Improving Hydrologic Disaster Forecasting and Response for Transportation by Assimilating and Fusing NASA and other Data Sets

Final report

April, 2017

Principal Investigator: ***Xu Liang***
University of Pittsburgh

Program Manager: ***Caesar Singh***
US Department of Transportation

Contract Number: ***0038672***

Table of contents

Table of contents 2
Glossary of terms 3
Executive summary 5
Chapter 1. Background and objectives 6
1.1 Background and problem statement 6
1.2 Objectives 7
Chapter 2. Research methodologies 8
2.1 Data acquisition 10
2.2 Data fusion 11
2.3. Hydrologic modeling 12
2.4 Severity assessment 14
Chapter 3. Findings and conclusions 16
Chapter 4. Ongoing work 24
Appendix – References 25

Glossary of terms

This section contains a glossary of technical or specialized terms, as well as acronyms, that are used throughout the report and possibly in future reports.

AGU American Geophysical Union (https://sites.agu.org/)

AMSR-E NASA's Advanced Microwave Scanning Radiometer – Earth Observation System (http://wwwghcc.msfc.nasa.gov/AMSR/)

C Functional programming language

C++ Object-oriented extension to C

Data assimilation The process of combining data values from different sources, typically from models and measurements, to produce more reliable estimates of the initial conditions of a model to be used for forecasting purposes

DHSVM "Distributed Hydrology Soil Vegetation Model": High resolution land-surface modeling engine for hydrologic simulations (http://www.hydro.washington.edu/Lettenmaier/Models/DHSVM)

GES-DISC NASA's Goddard Earth Science Data and Information Services Center (http://daac.gsfc.nasa.gov/)

GIS Geographic Information System

GRASS Geographic Resources Analysis Support System (https://grass.osgeo.org/)

GUI Graphical User Interface

IUPUI Indiana University/Purdue University of Indianapolis

LPRM NASA's Land Parameter Retrieval Model, which include satellite-based soil moisture estimates

HDFR Hydrologic Disaster Forecasting and Response

MongoDB An open-source non-SQL document database management system (http://www.mongodb.org/)

NASA National Aeronautics and Space Administration of the United States (http://www.nasa.gov/)

NLDAS North America's Land Data Assimilation Systems: dataset provided by the GES-DISC. It features meteorological and hydrological maps created from land-based observations. URL: http://ldas.gsfc.nasa.gov/nldas/

NOAA National Oceanic and Atmospheric Administration of the United States (http://www.noaa.gov/)

Noah Low resolution land-surface model for hydrologic simulation (http://www.emc.ncep.noaa.gov/mmb/gcp/noahlsm/)

NSIDC National Snow & Ice Data Center (http://nsidc.org/)

NWS NOAA´s National Weather Service (http://www.weather.gov/)

OPeNDAP Open-source Project for a Network Data Access Protocol, a series of software tools and standards for the transferring of scientific data (https://www.opendap.org/)

OPTIMISTS Optimized PareTo Inverse Modeling through Integrated Stochastic Search: proposed hybrid Bayesian/variational data assimilation algorithm

PostGIS Extension for the PostgreSQL database system to include GIS capabilities (http://postgis.net/)

PostgreSQL An open-source multi-platform relational database management system (http://www.postgresql.org/)

PyGRASS Python application programming interface for GRASS GIS (https://grasswiki.osgeo.org/wiki/Python/pygrass)

Python An interpreted high-level programming language (https://www.python.org/)

Qt IDE for C++ development and a series of logical and graphical libraries (http://qt-project.org/)

Return period The average time between events of a given level of intensity. Commonly used to measure the severity of hydrologic events with typical values between 2 and 100 years.

RWIS PennDOT's Roadway Weather Information System

SAGA "System for Automated Geoscientific Analyses" GIS (http://www.saga-gis.org/)

SMAP NASA's Soil Moisture Active-Passive mission (http://smap.jpl.nasa.gov/)

SSW NASA's Simple Subset Wizard (http://disc.gsfc.nasa.gov/ssw/)

USGS United States Geological Survey (http://www.usgs.gov/)

VIC Large-scale land-surface modeling engine for hydrologic simulations (http://vic.readthedocs.io/en/develop/)

Vistrails Open-source scientific workflow platform (https://www.vistrails.org)

Executive summary

In this 3-year project, the research team developed the Hydrologic Disaster Forecast and Response (HDFR) system, a set of integrated software tools for end users that streamlines hydrologic prediction workflows involving automated retrieval of heterogeneous hydro-meteorological data from multiple sources in near real-time, the computation of critical variables to assess and forecast hydrologic disasters using modern distributed hydrologic model, and data assimilation techniques. The system is intended to be deployed as a decision-support tool in operations where extensive areas need to be monitored for extreme weather events and/or where accurate hydrologic predictions are required.

The HDFR has been developed and built as a series of modules grouped under four categories:

1. ***Data:*** These modules allow to automatically download information from multiple servers hosted by data providers such as government agencies, comprising meteorological and hydrological observations from land and space-borne sensors, and model predictions such as weather forecasts.
2. ***Fusion:*** Enable the combination or "fusing" of observations and/or simulations from different instruments and/or models for generating more accurate estimates.
3. ***Modeling:*** Allow the creation of hydrologic models and provide tools to estimate their parameters and initial conditions to maximize the correspondence of the simulations with the observations for improved predictive power.
4. ***Severity:*** Contrast current or forecasted conditions with historical observations to assess threat levels and allow for efficient response actions.

Most of these modules were incorporated into the Geographic Resources Analysis Support System (GRASS), a popular open-source geographic information system, so that complex simulation workflows (from data acquisition to model result analysis and visualization) can be executed in a unified environment without requiring numerous external tools. Within GRASS, information is organized in a unified place with multiple options for data import and export, and for interoperability between the HDFR's modules and other general-purpose routines.

The research team consisted of researchers from the University of Pittsburgh, Indiana University - Purdue University Indianapolis (IUPUI), and NASA's Goddard Earth Science Data and Information Services Center (GES-DISC/ADNET). The team also partnered with the Pennsylvania Department of Transportation (PennDOT) for part of the cost sharing and for evaluating the incorporation of the HDFR into transportation infrastructure monitoring operations (for example for the determination of threatened bridges following severe weather).

While most of the individual modules of the HDFR were completed and tested, a late start date of the matching fund with PennDOT leads to delays in the full completion of the HDFR's development. The team will therefore continue the work on the HDFR system throughout 2017. In this report the completed activities of this project for each of the tasks originally identified are described, together with pending activities to be delivered in early 2018.

Chapter 1. Background and objectives

1.1 Background and problem statement

The complexity of Earth systems has always denied a sufficient level of understanding to enable sustainable management of water resources as well as effective protection from natural threats—especially in the face of increasing human needs. These deficiencies are of special concern in the transportation sector, where government agencies are tasked with the monitoring and maintenance of huge amounts of infrastructure spread over large regions. For example, the Pennsylvania Department of Transportation (PennDOT) oversees over 3,000 scour-critical bridges, scattered throughout the state, which could be affected during severe weather conditions. Therefore, current severity analysis procedures lack in both the level of accuracy and promptness required to layout agile and specific response actions during and after the extreme conditions.

Modern technologies offer powerful tools to assist in addressing these challenges: remote sensing missions provide an increasingly broad window into current environmental conditions; powerful computer systems allow running detailed geophysical simulations; and advanced Artificial Intelligence/Machine Learning algorithms help in reducing the uncertainty in these models' numerous unknowns. These advances combined should be able to make profound impacts on decision-making processes related to water resources management, and to disaster prevention and response.

However, there exists a major gap between nationwide and global operation efforts and the prompt accessibility to these tools. On one end, data products from federal agencies are available through heterogeneous sources, transfer protocols, and data formats—a hindrance that distances information from its end users. On the other end, when scientists and engineers turn their attention to local studies, they similarly find a high cost in learning and implementing tools for creating, configuring, and optimizing models. This occurs despite the plethora of available open-source modeling engines and the wealth of calibration and data assimilation methods in the scientific literature. An additional challenge exists in the connection between data and meaningful improvements to modeling efforts.

