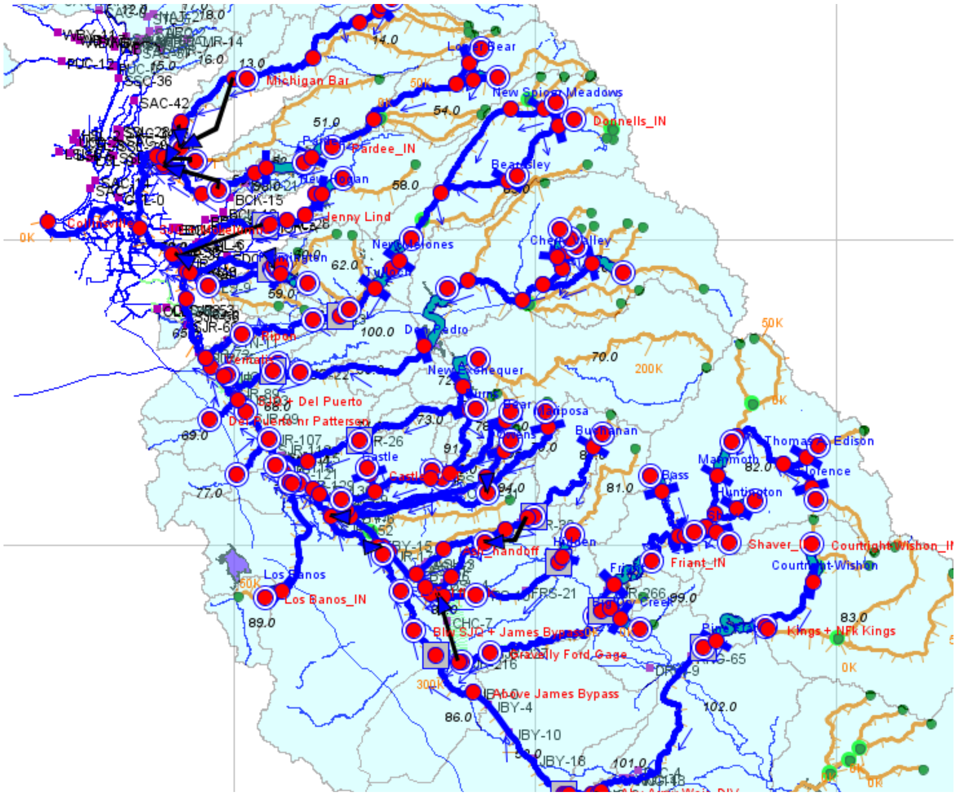


Successfully overcoming big data and big analysis challenges for the California Central Valley Hydrology Study

At left: San Joaquin River Basin reservoir operation model schematic. The model includes 35 reservoirs, 74 inflow locations, and 166 routing reaches.

By David Ford, PhD, PE, D.WRE and Nathan Pingel, PE, D.WRE

© 2016 David Ford Consulting Engineers, Inc.

Abstract

Demands on analysis to support flood risk management decision making have grown significantly. This growth is an outcome of (1) acknowledging the uncertainty about drivers of the decisions and wanting to account for that in assessing options; (2) appreciating the necessity of decision making in a system context, considering the direct and indirect impacts of choices; and (3) demanding more information on consequences and outcomes to guide our decision making. To meet the demands, we have taken advantage of advances in information technology in innovative and creative ways, including computing in the cloud, computing with more efficient algorithms, and managing and displaying big data sets in efficient ways.

In this white paper, we describe the strategy we employed to manage the "big data, big analysis" challenge of developing flood-frequency information and typical flood hydrographs for more than 200 locations in the Sacramento and San Joaquin River basins. We also describe strategies for applying the resulting "big information" for flood risk management alternatives analysis.

KEYWORDS

Big data; big analysis; flow-frequency analysis; hydrology; system analysis; California; Central Valley; flood management

What is "big data?"

The term "big data" refers to data or information sets of high volume, high velocity (frequently changing), and/or high variety or variability (Gartner 2012). Flood risk management decision making involves all 3 types of big data and information:

- High volume: Fine resolution of terrain, assets at risk, rainfall.
- High velocity: Constantly changing weather and water conditions.
- High variety: Flood risk components are diverse in type and vary with location. In addition, estimates of each are uncertain. Flood risk components include flood hazard, flood control system performance, exposure and vulnerability of people and property in the floodplain, and consequence.

