Journal of Hydrology 519 (2014) 2978–3000

Contents lists available at ScienceDirect

Journal of Hydrology

journal homepage: www.elsevier.com/locate/jhydrol

Evaluation of real-time hydrometeorological ensemble prediction on hydrologic scales in Northern California

Konstantine P. Georgakakos *, Nicholas E. Graham, Theresa M. Modrick, Michael J. Murphy Jr. [1], Eylon Shamir, Cristopher R. Spencer, Jason A. Sperfslage

Hydrologic Research Center, 12555 High Bluff Drive, Suite 255, San Diego, CA 92130, USA

ARTICLE INFO

Article history:
Available online 24 May 2014

Keywords:
Ensemble forecasting
Water resources management
Hydrometeorology
Forecast validation
Northern California

SUMMARY

The paper presents an evaluation of real time ensemble forecasts produced during 2010–2012 by the demonstration project INFORM (Integrated Forecast and Reservoir Management) in Northern California. In addition, the innovative elements of the forecast component of the INFORM project are highlighted. The forecast component is designed to dynamically downscale operational multi-lead ensemble forecasts from the Global Ensemble Forecast System (GEFS) and the Climate Forecast system (CFS) of the National Centers of Environmental Prediction (NCEP), and to use adaptations of the operational hydrologic models of the US National Weather Service California Nevada River Forecast Center to provide ensemble reservoir inflow forecasts in real time. A full-physics 10-km resolution (10 km on the side) mesoscale model was implemented for the ensemble prediction of surface precipitation and temperature over the domain of Northern California with lead times out to 16 days with 6-hourly temporal resolution. An intermediate complexity regional model with a 10 km resolution was implemented to downscale the NCEP CFS ensemble forecasts for lead times out to 41.5 days. Methodologies for precipitation and temperature model forecast adjustment to comply with the corresponding observations were formulated and tested as regards their effectiveness for improving the ensemble predictions of these two variables and also for improving reservoir inflow forecasts. The evaluation is done using the real time databases of INFORM and concerns the snow accumulation and melt seasons. Performance is measured by metrics that range from those that use forecast means to those that use the entire forecast ensemble.

The results show very good skill in forecasting precipitation and temperature over the subcatchments of the INFORM domain out to a week in advance for all basins, models and seasons. For temperature, in some cases, non-negligible skill has been obtained out to four weeks for the melt season. Reservoir inflow forecasts exhibit also good skill for the shorter lead-times out to a week or so, and provide a good quantitative basis in support of reservoir management decisions pertaining to objectives with a short term horizon (e.g., flood control and energy production). For the northernmost basin of Trinity reservoir inflow forecasts exhibit good skill for lead times longer than 3 weeks in the snow melt season. Bias correction of the ensemble precipitation and temperature forecasts with fixed bias factors over the range of lead times improves forecast performance for almost all leads for precipitation and temperature and for the shorter lead times for reservoir inflow. The results constitute a first look at the performance of operational coupled hydrometeorological ensemble forecasts in support of reservoir management.

1. Introduction

The Integrated Forecast and Reservoir Management (INFORM) demonstration project was conceived to demonstrate increased water-use efficiency in Northern California water resources operations through the innovative application of climate, hydrologic and decision science (Georgakakos et al., 2005, 2000; Carpenter and Georgakakos, 2001; Yao and Georgakakos, 2001). The particular objectives of INFORM are to (a) implement an integrated forecast-management system for the primary Northern California reservoirs, both for individual reservoirs as well as system-wide; (b) demonstrate the utility of climate, weather and hydrologic forecasts through near-real-time tests of the integrated system with actual data; and (c) align the forecast component of INFORM to

* Corresponding author. Also, at: Scripps Institution of Oceanography, UCSD, La Jolla, CA 92093, USA. Tel.: +1 (858) 461 4560.
E-mail address: KGeorgakakos@hrc-lab.org (K.P. Georgakakos).
[1] Current address: Monash Weather and Climate Group, School of Mathematical Sciences, Monash University, Melbourne, Victoria 3800, Australia.

http://dx.doi.org/10.1016/j.jhydrol.2014.05.032
0022-1694/

existing operational models and practices in the region to facilitate an eventual smooth transition to operations.

The Northern California river and reservoir system serves many vital water uses, including providing two-thirds of the state's drinking water, irrigating 7 million acres of the world's most productive farmland, and being home to hundreds of species of fish, birds, and plants. In addition, the system protects Sacramento and other major cities from flood disasters and contributes significantly to the production of hydroelectric energy. The Sacramento-San Joaquin Delta provides a unique environment and is California's most important fishery habitat. Water from the Delta is pumped and transported through canals and aqueducts south and west serving the water needs of many more urban, agricultural, and industrial users.

Fig. 1 shows the drainage basins of the region of interest in Northern California delineated by the U.S. National Weather Service (NWS) California Nevada River Forecast Center (CNRFC). The drainage basins are on the American, Yuba, Feather, Sacramento, and Trinity Rivers and their tributaries. The Folsom, Oroville, Shasta, New Bullards Bar and Englebright reservoirs on the Sacramento River tributaries are included in the INFORM system, together with the Trinity Reservoir (Clair Engle Lake) on the Trinity River. Forecasting of the precipitation and temperature in these drainage basins and of the resulting inflow into these reservoirs are part of the INFORM demonstration project activities.

INFORM meteorological-component forecast models use as input the operational ensemble forecasts of the National Centers for Environmental Prediction (NCEP) of the National Oceanic and Atmospheric Administration (NOAA). The INFORM models downscale these forecasts for the watersheds of the region of interest in real time to produce high resolution ensemble precipitation and surface air temperature forecasts and ensemble forecasts of reservoir inflows. The innovative aspects of the formulation aim to provide the ability to generate continuous dynamically-downscaled forecasts with high temporal and spatial resolution with lead times from 6 to 41.5 days.