1.2 Objectives

The main objective of this project is to develop a software system that streamlines hydrologic prediction workflows in order to improve forecast and response operations in large regions regarding hydrologic-related threats that could jeopardize the transportation infrastructure. Users of the system will be able to automatically download data from multiple sources, create and configure distributed hydrologic models, use data to optimize their parameters and state variables, assess the severity of past and future events, and visualize and export results in standard formats, all using a single graphical user interface (GUI). With the proposed system, named Hydrologic Disaster Forecast and Response (HDFR), users will not have to rely on additional tools for data acquisition, analysis, and visualization.

Specific objectives include:

1. Develop an open-source software system that can store information of different types (points in space, polygons, time series, 2D and 3D grids) in a centralized manner (temporary storage), and that can persist this information from a session to the next (permanent storage).
2. Create a module for the HDFR that can connect with NASA's Simple Subset Wizard[1] (SSW), a web portal that allows accessing and preprocessing (sub-setting to a desired spatiotemporal extent) data from multiple data centers at NASA and beyond.
3. Develop a series of modules to access remote hydro-meteorological data from a number of different online portals hosted by data providers (almost exclusively government agencies at present).
4. Extend the Multiscale Kalman Smoother (MKS) algorithm for data fusion [1] and incorporate it within HDFR workflows.
5. Test the modified MKS module with some of the data modules developed.
6. Couple two hydrologic modeling engines with the HDFR so that users can run simulations of watersheds for more accurate and varied (i.e., for more variables like soil moisture and evapotranspiration) predictions.
7. Allow HDFR users to calibrate models so that they adequately match observations.
8. Develop a data assimilation algorithm that enables users to adequately initialize the state variables of their models so that more accurate forecasts can be run.
9. Develop a series of modules that allow assessing the severity of current or future extreme events, mainly for precipitation and streamflow but also for other variables.
10. Integrate the HDFR with PennDOT's Intelligent Transport System (ITS) Traffic Management Centers (TMCs) (in a loose form).

[1] https://disc.gsfc.nasa.gov/SSW/

Chapter 2. Research methodologies

During the first stages of the project the team experimented with two main alternatives:

1. Developing its own data system in a custom C++ program, which would address the permanent storage requirement by connecting to a customized database. The team experimented using Qt[2] libraries, PostgreSQL[3] (with or without PostGIS[4]), and MongoDB[5].
2. Extending an existing software system that already contained (partial) solutions for temporal and permanent storage of most of the data types that were expected to be used, to address the specific needs. The team investigated Geographic Information Systems (GISs) such as SAGA[6] and GRASS[7] for this purpose, together with the Vistrails[8] workflow engine.

After this exploratory phase the team settled for using GRASS (Geographic Resources Analysis Support System) GIS for the following reasons:

- It has a robust built-in data model that accommodated most of the project's needs, together with multiple import and export formats for interoperability, and multiple choices for persisting the information.
- The familiarity of PennDOT's team with GISs.
- A graphical user interface (GUI) able to display multiple types of geographic information.
- A large library of modules which could be utilized to support some of the HDFR's planned capabilities, mostly related to data formatting and preprocessing tools.
- The availability of multiple alternatives to develop extensions to the base functionality.
- The support for temporal datasets, in terms of representation, operability, and visualization.
- A mature community to rely on for support on the usage of existing tools and the development of new ones.

The HDFR was thereafter developed as a series of plugins or extensions to GRASS GIS. These extensions were developed in Python using GRASS' PyGRASS[9] application programming interface, often connecting to external software developed in C++. These extensions represent the different modules that make up the HDFR and each of them includes a form available from the GRASS GUI for the user to specify their inputs and parameters.

Individual modules are integrated into the HDFR by writing two pieces of software: one which contains the logical part of the module (its main functionality), and the other (the wrapper part) which connects the logical part in a format that can be loaded and understood

[2] https://qt-project.org/
[3] https://www.postgresql.org/
[4] http://www.postgis.net/
[5] https://www.mongodb.com/
[6] http://www.saga-gis.org/
[7] https://grass.osgeo.org/
[8] https://www.vistrails.org/
[9] https://grasswiki.osgeo.org/wiki/Python/pygrass

by the GRASS GUI. The logical part is, in many instances, implemented in C++ and then compiled into an executable file. In the case of the DHSVM and VIC modules, the actual functionality is provided by the executable file of the third party developer. The wrapper part of the modules is written in Python using PyGRASS. These parts are usually much simpler as they only contain mappings between the parameter assignments provided by the user and the executables' parameters, and between the obtained results (usually in the form of files in the local system) and GRASS storing and visualization tools. Additionally, each module must define a set of fields to be displayed in a form so that the users can select the desired parameter assignments.

Modules in the HDFR are also designed in a way that they easily integrate with one another. Figure 1 illustrates the relationships between the different modules of the HDFR from the perspective of operational hydrologic prediction. The data modules provide information that can be first fused and then used for creating and tuning hydrologic models. Models are optimized first through offline model parameter calibration, and then online in the operational pipeline through data assimilation. The obtained predictions can then be analyzed to determine their severity.

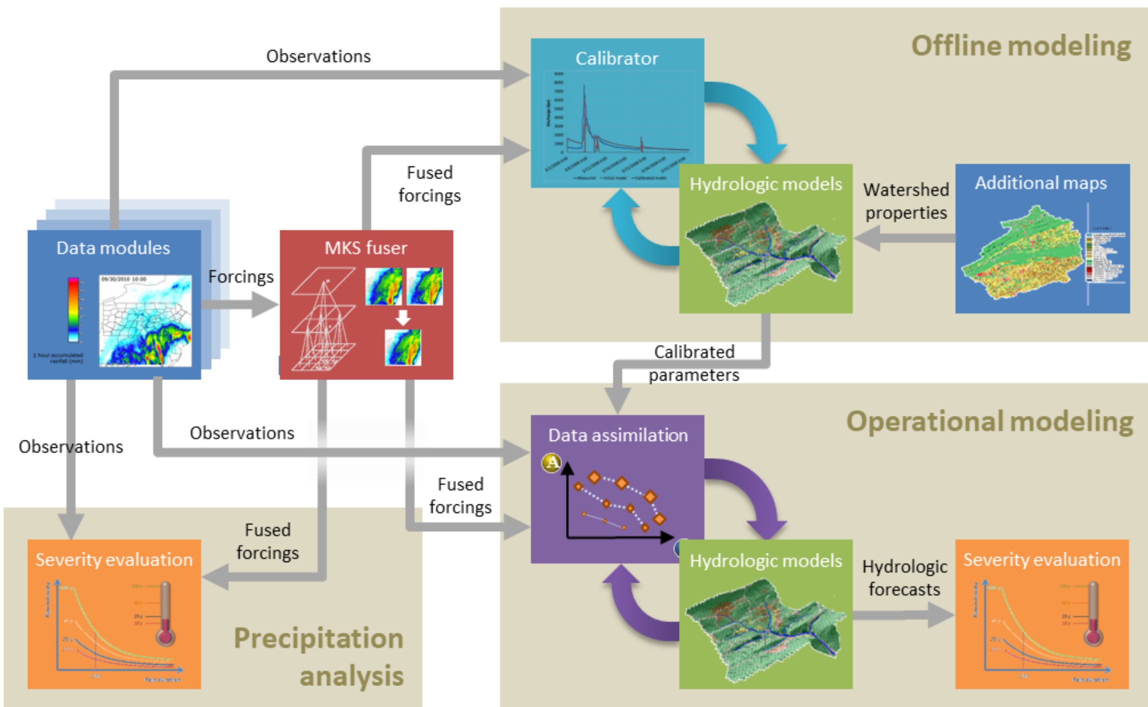

Figure 1. Illustration of the relationships among the modules of the HDFR. Grey arrows represent information flow from one component to another. Modules grouped under "Offline modeling" are expected to be used in the preparation of information to be consumed by those grouped under "Operational modeling."

The following subsections describe each of the main components of the HDFR and their related modules.

2.1 Data acquisition

Without a unifying system like the HDFR, users usually would have to resort to dealing with the specificities of each source of information in terms of the portal where it is available, the transfer protocol, the file formats, and the geographical projection. This often necessitates the use of multiple software packages. With the developed modules in the HDFR, these tasks are not only greatly simplified but can be realized from machine to machine in an automated way in near real-time fashion. Table 1 lists all of the data modules that make part of the HDFR, the type of information available through them, the spatiotemporal extent, and their current level of completion.