Overcoming big data, big analysis challenges

With big data and information comes the need for big analysis. Traditional data management, analysis, and visualization strategies are inadequate for these. For some studies, simplified analyses may be used to produce low-resolution results, which can be appropriate depending on the application. However, in cases where improved accuracy and insight are important, big analysis is required. Use of the latest technology and data management strategies permits us to conduct these big analyses effectively and efficiently. This includes:

- Use of parallel computing, cloud computing, and other computing strategies.
- Development of custom applications to automate data query and storage, initiation of simulations, and results processing.
- Use of geographic information systems (GIS) to display a large volume of data and results in a meaningful way.
- Development of data management plans specific to the application. The data management plan identifies all analysis requirements and potential issues and provides strategies for meeting those requirements and resolving those issues. The planning process addresses the full scope of the analysis from start to finish and identifies opportunities for efficiency.

Central Valley Hydrology Study: A big data, big analysis challenge

The Central Valley Hydrology Study (CVHS) exemplifies a big data, big analysis challenge overcome with technology and data management strategies to produce high-resolution results to inform decision making for a large, complex watershed.

CVHS was a multi-year effort of the California Department of Water Resources (DWR) and the US Army Corps of Engineers (USACE), Sacramento District (SPK) to update flood hazard information for the Central Valley (DWR 2015). CVHS developed unregulated volume-frequency curves and regulated peak flow-frequency curves at over 200 locations. This information was then applied for risk analysis of several flood risk management alternatives for the Central Valley Flood Protection Plan (CVFPP) (DWR 2016).

CVHS was a big study with:

- Big basins. The hydrologic study area included 26,300 sq mi of the Sacramento River Basin and 16,700 sq mi of the San Joaquin River Basin as well as a portion of the Tulare Basin.
- Big flood control system. A complex network of reservoirs, levees, diversions, bypasses, and weirs control flows in the system. The 72 reservoirs modeled for the study include Shasta Lake with about 4.5 million ac-ft of storage, Oroville with 3.5 million ac-ft, Folsom Reservoir with 1 million ac-ft, and New Bullards Bar with 1 million ac-ft. About 1,600 miles of California’s State Plan of Flood Control (SPFC) levees reduce flood risk from the Sacramento and San Joaquin rivers and their tributaries (DWR 2012).
- Big flows. The watershed is subject to abundant rain and drains snowmelt from the Sierra Nevada mountain range. Floods are caused by intense rainfall, normal snowmelt, and rain on snow. During the 1997 event, peak inflow to Shasta Lake was 236,734 cfs, Lake Oroville 302,013 cfs, and Folsom Reservoir 252,538 cfs (CNRFC undated).
- Big flood risk. The population at risk in the Central Valley is about 1 million, and assets at risk are about $70 billion (DWR 2012).
- Big data. The study involved processing of stream gage data throughout the basins. For some gages, more than 100 years of data was processed. In addition, for hydraulic modeling, DWR collected 7,800 sq mi of Light Detection and Ranging (LiDAR) data and aerial photos for 9,000 sq mi, and conducted field surveys for 3,000 cross sections included in the hydraulic model and bathymetric surveys for 2,500 cross sections.
- Big analysis tools. To complete the hydrologic and risk analysis study, a full suite of applications was required: HEC-HMS for precipitation-runoff modeling, HEC-ResSim for reservoir operation simulation, HEC-RAS for hydraulic routing, HEC-FDA for risk analysis, HEC-SSP for statistical analysis, HEC-DSS to serve as a common database system, and HEC-DSSVue to display data and results. The system-wide HEC-ResSim models included 72 reservoirs and 404 control points. The system-

wide HEC-RAS channel models covered approximately 1,650 miles of streams and had more than 9,000 cross sections. The FLO-2D overland hydraulic routing models covered nearly 6,000 sq mi of the floodplain.

