill could be equal to a statistic describing forecast performance, such as the correlation of the forecast with observations. In the latter case, it could either refer to forecast performance according to a single metric, such as our case of the Brier Score. Skill is often, but not exclusively, expressed as the relative representation that compares the forecast performance of a particular forecast prediction to that of a reference, benchmark prediction—a formulation called a 'Skill Score':

$$SS = \frac{A_{forc} - A_{ref}}{A_{perf} - A_{ref}}.$$

For our case, our skill metric "A" is the Brier Score (BS, described above), and the "reference forecast" is the climatological probability of the event occurring (for the upper 75% events, this "climatological probability" is 0.25). In this case, a perfect forecast results in a forecast skill metric of zero, and skill score value of 1.0. A forecast with equal skill to the reference forecast would have a skill score of 0.0, and a forecast which is less skillful than the reference forecast would have unbounded negative skill score values.

The Rank Histogram

The rank histogram is not a verification method per se, but rather a diagnostic tool to evaluate the *spread* of an ensemble. The underlying assumption is that the ensemble member forecasts are distributed so as to delineate ranges or "*bins*" of the predicted variable such that the probability of occurrence of the observation within each bin is equal. For each specific forecast, the bins are determined by ranking the ensemble member forecasts from lowest to highest. The interval between each pair of ranked values forms a bin. If there are N ensemble members, then there will be N+1 bins. The outer bins, lowest and highest – valued, are open-ended.

Rank histograms are prepared by determining which of the ranked bins the observation falls into for each case, and plotting a histogram of the total occurrences in each bin, for the full verification sample. It is desirable to use a large sample of cases so that there is likely to be some occurrences in each of the bins. The examples shown in the figure below highlight ensemble forecasts that are too wide (left panel) and too narrow (right panel).

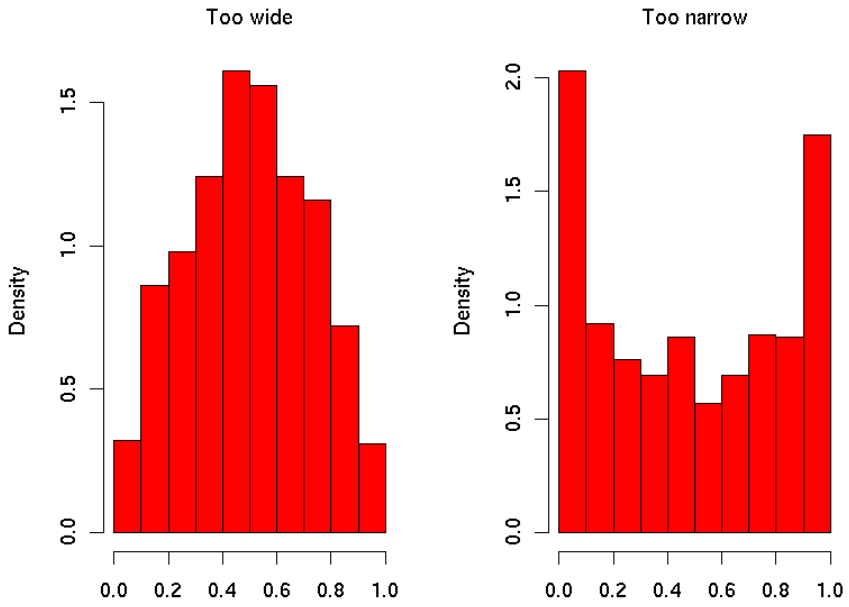


Figure 40: Troubled rank histograms: panel on the left shows an ensemble forecast system whose ensembles are too dispersed (the observation falls too often in the middle of the distribution), while the panel on the right would be for a too narrow ensemble forecasting system. For a perfect ensemble forecasting system, rank histograms would be "flat" (i.e. the observation falling into each bin the same number of times).

References

Amarnath, G., Umer, Y. M., Alahacoon, N., and Inada, Y.: Modelling the flood-risk extent using LISFLOOD-FP in a complex watershed: case study of Mundeni Aru River Basin, Sri Lanka, Proc. IAHS, 370, 131-138, doi:10.5194/piahs-370-131-2015, 2015.