The original plan was to create a module to access PennDOT's Roadway Weather Information System (RWIS) data, but the RWIS system is still under development and its data products are unavailable. Similarly, the team decided not to develop a module to download terrain elevation information given that the data are virtually static and only require being downloaded once for a specific area. On the other hand, additional modules, that were not initially proposed, were developed; namely, TRMM, GPM, SMAP, and NAM.

Table 1. List of data modules in the proposed system.

Dataset	Variables; source	Spatial extent	Temporal extent	LC*
NASA's NLDAS-2	Multiple meteorological; land stations, models	Continental U.S., 12 km grid	Hourly, 4 d lag, since 1979	2
NASA's GPM	Precipitation; satellite radar, radiometer	Global, 0.1° grid	30 min, 6 h lag, since 2014	2
NASA's SMAP	Soil moisture; satellite radiometer	Global, 40 km grid	Daily, 50 h lag, since 2015	1
NASA's MODIS snow	Snow data; satellite	Global, 0.05° grid	Daily, 3 h lag, since 2000	3
NASA's LPRM	Soil moisture; satellite	Global, 10 km grid	Daily, 1 d lag, since 2012	0
NASA's TRMM	Precipitation; satellite radiometer	Global, 0.25° grid	3-hour, 1 d lag, 1998-2013	2
USGS' water data	Gage height, discharge; hydrometric stations	United States, > 10,000 sites	Per site	2
NWS RFC precipitation	Precipitation; multi-sensor (primarily NEXRAD)	Continental U.S., 4 km grid	Hourly, 1 h lag, since 2013	3
NOAA's SNODAS	Snow data; multi-sensor	United States, 0.7 km grid	Daily, 4 h lag, since 2003	3
NOAA's METAR	Precipitation; rain gauges	Global	Per site	2
NWS' GFS	Multiple hydro-meteorological; atmospheric/land-surface model	Global, 1° grid	3-hour, every 6 h, 192 h lead time	3
NWS' NAM	Multiple hydro-meteorological; atmospheric/land-surface model	North America, 12 km grid	Hourly, every 3 h, 60 h lead time	2

* "LC" stands for level of completion: 0-no progress, 1-module can download and interpret data; 2-module allows selection of custom spatial domains; 3-module has been integrated into GRASS GIS.

The team members at NASA's GES-DISC/ADNET supported the project by implementing access to additional products that were required for the project through the SSW. The work focused on incorporating two sets of soil moisture data products, SMAP and LPRM. The suite of LPRM products are archived at the GES DISC. Of those added, the LPRM-AMSR2 products are forward-processed and, thus, most relevant. The SMAP products are also forward-processed.

Adding the SMAP products, however, turned out to be complicated because the way they are archived at NSIDC is not entirely compatible with the normal SSW process, specifically, the way geolocation is represented. A new OPeNDAP agent had to be developed to work with NSIDC's OPeNDAP implementation for SMAP. Currently, while seven of the eight specified SMAP products have been added to the SSW, the issues remain and are being worked on. The eighth product probably will need some corrective work related to the HDF5 handler by the HDF Group. The team will continue to work on these remaining issues until they are resolved.

2.2 Data fusion

Data fusion consists of taking multiple sources of information on the same variable but with different spatial resolutions and accuracy levels (e.g., precipitation gauges and Doppler radars), and producing a combined estimate that is expected to better represent reality. The MKS-based fusion module in the HDFR is intended to allow users to perform analyses of precipitation events (or of other variables such as temperature and snow cover) by taking advantage of the multiple sources of information simultaneously, thus resulting in reduced bias and uncertainty. For example, precipitation analyses can be performed after fusing data from NLDAS (rain gauges), NWS precipitation (land radar), and GPM/TRMM (satellite radar). Snow analyses can be performed after fusing SNODAS and MODIS data.

The MKS produces fused estimates by propagating information between levels in a fixed scale hierarchy in two sweeps, one upward and one downward. It also allows to modify the accuracy level of all data sources such that the resulting fused estimate has a maximum consistency with the observations after performing the double-sweep process. Because of its formulation, the original MKS algorithm has a few constraints. Table 2 lists these constraints and the modifications performed during the development of the HDFR to lift them. The MKS module, now integrated into GRASS, also allows performing all input and output transfers in memory without having to resort to a database as it was done in the original version.

Table 2. Constraints of the MKS algorithm and the modifications performed to lift them.

Constraint	**Modification**
Applicability only to square regions.	Regions of arbitrary shape now can be used and the HDFR automatically fits them into a target square to perform the analysis. The MKS now admits point information as input as well.
All sources of information must cover the exact same area.	Input information can have arbitrary resolutions and extents, with the HDFR performing the adequate preprocessing steps (i.e., resampling and trimming) before invoking the MKS.
Input information must be gridded and must conform to a specific resolution hierarchy (i.e., with 1^2, 2^2, 4^2, 8^2, 16^2 … number of cells).	Regions of arbitrary shape can now be used and the HDFR automatically fits them into a target square to perform the analysis.
Output information is only available at those same resolution scales.	Outputs can now be resampled to arbitrary projections as defined by the user.
Only one source of information can be defined for each scale.	MKS now allows multiple inputs at each scale.

2.3. Hydrologic modeling

The team initially considered using the Noah[10] engine in any of its multiple versions for large and medium watersheds. However, after finding multiple problems, such as the difficulty of compiling several of the most complete versions of Noah on Windows—which is the PennDOT's choice for an operating system, the team settled for using the Variable Infiltration Capacity (VIC) [2] modeling software instead. VIC is an open-source C package that allows modeling an array of squared soil columns that together represent the watersheds. In each column, water can move between the atmosphere, vegetation, snow pack, surface, and a series of soil layers, allowing for heterogeneous definitions of the surface's characteristics. The version selected also includes a coupled algorithm to perform routing simulations through the land surface and through the channel network [3]. Some modifications on VIC's source code were done, principally to allow the conditions of the channels to be stored, in order to be able to interrupt simulations to be continued afterwards. This is especially important for allowing data assimilation (see below).

Although optimized forecasting simulations using VIC are able to be performed, the corresponding modules are not coupled with the HDFR on GRASS given PennDOT's priorities on small bridges (as will be explained in the Results and Discussion chapter) which would be managed using the other modeling engine. Moreover, the addition of the NAM forecasting module, which operates at a similar spatial and temporal resolution as was planned for the VIC module, would mostly be able to fill in this gap. Nonetheless, PennDOT has been made aware of the future benefits of this work.

[10] http://www.ral.ucar.edu/research/land/technology/lsm.php

For those small watersheds, the Distributed Hydrology Soil Vegetation Model (DHSVM) [4] is used as originally proposed, which is also an open source package developed in C. However, DHSVM is tuned for much higher resolutions (smaller than 100 m). Like VIC, soil water in DHSVM is moved vertically within each soil column; and the lateral water movement is realized through routing. Unlike VIC, the routing process is already part of DHSVM and one does not need an external routing model. However, the DHSVM required much larger efforts to modify in order to meet the project's needs. The modifications included:

- Fixing of numerous bugs in the open source codes of DHSVM (e.g., existence of unrealistic sinks, unrealistic identification of channel locations, calculation of surface water depth, and bugs related to runoff generation and overland flow routing).
- Modification of the original 4-direction algorithm to the 8-direction algorithm in the routing scheme of DHSVM. This improvement makes the flow direction consistent with the most modern GIS methods.
- Development of a new routing module for DHSVM using a full implicit or a simplified implicit MacComack method to make it computationally feasible to deal with watersheds of a reasonably large size.
- Correct initialization and computation of the water storage in the channel network.

After performing these corrections/modifications/developments, and a set of tests over two watersheds, a GRASS module was created to interact with the DHSVM. From GRASS's GUI, the user can determine the outlet of the target watershed, and the module will automatically generate all the input files based on the information of elevation map (including all of the channel network's properties). After defining input maps for the soil and vegetation types, and selecting the desired meteorological forcing (which can be from several of the data modules), the DHSVM is run without requiring the time consuming and tiresome pre-processing which involves the manual determination and formatting of all input information—which is made worse due to the lack of an associated GUI.