With every facet of the study being big, CVHS faced 5 major challenges:

1. Storing and transferring big data.
2. Processing big data and results.
3. Simulating a big number of events and scenarios.
4. Displaying a big set of results.
5. Sharing a big amount of information to a big number of stakeholders.

How CVHS overcame the big data, big analysis challenge

With detailed planning and use of technology, CVHS overcame each of the 5 challenges.

Challenge 1: Storing and transferring big data

CVHS required data sharing between SPK and Ford Engineers on a daily basis over the course of the study. File sizes were often in gigabytes. Sending these files using an ftp site would be inefficient due to long uploading and downloading times, and picking up and dropping off portable hard drives was not practical.

To overcome this challenge, we set up a data repository system using Subversion and TortoiseSVN, which is an open source client-server system. Both SPK and Ford Engineers could add, view, edit, and delete files rapidly on the server and use a shared file structure, as if they were working on a common local PC. The data repository also allowed for version tracking, working as a library system where a model was checked out, updated, returned to the repository, and checked out and replaced again, as illustrated in Figure 1.

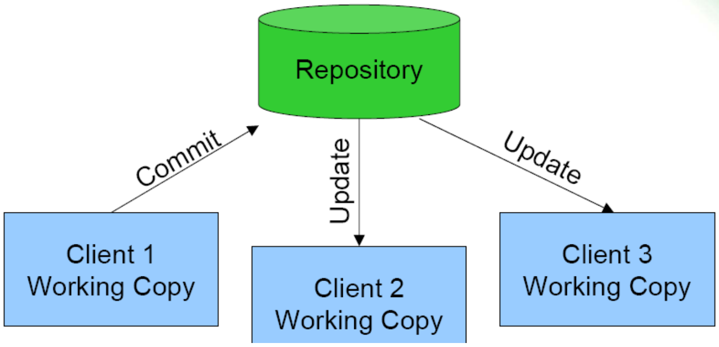


Figure 1. We set up a data repository with a versioning system to transfer and manage the high volume of data, models, and results for the study.

Challenge 2: Processing big data and results

CVHS required processing of thousands of daily or hourly stream gage readings for the period of record at locations throughout the Sacramento and San Joaquin River basins.

These boundary condition flows were routed through an HEC-RAS hydraulic system model reflecting the unregulated condition, absent reservoir and floodplain storage, for both basins. The results of this simulation were analyzed statistically, yielding unregulated volume-frequency curves for multiple durations at each of the 200 analysis locations.

To develop regulated flow-frequency curves needed for planning, the unregulated flows and scaled versions of those are routed through an HEC-ResSim reservoir operation system model and an HEC-RAS system model reflecting the regulated condition (with reservoir and floodplain storage) for both basins. The results are referenced back to the unregulated volume-frequency curves to fit the regulated flow-frequency curves for the 200 analysis locations.