Arulampalam, M. S., Maskell, S., Gordon, N., & Clapp, T. (2002). A tutorial on particle filters for online nonlinear/non-Gaussian Bayesian Tracking. *IEEE Transactions on Signal Processing, 50*, 174-188.

Biancamaria, S., F. Hossain, and D. P. Lettenmaier, 2011: Forecasting transboundary river water elevations from space: RIVER WATER HEIGHT FORECAST FROM SPACE. *Geophysical Research Letters*, **38**, n/a–n/a, doi:10.1029/2011GL047290.

Brakenridge, G. R., Nghiem, S. V., Anderson, E., & Chien, S. (2005). Space-based measurement of river runoff. *EOS, Transactions of the American Geophysical Union, 86*(19), 185-188.

Brakenridge, G. R., Nghiem, S. V., Anderson, E., & Mic, R. (2007). Orbital microwave measurement of river discharge and ice status. *Water Resources Research, 43*, W04405. doi: 10.1029/2006WR005238

Bremnes, J. B., 2004: Probabilistic forecasts of precipitation in terms of quantiles using NWP model output. *Mon. Wea. Rev.,* **132**, 338–347.

CEGIS. (2006). Sustainable end-to-end climate/flood forecast application through pilot projects showing measurable improvements. *CEGIS Base Line Report, 78 pp*.

Ciach, G. J., and W. F. Krajewski, 1999: On the estimation of radar rainfall error variance. *Advances in Water Resources*, **22**, 585–595.

Clark, M. P., Rupp, D. E., Woods, R. A., Zheng, X., Ibbitt, R. P., Slater, A. G., Schmidt, J., Uddstrom, M. J. (2008). Hydrocological data assimilation with the ensemble Kalman filter: Use of riverflow observations to update states in a distributed hydrological model. *Advanced Water Resource, 31*, 1309-1324.

De Groeve, T., R. G. Brakenridge, and S. Paris, 2015: Global Flood Detection System: Data Product Specifications. https://floodobservatory.colorado.edu/Technical%20Note%20GFDS%20Data%20Products%20v1.pdf (Accessed February 10, 2016).

Habib, E., and W. F. Krajewski, 2002: Uncertainty analysis of the TRMM ground-validation radar-rainfall products: Application to the TEFLUN-B field campaign. *Journal of Applied Meteorology*, **41**, 558–572.

Hamill, T. M., 2012: Verification of TIGGE multimodel and ECMWF reforecast-calibrated probabilistic precipitation forecasts over the contiguous United States*. *Monthly Weather Review*, **140**, 2232–2252.

Hirpa, F. A., T. M. Hopson, T. De Groeve, G. R. Brakenridge, M. Gebremichael, and P. J. Restrepo, 2013: Upstream satellite remote sensing for river discharge forecasting: Application to major rivers in South Asia. *Remote Sensing of Environment*, **131**, 140–151, doi:10.1016/j.rse.2012.11.013.

Hopson, T. M., and P. J. Webster, 2010: A 1–10-Day Ensemble Forecasting Scheme for the Major River Basins of Bangladesh: Forecasting Severe Floods of 2003–07*. *Journal of Hydrometeorology*, **11**, 618–641, doi:10.1175/2009JHM1006.1.

Hopson, T. M., 2005: Operational Flood-Forecasting for Bangladesh. Ph.D. Thesis, Univ. of Colorado, 225 pp.

Huffman, G. J., Bolvin, D. T., Nelkin, E. J., Wolff, D. B., Adler, R. F., Gu, G., . . . Stocker, E. F. (2007). The TRMM multisatellite precipitation analysis (TMPA): Quasi-global, multiyear, combined-sensor precipitation estimates at fine scales. *Journal of Hydrometeorology, 8*(1), 38-55.