For the purpose of allowing HDFR users to calibrate models created to be run on VIC and the DHSVM so that they adequately match observations, and given that uncalibrated models often produce very large errors, an evolutionary multi-objective optimization algorithm was extended and coupled with both VIC and the DHSVM. The implemented method allows selecting the parameter values of the models that minimize the discrepancies with the observations (mainly of streamflow). The calibration algorithm implements a state of the art technique which consists of using an ensemble of multiple low-level optimization algorithms (in this case a genetic algorithm, a hybrid between ant colony optimization and Metropolis-Hasting sampling, and a non-convex gradient descent method) that are invoked alternately and adaptively to better maneuver the solution space at different stages of the calibration process.

While calibration allows to determine parameter sets that lead to adequate behavior of the model, a big portion of uncertainty is still present within the initial conditions or initial states. It is also important to acknowledge that the number of observations generally available and the inherently incomplete representations in hydrologic models yields a large degree of uncertainty in any estimates that should not be underestimated. For this reason, data assimilation should not produce deterministic estimates of such state variables, but probabilistic ones instead.

Therefore, the team implemented a novel data assimilation algorithm which is named OPTIMISTS (after Optimized PareTo Inverse Modelling through Integrated STochastic Search). This algorithm replaces the originally proposed hybrid dual-state data assimilation framework. OPTIMISTS hybridizes the two most popular families of data assimilation techniques in the modern literature in an attempt to combine the more advantageous characteristics of them both: Bayesian data assimilation, which produces probabilistic estimates sequentially from which states can be randomly sampled; and variational data assimilation, which uses optimization algorithms to create deterministic state estimates that minimize errors. A manuscript was submitted to the Hydrology and Earth System Sciences journal describing OPTIMISTS and a series of forecasting tests in detail [5].

2.4 Severity assessment

The first module that was created is focused on precipitation alone, and it is meant to replace the current system used at PennDOT to identify bridges that require inspection due to potential scouring damage after extreme weather conditions. The module receives a precipitation event as input (which could be from any of the sources described before or a fusion of them) and computes the return period of the event at every precipitation pixel. The return period corresponding to an observed or predicted storm is interpolated based on the regionally-distributed Intensity-Duration-Frequency (IDF) curves developed by PennDOT[11]. The bridges to inspect, from which there are over 3,000 possible candidates in the state of Pennsylvania (shown in Figure 2), are then selected by comparing their degree of vulnerability to scouring with the return periods corresponding to their drainage areas (or watersheds).

This module works in very similar ways to the system PennDOT currently uses. However, a second "multi-duration" module that was developed analyzes the storm by taking into account multiple precipitation accumulation periods and selects the most severe one. In this way, not only can the very fast storm events (such as those responsible for flash floods) be detected automatically, but also those associated with longer storm periods. At present, PennDOT's system requires a labor-intensive process in order to identify the most severe accumulation period for each storm at each bridge location. This significant reduction in necessary manpower, combined with the HDFR's ability to automatically obtain precipitation information from multiple sources (including forecasts) and to prepare better estimates through fusion, provides significant advantages over the existing system.

For other variables, a module that enables the execution of frequency analyses on arbitrary time series was developed to establish severity curves similar to the IDF ones in the case of precipitation. The module first samples the most extreme events in the multi-annual time series using standard methods (e.g., annual maxima, partial-series maxima, and exceedance maxima) and then adjusts a probability distribution that can then be sampled for interpolating or extrapolating the return period of observed events. The probability distribution with the least overall fitting error, among maximum likelihood estimates of Log-Normal, Gumbel, and Pearson Type-III distributions, is selected. Time series from any of the data modules can be used for this purpose.

---

[11] Described in PennDOT's Drainage Manual, 2010, Chapter 7, Appendix A (https://www.dot.state.pa.us/public/bureaus/design/PUB584/PDMChapter07A.pdf).

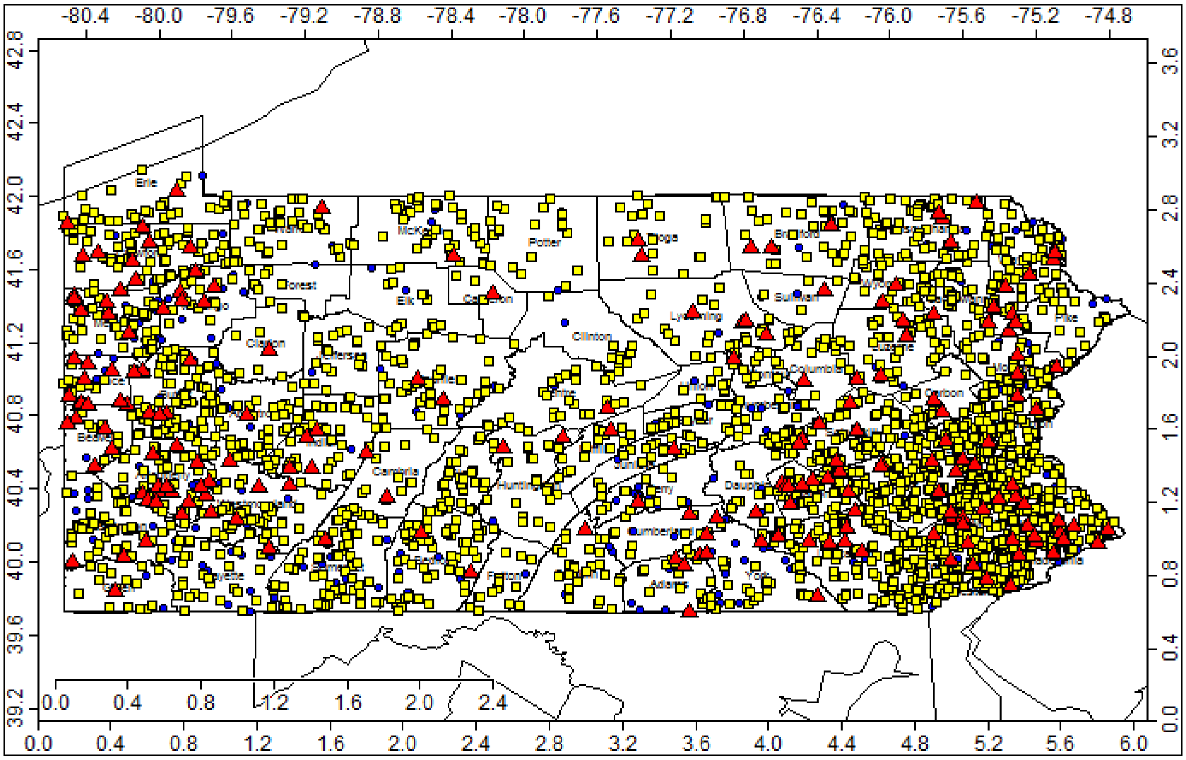

Figure 2. PennDOT's scour critical bridges. Blue circles: low vulnerability; yellow squares: medium vulnerability; red triangles: high vulnerability.

Finally, the team developed a module to approximate the streamflow return period of ungauged watersheds based only on observed precipitation and some general characteristics related to ungauged watersheds. The approximation was made through means of a coupled double regression (one for estimating the amount of water stored in the watershed and the other for estimating the outflow based on said storage) adjusted using the streamflow return values computed for the 15 test watersheds. The regression equations make use of the watersheds' area, slope, flow path length, soil porosity, and forest coverage values.

Chapter 3. Findings and conclusions

As explained in the previous chapter, the team decided to develop the HDFR as an extension to GRASS GIS, a system which offers solutions to many of them: data with multiple formats (points, polygons, time series, rasters, and temporal rasters), and tools to import, create, modify, analyze, visualize, store, and export the data and/or model simulation results. GRASS also offers multiple static and dynamic interactive visualization tools for use by HDFR. Figure 3 shows several example screenshots of the HDFR under GRASS.

The HDFR allows users to perform weather analysis workflows, involving the download, fusion, and assessment of different sources of information. As an example, precipitation information from May 16th, 2014 was downloaded from the NWS (radar) and NLDAS (land stations) servers. The downloaded precipitation was then fused using MKS, producing an estimate with better quality, and this fused precipitation was finally used to estimate the return period of the storm. The results and their associated processes are shown in Figure 4.