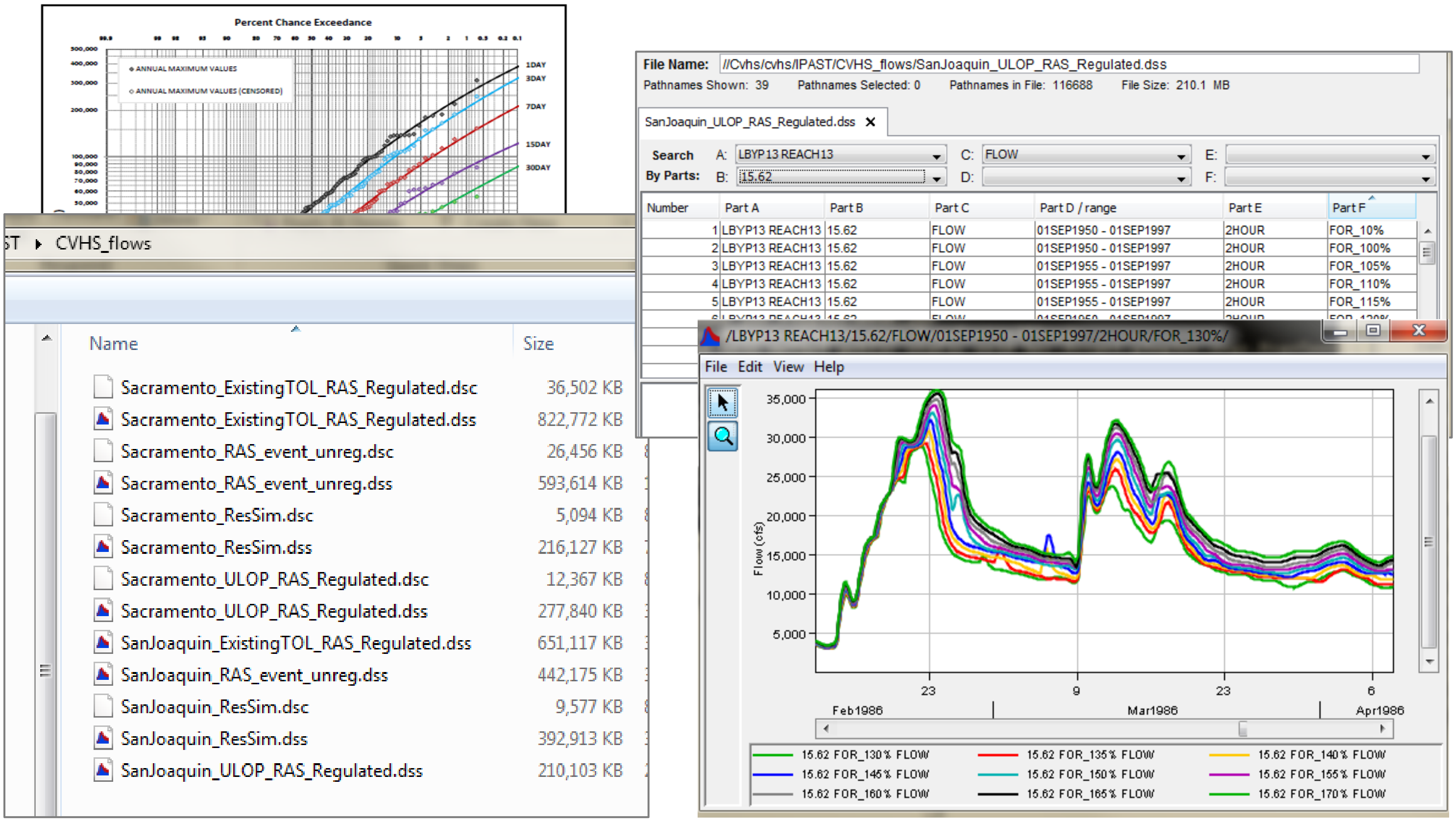
This iterative process done manually would take many weeks, and CVHS would not meet its demanding schedule. Thus, Ford Engineers developed a custom application to facilitate these tasks.

The application, the Information Processing and Synthesis Tool (IPAST) (Ford Engineers 2015), automates data extraction, data processing, results extraction, and results processing. Specifically IPAST achieves the following:

- Integrates seamlessly with the simulation models through HEC-DSS, USACE's data management system.
- Manages the big datasets from the hundreds of simulations and hundreds of locations, extracting unregulated flow and volume annual maximums and fitting unregulated flow-frequency curves using state-of-the-art procedures developed by USACE and the US Geological Survey (USGS).
- Extracts unregulated and corresponding regulated flows and creates in a consistent manner an unregulated-to-regulated flow transform and a regulated flow-to-stage transform for hundreds of sites in the basins.
- Computes various statistics from historical volume, flow, and stage records and from extracted simulation results.
- Allows the user to control the sequence and method of analysis and to view all results in an easy-to-understand graphical user interface (GUI), shown in Figure 2, without using the complex simulation applications directly.

The product of IPAST use is thousands of data values and results digested into concise, meaningful information to be reviewed and deemed final by an engineer. In addition, by automating the process, human error is eliminated.

We developed IPAST in the .NET platform, with an underlying SQL Server Compact database. IPAST has been approved through the USACE Hydrology, Hydraulics, and Coastal Community of Practice software validation process for projects within SPK.