Huffman, G. J., R. F. Adler, S. Curtis, D. T. Bolvin, and E. J.Nelkin (2005). Global rainfall analyses at monthly and 3-hr time scales. Measuring Precipitation from Space: EURAINSAT and the Future, V. Levizzani, P. Bauer, and J. F. Turk, Eds., Springer, 722 pp.

IPCC, 2013: Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change [Stocker, T.F., D. Qin, G.-K. Plattner, M. Tignor, S.K. Allen, J. Boschung, A. Nauels, Y. Xia, V. Bex and P.M. Midgley (eds.)]. Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA, 1535 pp.

Jongman, B., H. C. Winsemius, J. C. J. H. Aerts, E. Coughlan de Perez, M. K. van Aalst, W. Kron, and P. J. Ward, 2015: Declining vulnerability to river floods and the global benefits of adaptation. *Proceedings of the National Academy of Sciences*, **112**, E2271–E2280, doi:10.1073/pnas.1414439112.

Joyce, R. J., J. E. Janowiak, P. A. Arkin, and P. Xie (2004). CMORPH: A method that produces global precipitation estimates from passive microwave and infrared data at high spatial and temporal resolution. J. Hydrometeor., 5, 487–503.

Kleuskens, M., Westerhoff, R. S., & Huizinga, J. (2011). *Operational flood mapping: A pilot study in the Mekong Area*. Paper presented at the Proceedings of the International Symposium of Remote Sensing of the Environment, Sydney, Australia.

Krajewski, W. F., G. J. Ciach, and E. Habib, 2003: An analysis of small-scale rainfall variability in different climatic regimes. *Hydrological Sciences Journal*, **48**, 151–162, doi:10.1623/hysj.48.2.151.44694.

Koenker, R. and W. Portnoy, 1997: The Gaussian Hare and the Laplacean Tortoise: Computability of Squared-error vs. Absolute Error Estimators. Statistical Science, 12, 279-300.

Koenker, R. W. and G. W. Bassett, 1978: Regression Quantiles. Econometrica, 46, 33-50.

Lee, H. S., Seo, D. J., Lui, Y., Koren, V., McKee, P., & Corby, R. (2012). Variatinal assimilation of riverflow into operational ditributed hydologic models: Effect of spatiotemporal scale adjustment. *Hydrology and Earth System Sciences, 16*, 2233-2251. doi: 10.2194/hess-16-2233-2012

Mitra, A. K., G. R. Iyengar, V. R. Durai, J. Sanjay, T. N. Krishnamurti, A. Mishra, and D. R. Sikka, 2011: Experimental real-time multi-model ensemble (MME) prediction of rainfall during monsoon 2008: large-scale medium-range aspects. *Journal of earth system science*, **120**, 27–52.

Montanari, M., Hostache, R., Matgen, P., Schumann, G., Pfister, L., & Hoffman, L. (2009). Calibration and sequential updating of a coupled hydrolic-hydrualic model using remote sensing-derived water stages. *Hydrology and Earth System Sciences, 13*, 367-380. doi: 10.5194/hess-13-367-2009

Moradkhani, H., Hsu, K., Gupta, H. V., & Sorooshian, S. (2005). Uncertainty assessment of hydrologic model states and parameters: Sequential data assimilation using particle filter. *Water Resources Research, 41*, W05012. doi: 10.1029/2004WR003604

Mylne, K. R., 1999: The use of forecast value calculations for optimal decision making using probability forecasts. Preprints, 17th Conf. on Weather Analysis and Forecasting. Denver, CO, Amer. Meteor. Soc., 235-239.

Prakash, S., A. K. Mitra, I. M. Momin, D. S. Pai, E. N. Rajagopal, and S. Basu, 2015: Comparison of TMPA-3B42 Versions 6 and 7 Precipitation Products with Gauge-Based Data over India for the Southwest Monsoon Period. *Journal of Hydrometeorology*, **16**, 346–362, doi:10.1175/JHM-D-14-0024.1.

Ranade, A., A. K. Mitra, N. Singh, and S. Basu, 2014: A verification of spatio-temporal monsoon rainfall variability across Indian region using NWP model output. *Meteorology and Atmospheric Physics*, **125**, 43–61, doi:10.1007/s00703-014-0317-5.