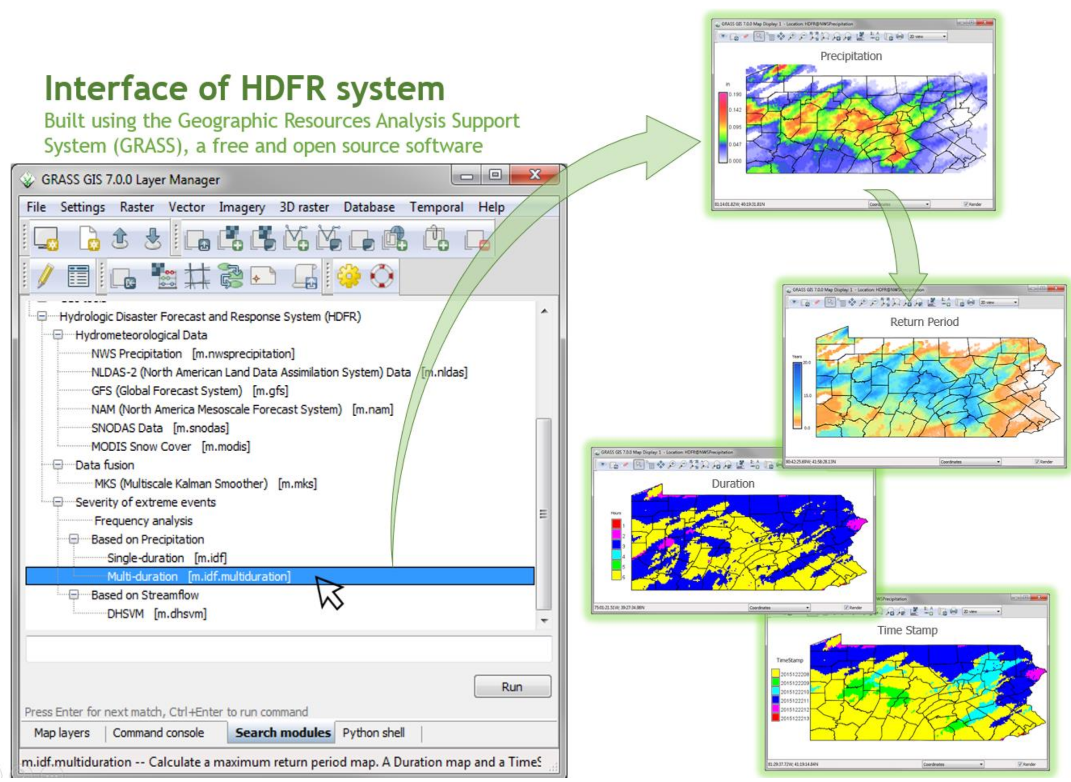

Figure 3. Screenshots of the HDFR under GRASS. Left: the HDFR module menu. Right: examples of data sets either downloaded using the data modules or computed using analysis modules.

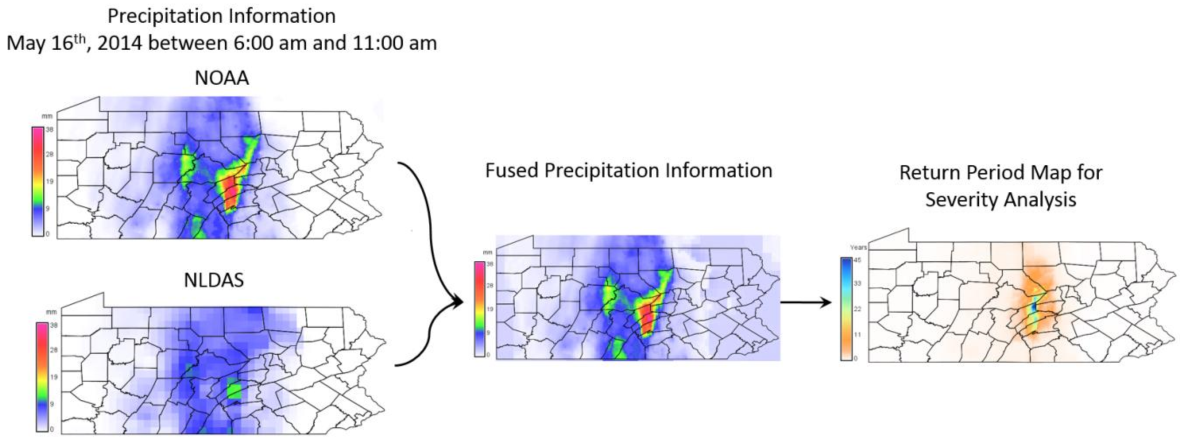

Figure 4. HDFR Workflow where precipitation information is downloaded, fused, and analyzed based on its return period. This precipitation event shows a return period of 45 years.

The HDFR is also able to perform such analyses using forecasted precipitation. This allows the user to perform risk assessments before the occurrence of precipitation events. An example of this is shown Figure 5, where forecast information from the North American Mesoscale (NAM) System is fused with forecast information from the Global Forecast System (GFS).

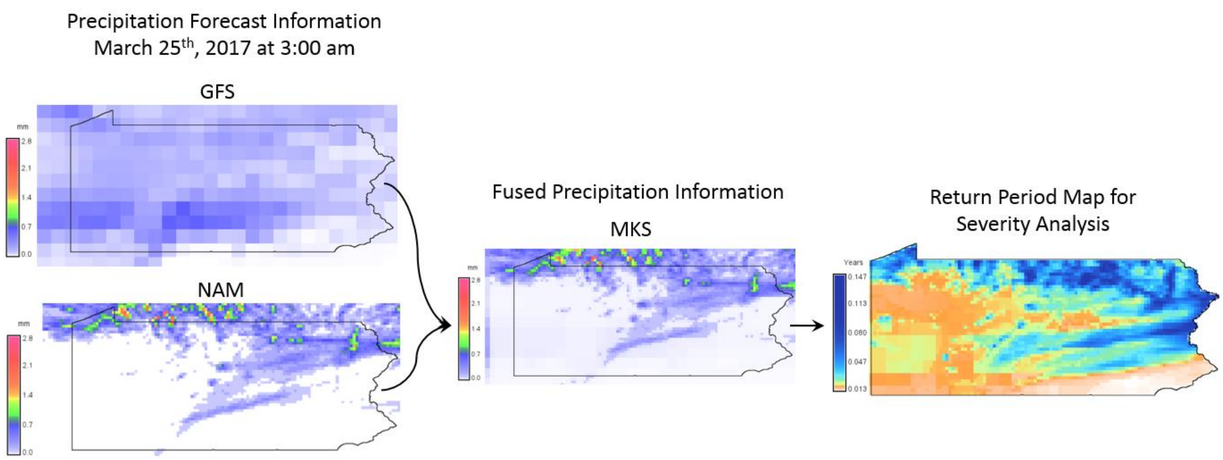

Figure 5. HDFR Workflow where forecast precipitation information is downloaded, fused and analyzed based on its return period.

Figure 6 shows one of the test models created for the DHSVM, the one for the small Indiantown Run watershed in southeastern Pennsylvania. The model was calibrated with the HDFR's algorithm using two years' worth of streamflow information. Figure 7 shows the comparison between streamflow time series for the uncalibrated and the calibrated models. It can be seen that the calibration process considerably reduces the discrepancies with the observations.