Name	Size
Sacramento_ExistingTOL_RAS_Regulated.dsc	36,502 KB
Sacramento_ExistingTOL_RAS_Regulated.dss	822,772 KB
Sacramento_RAS_event_unreg.dsc	26,456 KB
Sacramento_RAS_event_unreg.dss	593,614 KB
Sacramento_ResSim.dsc	5,094 KB
Sacramento_ResSim.dss	216,127 KB
Sacramento_ULOP_RAS_Regulated.dsc	12,367 KB
Sacramento_ULOP_RAS_Regulated.dss	277,840 KB
SanJoaquin_ExistingTOL_RAS_Regulated.dss	651,117 KB
SanJoaquin_RAS_event_unreg.dss	442,175 KB
SanJoaquin_ResSim.dsc	9,577 KB
SanJoaquin_ResSim.dss	392,913 KB
SanJoaquin_ULOP_RAS_Regulated.dss	210,103 KB

File Name: //Cvhs/cvhs/IPAST/CVHS_flows/SanJoaquin_ULOP_RAS_Regulated.dss

Pathnames Shown: 39 Pathnames Selected: 0 Pathnames in File: 116688 File Size: 210.1 MB

SanJoaquin_ULOP_RAS_Regulated.dss

Search By Parts: A: LBYP13 REACH13 B: 15.62 C: FLOW D: E: F:

Number	Part A	Part B	Part C	Part D / range	Part E	Part F
1	LBYP13 REACH13	15.62	FLOW	01SEP1950 - 01SEP1997	2HOUR	FOR_10%
2	LBYP13 REACH13	15.62	FLOW	01SEP1950 - 01SEP1997	2HOUR	FOR_100%
3	LBYP13 REACH13	15.62	FLOW	01SEP1955 - 01SEP1997	2HOUR	FOR_105%
4	LBYP13 REACH13	15.62	FLOW	01SEP1955 - 01SEP1997	2HOUR	FOR_110%
5	LBYP13 REACH13	15.62	FLOW	01SEP1955 - 01SEP1997	2HOUR	FOR_115%

/LBYP13 REACH13/15.62/FLOW/01SEP1950 - 01SEP1997/2HOUR/FOR_130%/

Figure 2. We developed a custom application, the Information Processing and Synthesis Tool (IPAST), to process the high volume of data and results.

Challenge 3: Simulating a big number of events and scenarios

As described in the previous section, historical event datasets and scaled versions of those datasets were routed through Sacramento and San Joaquin River basin HEC-ResSim and HEC-RAS system models configured to represent the regulated condition. The results were used to develop regulated flow-frequency curves.

CVHS required analysis of 4 historical events, each of which was scaled with 37 multipliers. This yielded 148 alternatives that needed to be simulated with each model. With 1 HEC-ResSim model and 1 HEC-RAS model for 2 basins (4 models), 592 total simulations were required. Each simulation took between 45 and 180 minutes, depending on the model and alternative.

CVHS produced flood hazard information for the CVFPP. When applying the CVHS procedure for the CVFPP, analysis of an alternative plan for managing flood water in the system could require over 4 weeks of continuous processing if the simulations were run sequentially on a single processor. Time for configuration of the simulations would increase this time. The result would have been an unacceptable computing burden, considering the need to analyze many alternative plans.