Revilla-Romero, B., H. E. Beck, P. Burek, P. Salamon, A. de Roo, and J. Thielen, 2015: Filling the gaps: Calibrating a rainfall-runoff model using satellite-derived surface water extent. *Remote Sensing of Environment*, **171**, 118–131, doi:10.1016/j.rse.2015.10.022.

Richardson, D. S., 2000: Skill and economic value of the ECMWF ensemble prediction system. Quart. J. Roy. Meteor. Soc., 126, 649-668.

Rosenberg, E. A., A. W. Wood, and A. C. Steinemann, 2013: Informing Hydrometric Network Design for Statistical Seasonal Riverflow Forecasts. *Journal of Hydrometeorology*, **14**, 1587–1604, doi:10.1175/JHM-D-12-0136.1.

Seo, D. J., Cajina, L., Corby, R., & Howieson, T. (2009). Automatic state updating for operational riverflow forecasting via variational data assimilation. *Journal of Hydrology, 367*, 255-275.

Song, X. Y., & Lee, S. Y. (2004). Bayesian analysis of two-level nonlinear structural equation models with continuous and polytomous data. *British Journal of Mathematical and Statistical Psychology, 57*, 29-52.

Sunilkumar, K., T. Narayana Rao, K. Saikranthi, and M. Purnachandra Rao, 2015: Comprehensive evaluation of multisatellite precipitation estimates over India using gridded rainfall data: EVALUATION OF MPES OVER INDIA. *Journal of Geophysical Research: Atmospheres*, **120**, 8987–9005, doi:10.1002/2015JD023437.

Syvitski, J. P. M., & Brakenridge, G. R. (2013). Causation and avoidance of catastrophic flooding along the Indus river, Pakistan. *GSA Today, 23*(1). doi: 10.1130/GSATG1165A.1131

Webster, P. J. (2013). Improve weather forecasts for the developing world. *Nature, 493*, 17-19.

Webster, P. J., Jian, J., Hopson, T. M., Hoyos, C. D., Agudelo, P., Chang, H.-R., . . . Subbiah, A. R. (2010). Extended-range probabilistic forecasts of Ganges and Brahmaputa floods in Bangladesh. *Bulletin of the American Meteorological Society, 91*, 1493-1514.

Webster, P. J., T. Hopson, C. Hoyos, A. Subbiah, H-. R. Chang, R. Grossman, 2006: A three-tier overlapping prediction scheme: Tools for strategic and tactical decisions in the developing world. In Predictability of Weather and Climate, Ed. T. N. Palmer, Cambridge University Press.

Weerts, A. H., & El Serafy, G. Y. H. (2006). Particle filtering and ensemble Kalman filtering for state updating with hydrological conceptual rainfall-runoff models. *Water Resources Research, 42*, W09403. doi: 10.1029/2005WR004093

Westerhoff, R. S., Huizinga, J., Kleuskens, M., Burren, R., & Casey, S. (2010). *Operational satellite-based flood mapping using the Delft-FEWS System.* In Proceedings of the ESA Living Planet Symposium, Bergen, Norway, June 28-July 2, 2010. Retrieved Jan 10, 2013, from http:/kennisonline.deltares.nl/product/22381

Wilks, D. S., 2001: A skill score based on economic value for probability forecasts. Meteor. Appl., 8, 209-219.

Wilks, D. S., 1995: Statistical Methods in the Atmospheric Sciences. Cambridge Press, 467 pp.

Xie P. P., B. Rudolf, U. Schneider, and P.A. Arkin, 1996: Gauge-based monthly analysis of global land precipitation from 1971 to 1994. *J. Geophys. Res. - Atmos.,* 101 (D14), 19023-19034.

Zhu, Y., Z. Toth, R. Wobus, D. Richardson, K. Mylne, 2002: The economic value of ensemble-based weather forecasts. Bull. Amer. Meteor. Soc., 83, 73-83.