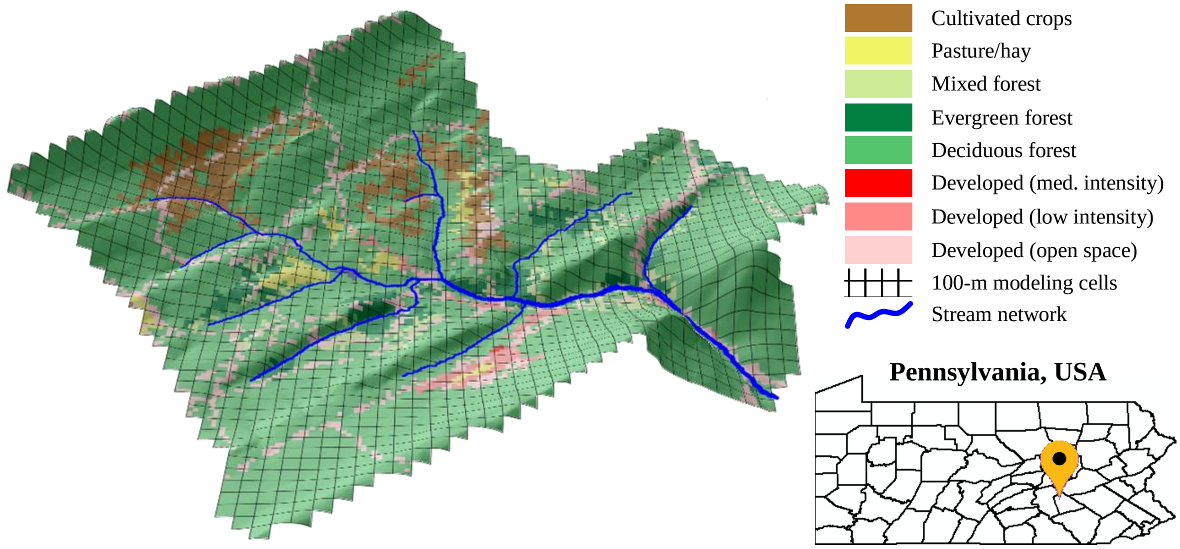

Figure 6. Illustration of the Indiantown Run watershed test model, showing the 100-m-size modeling cells (for a total of 1,472) and the assigned vegetation types.

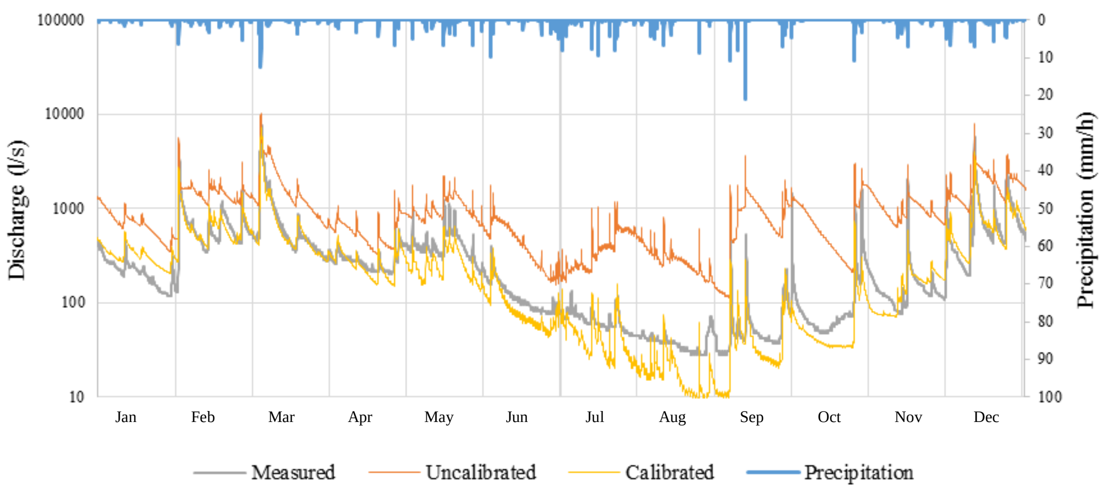

Figure 7. Streamflow time series comparison between calibrated and uncalibrated versions of the Indiantown Run model and the observations.

Figure 8 shows how OPTIMISTS was used to improve the streamflow forecasting skill of the Indiantown Run model in two different scenarios through improving the initial soil moisture states of the hydrological model.

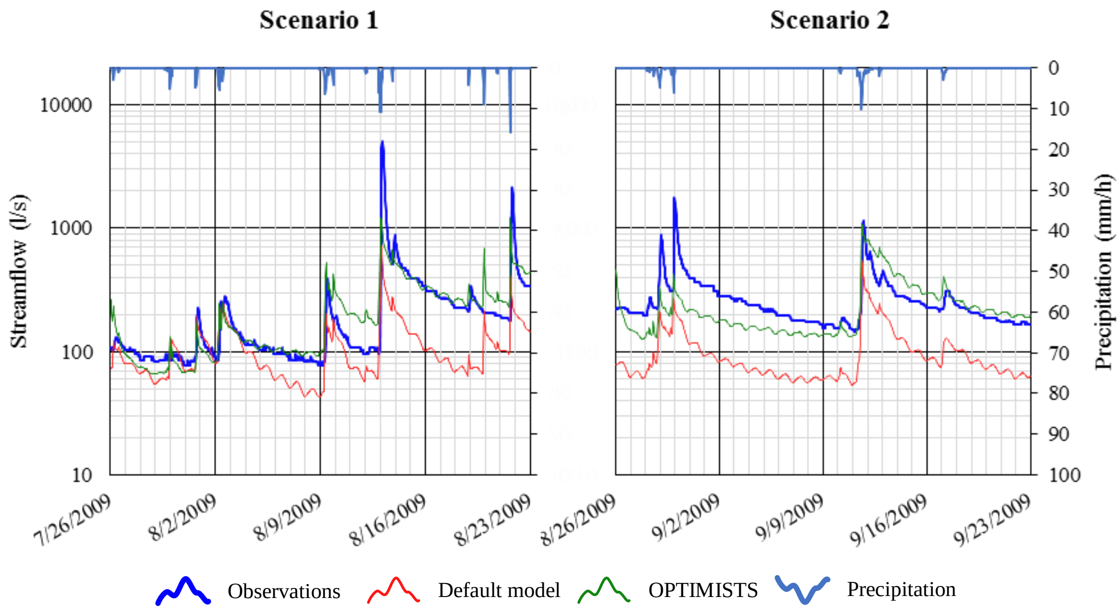

Figure 8. Streamflow time series comparison between the default Indiantown Run model and that produced after assimilation observed streamflow with OPTIMISTS. For each of the two scenarios, the first two weeks correspond to the assimilation period and the latter two correspond to the forecast period.

Additionally, a set of 15 gauged watersheds of different sizes that roughly correspond to PennDOT scour-critical bridges was selected to test the frequency analysis module for streamflow (using the USGS data), and for temperature extremes and snow cover (using the NLDAS data). The frequency analysis module was used to create IDF and QDF (streamflow duration frequency) curves for these test watersheds. For example, Figure 9 and Table 3 show the QDF curves for the Little Lehigh Creek watershed near Allentown, Pennsylvania.

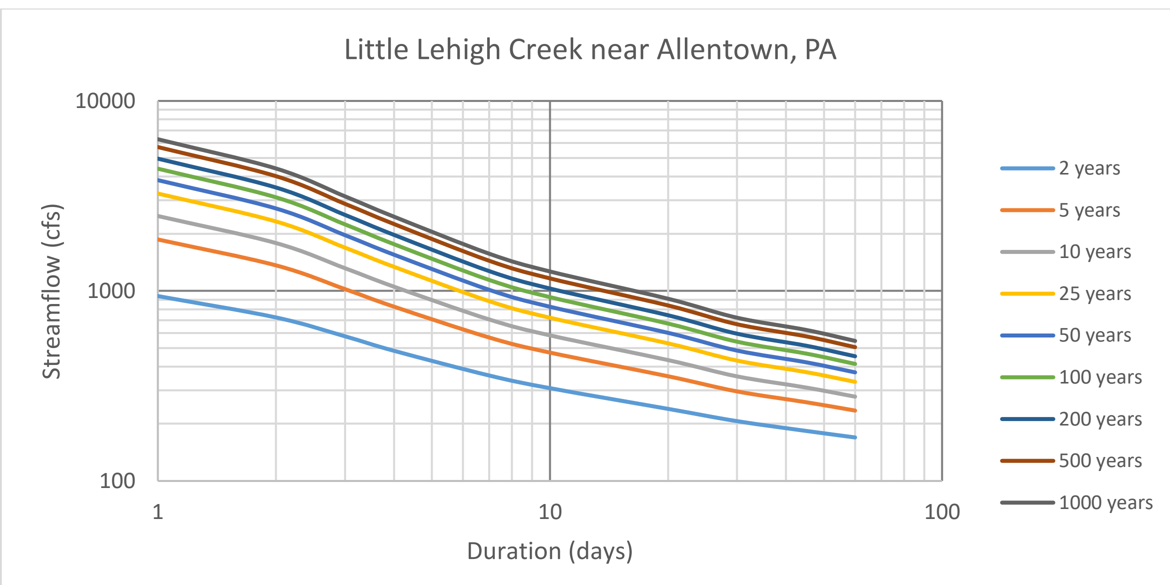

Figure 9. Streamflow-duration-frequency (QDF) curve for USGS site 01451500.