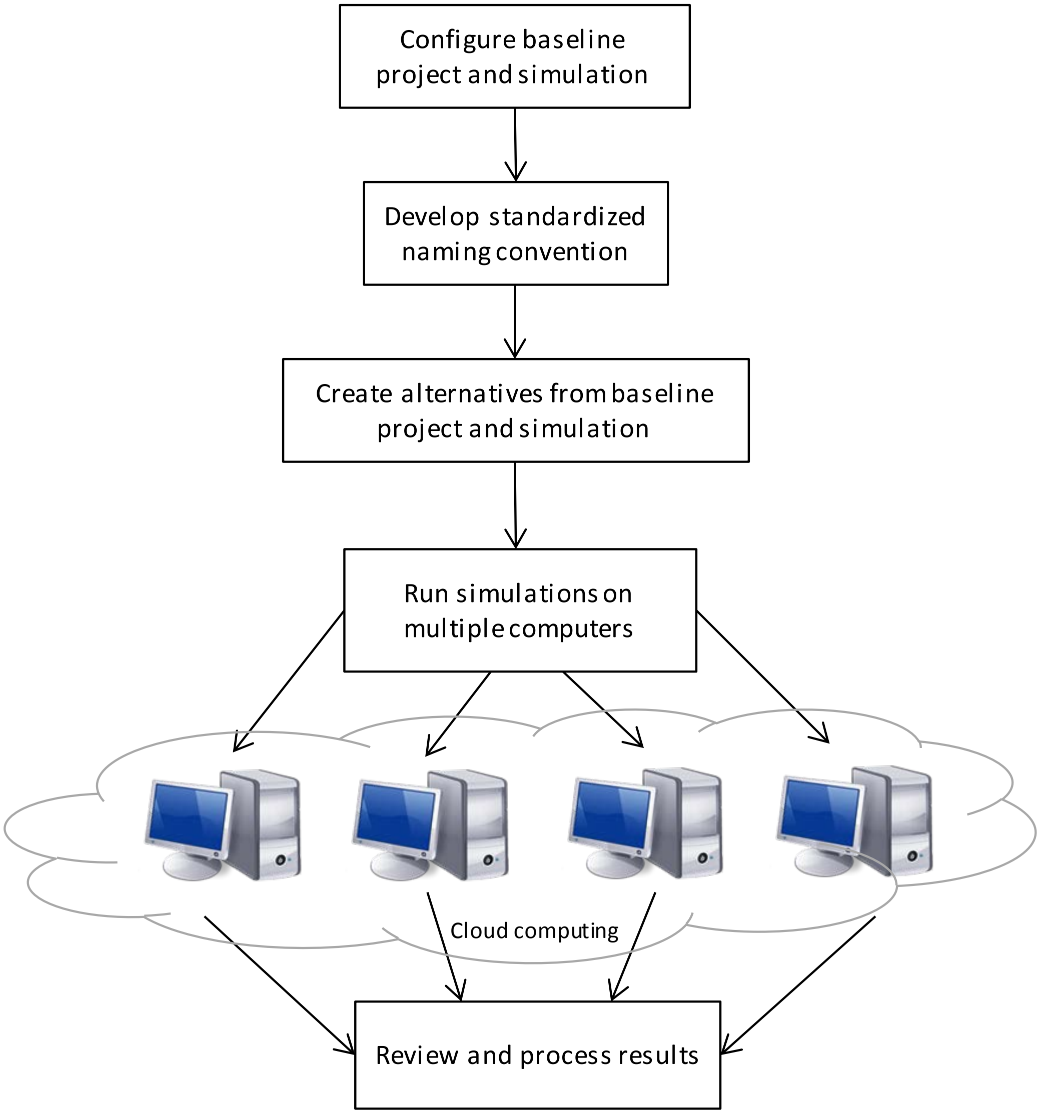

Configure baseline project and simulation
Develop standardized naming convention
Create alternatives from baseline project and simulation
Run simulations on multiple computers
Cloud computing
Review and process results

Figure 3. We automated configuration and initiation of simulations over multiple computers, thus reducing the total processing time from more than 4 weeks to less than 1 week.

To permit the processing needed in a reasonable time, we developed a computational scheme that used multiple computers for the simulations, as illustrated in Figure 3. The required simulations were distributed to computers and executed simultaneously. Specialized applications managed the execution, tracked progress, started new simulations, and managed the transfer of results from 1 model to another. All of this minimized human intervention and permitted parallel processing, thus reducing the total processing time from more than 4 weeks to less than 1 week, while reducing the chance for errors when dealing with the many and large model output files.

Challenge 4: Displaying a big set of results

The Central Valley Floodplain Evaluation and Delineation (CVFED) Program and CVFPP used the flood hazard information developed by CVHS to inform floodplain mapping and risk analysis throughout the basins (DWR undated). The large set of results needed to be presented in a concise, meaningful way for a wide range of stakeholders.

To accomplish this, DWR used GIS to display the information in a spatially referenced way, as shown in Figure 4. Using GIS, the user can pan through the study area and see, for example, depth and velocity results from a levee breach scenario at any given location. In addition, GIS and the high volume of topographic data were used to develop the analysis models that represent the watershed at a fine resolution.

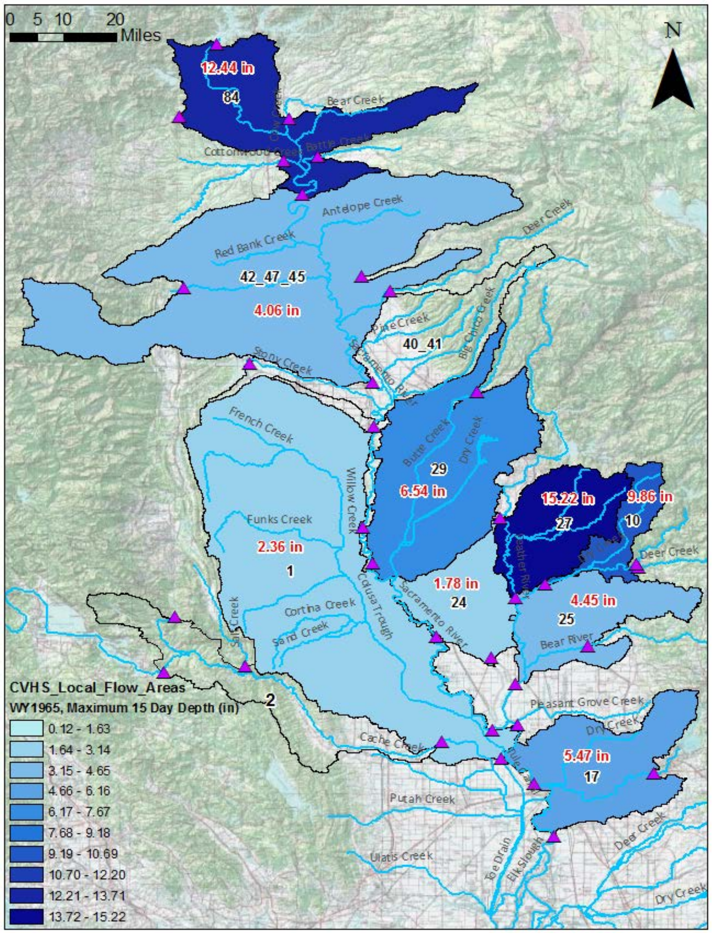
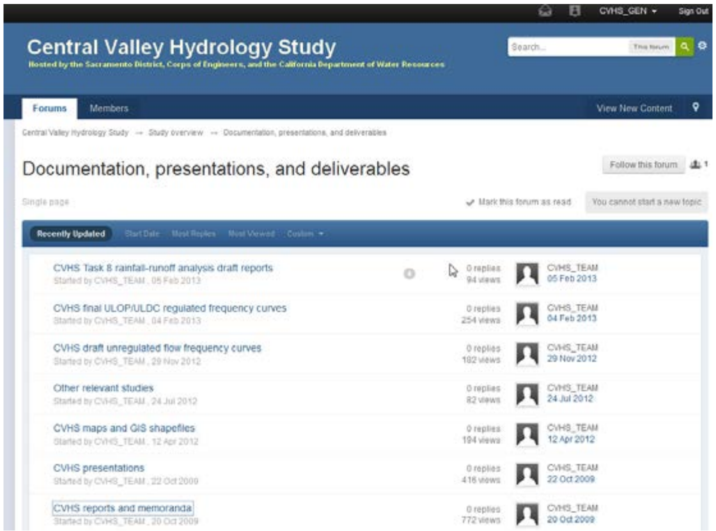


Figure 4. GIS was used to display the basin-wide results in a concise, meaningful way.

Challenge 5: Sharing a big amount of information to a big number of stakeholders

Finally, because of the study's large geographic scope and foundational importance to future studies, information dissemination was required throughout the study for many stakeholders.

Rather than sending hundreds of e-mails, we developed a Web forum, www.cvhydrology.org, shown in Figure 5. Here, we posted the numerous presentations, reports, memos, and deliverables of the study. Stakeholders easily downloaded the information, received automatic notifications when new information was available, and posted questions for the CVHS team to respond to. The Web forum provided for efficient communication and transparency throughout the multi-year study. The reports and memoranda section alone has been viewed 1,500 times.