Table 3. Streamflow-duration-frequency (QDF) curve for Little Lehigh Creek near Allentown, PA (USGS site 01451500). Values are in cubic feet per second (cfs).

		Duration (days)									
		1	2	3	4	7	10	20	30	45	60
Return period (years)	2	940	725	577	484	359	308	240	206	183	169
	5	1866	1364	1023	827	569	474	356	296	260	235
	10	2480	1787	1318	1053	708	584	433	355	310	278
	25	3255	2321	1690	1340	883	723	530	430	374	332
	50	3829	2718	1966	1552	1014	826	602	486	421	373
	100	4400	3111	2241	1763	1143	928	674	541	468	413
	200	4969	3503	2514	1973	1272	1030	745	596	515	453
	500	5719	4021	2875	2251	1442	1164	839	668	577	506
	1000	6286	4411	3147	2460	1570	1266	910	723	623	546

Comparisons were performed between the different methods for estimating the return period of extreme precipitation events on some of the test watersheds. The different methods are:

1. Multi-duration precipitation severity module (point IDF curves) based on a similar method PennDOT uses at present
2. Watershed precipitation severity (uses newly-developed areal IDF curves for the entire corresponding watershed)
3. Streamflow severity regression module
4. Streamflow return period (estimated from severity curves that were computed from USGS streamflow observations)

Figure 10, Figure 11, and Figure 12 show comparisons between the obtained return period for different storms and different watersheds. Taking the streamflow return period as the ground truth (black time series), it can be seen that the use of point IDF curves (green time series) becomes less accurate for larger watersheds. Note that given the large size of the watershed, there is a delay in the streamflow peak compared to that in the precipitation signal. In addition, the use of precipitation information alone can significantly underestimate the return period of the streamflow during the snow melt season as shown in Figure 11. The streamflow regression module (orange time series in Figure 12), on the other hand, appears to accurately predict peak values even though its accuracy during recession periods is not very consistent.

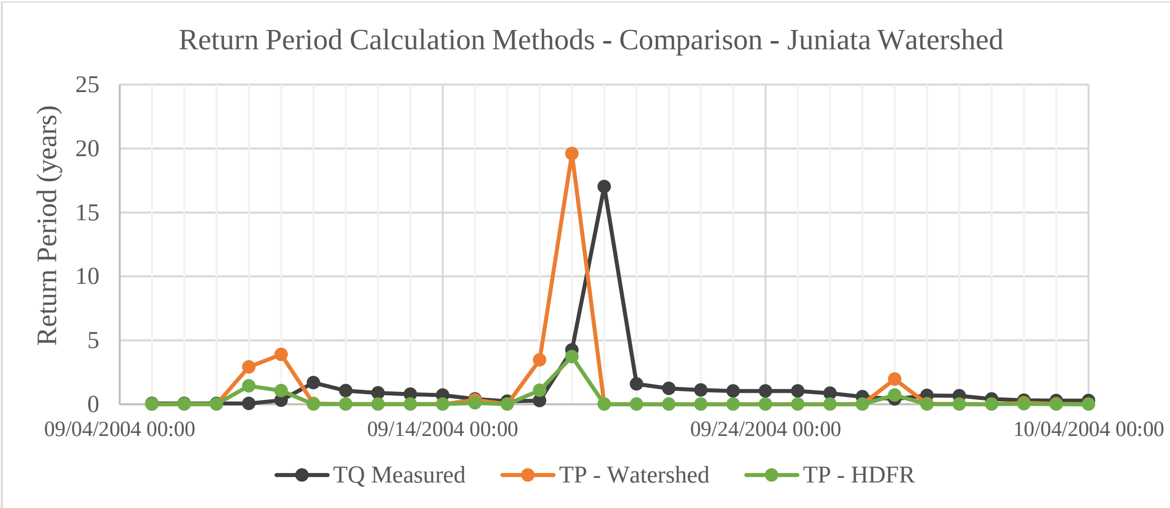

Figure 10. Comparison of return period (T) computed with different methods. Juniata Watershed. Outlet at USGS station 01567000.

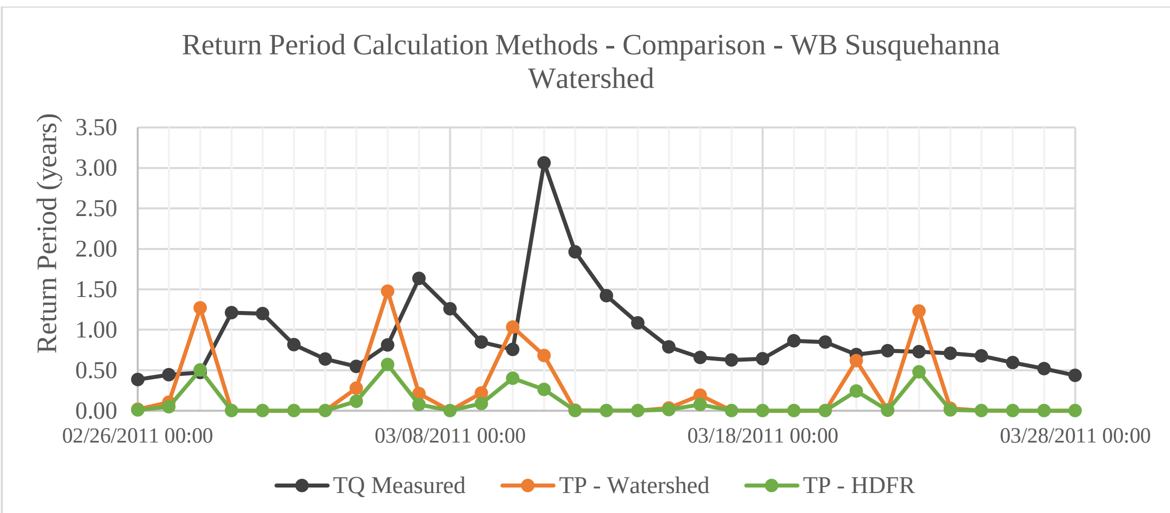

Figure 11. Comparison of return period (T) computed with different methods. West-Branch Susquehanna Watershed. Outlet at USGS station 01551500.

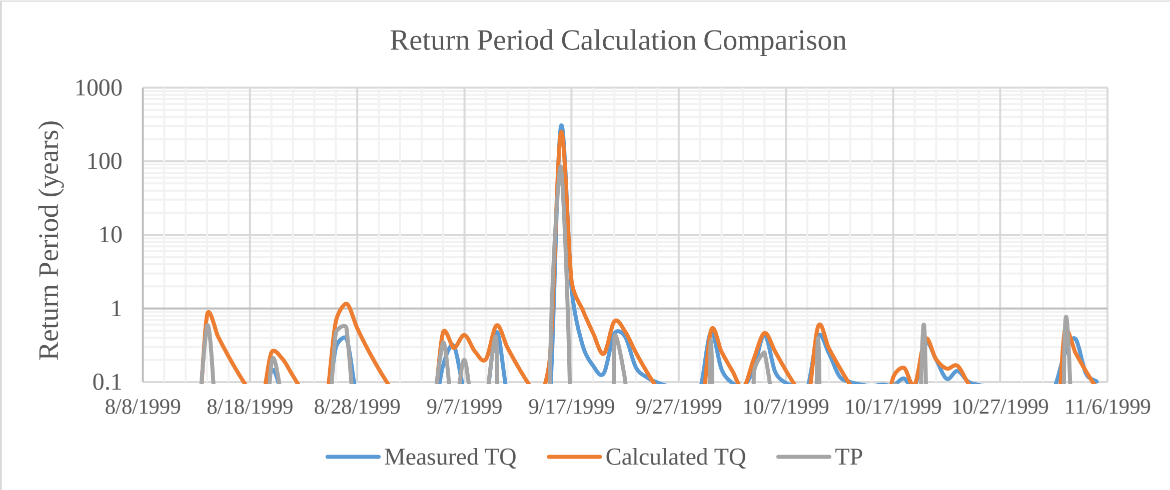

Figure 12. Comparison between real streamflow return period (TQ) and precipitation return period (TP).