Central Valley Hydrology Study

Hosted by the Sacramento District, Corps of Engineers, and the California Department of Water Resources

Central Valley Hydrology Study → Study overview → Documentation, presentations, and deliverables

Documentation, presentations, and deliverables

CVHS Task 8 rainfall-runoff analysis draft reports Started by CVHS_TEAM , 05 Feb 2013	0 replies 94 views	CVHS_TEAM 05 Feb 2013
CVHS final ULOP/ULDC regulated frequency curves Started by CVHS_TEAM , 04 Feb 2013	0 replies 254 views	CVHS_TEAM 04 Feb 2013
CVHS draft unregulated flow frequency curves Started by CVHS_TEAM , 29 Nov 2012	0 replies 182 views	CVHS_TEAM 29 Nov 2012
Other relevant studies Started by CVHS_TEAM , 24 Jul 2012	0 replies 82 views	CVHS_TEAM 24 Jul 2012
CVHS maps and GIS shapefiles Started by CVHS_TEAM , 12 Apr 2012	0 replies 194 views	CVHS_TEAM 12 Apr 2012
CVHS presentations Started by CVHS_TEAM , 22 Oct 2009	0 replies 416 views	CVHS_TEAM 22 Oct 2009
CVHS reports and memoranda Started by CVHS_TEAM , 20 Oct 2009	0 replies 772 views	CVHS_TEAM 20 Oct 2009

Figure 5. We developed a Web forum for sharing information and corresponding with a large number of stakeholders.

Keys to success

Keys to success in managing the CVHS big data, big analysis, which can be applied to other studies, include:

- Develop in a collaborative manner detailed work plans. CVHS was successful because challenges and methods to overcome them were identified early on in the data management plan, and each team member had a clear understanding of how to implement the data management plan.
- Set and follow strict data management standards and versioning protocols. Voluminous data, models, and results can create chaos if a uniform naming and cataloging system is not applied at the onset. CVHS was completed efficiently because of use of naming conventions and the versioning features in the Subversion-TortoiseSVN system.
- Maximize automation to minimize human error. Human error is inevitable with complex analyses and can result in costly delays and quality issues. CVHS invested in developing custom applications to automate iterative processes, which paid off in smooth execution of the analysis in a short time frame.
- Leverage GIS and graphics applications to view results. Decision makers require concise meaningful information and often the most effective way to present this information is visually. As CVFED and the CVFPP demonstrated, GIS is a way to present a high volume of results spatially so that the user can identify a specific area and see detailed results. The user can also see how those results fit in with results in other areas.
- Avoid letting perceived computational limitations "limit" the accuracy of study results. As demonstrated by CVHS, a wide range of technology can be used to overcome big data, big analysis challenges. Rather than change the desired level of detail for a study, tools are readily available or can be developed to facilitate the study.

References

- California Department of Water Resources (DWR) (2012). *2012 Central Valley Flood Protection Plan*. June.
- DWR (2015). *Central Valley Hydrology Study*. Final report prepared by SPK and Ford Engineers. Nov. 29. <www.cvhydrology.org>.
- DWR (2016). *Central Valley Flood Protection Plan 2017 Update*. Draft. December.
- DWR (undated). "Central Valley Floodplain Evaluation and Delineation Program." Web forum. <www.cvfed.org>.
- California Nevada River Forecast Center (CNRFC) (undated). "Heavy Precipitation Event: Southwest Oregon, Northern California, and Western Nevada, December 26, 1996 – January 3, 1997." <http://www.cnrfc.noaa.gov/storm_summaries/jan1997storms.php>.
- David Ford Consulting Engineers (Ford Engineers) (2015). "Information Processing and Synthesis Tool." White paper. <http://ford-consulting.com/wp-content/uploads/2016/05/IPAST_whitepaper_20151215.pdf>.
- Gartner (2012). "The Importance of 'Big Data': A Definition."<https://www.gartner.com/doc/2057415/importance-big-data-definition>.

David Ford Consulting Engineers

2015 J Street, Suite 200
Sacramento, CA 95811
Ph. 916-447-8779
support@ford-consulting.com