Despite the concerns about the use of point IDF curves (e.g., the green curve in Figure 10) instead of watershed-based IDF curves (e.g., the orange curve in Figure 10), PennDOT decided that the former ones were to be used in the near future for two reasons: first, because developing IDF curves for every watershed would be a very labor-intensive endeavor and, second, because the bridges most vulnerable to scouring are usually small ones—which are generally associated with small watersheds.

The last objective of this project involves the loose integration of the HDFR with PennDOT's Intelligent Transport System (ITS) Traffic Management Centers (TMCs). This objective is well underway to completion with the Centers' Bridge Engineering team already using a delivered test version of the HDFR system. The version included several data modules and one of the precipitation severity analysis modules. PennDOT's tests allowed us to solve some difficulties regarding the installation of the system. PennDOT was also able to execute a simple workflow for the estimation of the severity of precipitation events.

Direct/tight interaction/integration between the Centers' software and the HDFR was deemed undesirable by PennDOT because 1) their system is still under development and test, with multiple components working currently in isolation; and 2) they did not want any interference between the two systems that could slow the development of each. Therefore, only manual interactions between the systems, and the offline test and use of the HDFR system is preferred by PennDOT in the evaluation process.

A second approved delivery consisted of a series of curves to estimate the severity of hydro-meteorological events for a set of 14 test watersheds. The test watersheds correspond to 14 USGS stream gauges, and they are roughly equivalent to those of vulnerable scour-critical bridges. That is, the watersheds of the streams that these bridges cross are similar to the watersheds of the streams where these gauges are located. An example of such curves is shown in Figure 13 and Table 4. Both deliverables can be found as attachments to this report.

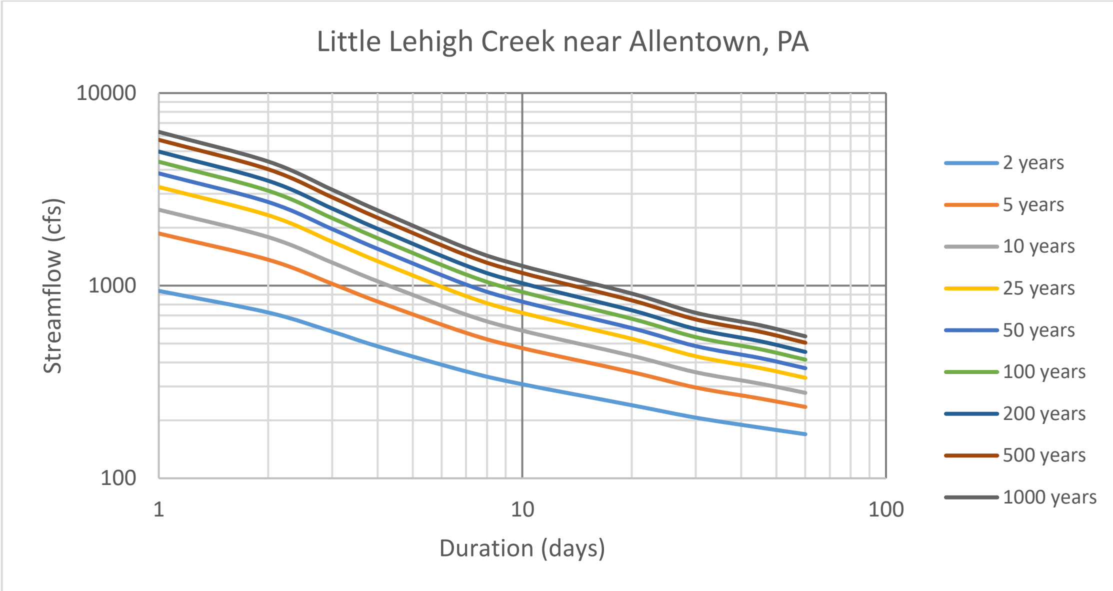

Figure 13. Streamflow-duration-frequency (QDF) curve for USGS site 01451500.

Table 4. Streamflow-duration-frequency (QDF) curve for Little Lehigh Creek near Allentown, PA (USGS site 01451500). Values are in cubic feet per second (cfs).

	Duration (days)									
Return period (years)	**1**	**2**	**3**	**4**	**7**	**10**	**20**	**30**	**45**	**60**
2	940	725	577	484	359	308	240	206	183	169
5	1866	1364	1023	827	569	474	356	296	260	235
10	2480	1787	1318	1053	708	584	433	355	310	278
25	3255	2321	1690	1340	883	723	530	430	374	332
50	3829	2718	1966	1552	1014	826	602	486	421	373
100	4400	3111	2241	1763	1143	928	674	541	468	413
200	4969	3503	2514	1973	1272	1030	745	596	515	453
500	5719	4021	2875	2251	1442	1164	839	668	577	506
1000	6286	4411	3147	2460	1570	1266	910	723	623	546

Chapter 4. Ongoing work

During the remainder of the collaboration with PennDOT under the scope of this project, which will last until April, 2018, the team will continue to work on the following activities. Stretch goals indicate those that they are desired but not required by the project's objectives. The activities are the following:

- Complete the development of the remaining data modules Table 1 to a level of completion of 3—a complete integration into the HDFR under GRASS GIS.
- (Stretch goal) Reduce the footprint of the MKS module by optimizing the original code to improve its efficiency. Such enhancement would ease the installation of the HDFR modules, reduce their size, and potentially increase the execution speed.
- While the DHSVM can be called from the corresponding HDFR module under GRASS, the simulation results cannot yet be accessed from the GIS interface. The module will be thus completed by allowing the user to specify which of the many simulation results to be retrieved and then allow those to be imported into the current GRASS project so that further analysis and visualization becomes possible.
- The two precipitation severity modules developed output the computed return period on a grid representation. They allow transferring these results to the bridges' watersheds through the areal averaging of these values to a provided set of polygons. However, GRASS does not support polygon layers with overlapping areas, which is the case for many small watersheds that are part of larger ones. The team plans to overcome this limitation either by modifying the way polygons are represented or by automating pre-processing methods to circumvent the overlapping constraint.
- Complete the integration of the frequency analysis and streamflow severity regression modules into the HDFR.
- (Stretch goal) Modify the streamflow severity regression module so that more adequate formulations are available for watersheds of different sizes.
- Deliver enhanced test versions of the HDFR to PennDOT and implement their feedback towards the final version of the HDFR which will be delivered to PennDOT in Spring 2018.
- Revise the submitted manuscript [5] and follow the process through publication.
- Write additional manuscripts related to the estimation of extreme values for precipitation and streamflow, dealing with parameter and initial state uncertainty simultaneously, high-dimensional data assimilation with OPTIMISTS, and a description of the entire HDFR system.

Appendix – References

[1] S. Wang, X. Liang, and Z. Nan, "How much improvement can precipitation data fusion achieve with a Multiscale Kalman Smoother-based framework?," *Water Resour. Res.*, vol. 47, no. 3, 2011.

[2] X. Liang, D. P. Lettenmaier, E. F. Wood, and S. J. Burges, "A simple hydrologically based model of land surface water and energy fluxes for general circulation models," *J. Geophys. Res.*, vol. 99, no. D7, p. 14415, 1994.

[3] Z. Wen, X. Liang, and S. Yang, "A new multiscale routing framework and its evaluation for land surface modeling applications," vol. 48, no. June, pp. 1–16, 2012.

[4] M. Wigmosta, B. Nijssen, and P. Storck, "The distributed hydrology soil vegetation model," *Math. Model. Small Watershed Hydrol. Appl.*, pp. 7–42, 2002.

[5] F. Hernández and X. Liang, "Hybridizing sequential and variational data assimilation for robust high-resolution hydrologic forecasting," *Hydrol. Earth Syst. Sci. Discuss.*, no. September, pp. 1–25, 2016.