In the present paper we summarize the formulations and procedures associated with the real-time ensemble predictions of basin-scale precipitation and temperature as well as of the ensuing reservoir inflows, and focus on the evaluation of their performance using observed data from the available real time databases. This evaluation of performance intends to illuminate the real time behavior of the forecast system that includes changes in the NCEP operational model output during the period of evaluation and a few missing forecasts due to real time connectivity problems. However, it is this type of evaluation that is useful for real time reservoir management as these are recurring problems with real time forecast systems.

Forecast lead times of interest in this work are from 6 h to 41.5 days and are pertinent to reservoir operations management. INFORM also produces longer lead-time forecasts (once a month out to 9 months with daily resolution) pertinent to reservoir operations planning but these are not evaluated herein (see relevant evaluations in Carpenter and Georgakakos (2001), and Yao and Georgakakos (2001), as well as in HRC-GWRI (2007, 2013)).

More complete descriptions of the activities, formulations, main findings and recommendations of INFORM are presented in HRC-GWRI (2007, 2013). Recent application of the INFORM formulations in climate change studies is in Georgakakos et al. (2012b,a). The present work complements the literature of operational forecast system evaluation that pertains to coupled hydrometeorological models, high resolution gridded ensemble precipitation and temperature forecasts in mountainous terrain with seasonal snow, and ensemble reservoir inflow forecasts used for improved multi-objective reservoir management worldwide (e.g., Collischonn et al., 2007; Olsson and Lindstrom, 2008; Vannitsem, 2008; McCollor and Stull, 2008; Renner et al., 2009; Cloke and Pappenberger, 2009; Janowiak et al., 2010; Achleitner et al.,

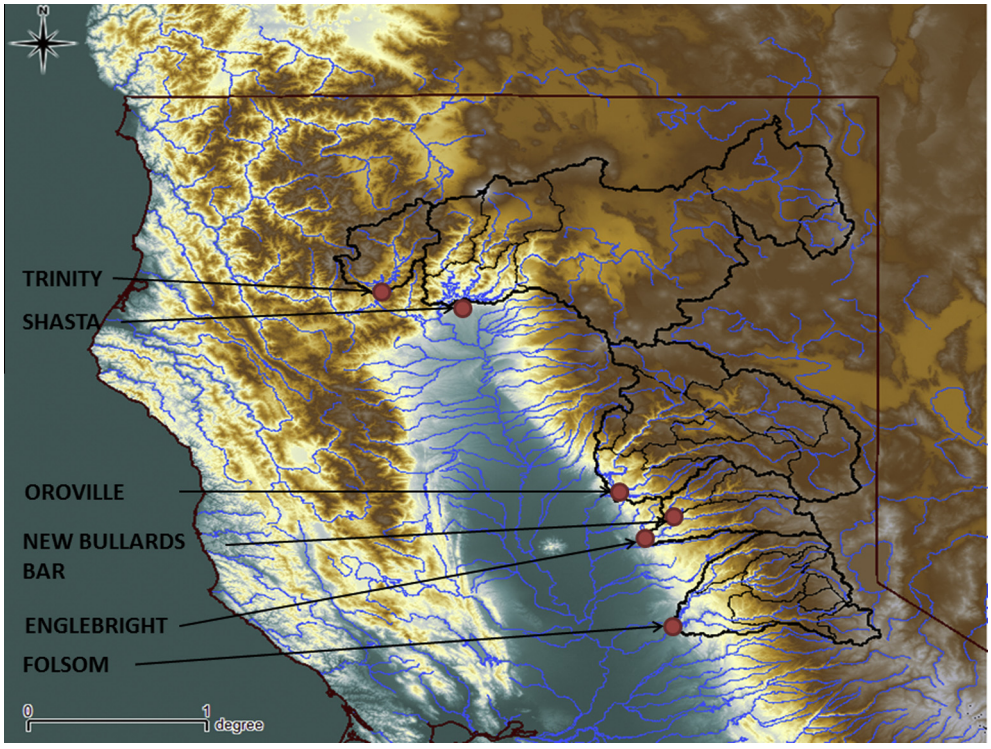

Fig. 1. Northern California hydrologic basins for the INFORM demonstration project. Watersheds draining into reservoirs are indicated by heavy black lines, subcatchments by thin black lines and main rivers by blue lines. The major reservoir locations and names are also indicated. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

2012; Boucher et al., 2012). The present paper contributes new information regarding the maximum lead times of reliable forecasts as it examines forecast lead times from the same operational forecast system in the range 1–41.5 days, in a series of medium to large reservoir drainage areas (from ~2000 km^2 to ~20,000 km^2) that span a 3-degree latitude range and an elevation range of more than 2500 m. The unique contribution of the present work rests on the novel approach followed to dynamically downscale operational large scale ensemble forecasts across the forecast horizons from 6 h to 41.5 days. The methodology is transferrable to other areas of mountainous terrain and seasonal snow.

2. Real-time forecast system

The conceptual design of the INFORM real-time forecast system (RTFS) is tailored to use operationally available data and has the following key characteristics:

(a) Ensemble forecasts of mean areal precipitation and temperature as well as reservoir inflow are based on dynamical downscaling methods.
(b) Hydrologic components (snow and soil water models, and runoff and flow routing models) used are adaptations of the operational hydrologic forecast models of the NWS California Nevada River Forecast Center (CNRFC).
(c) Updated estimates of snow and soil water content states from CNRFC are assimilated once a day to align the hydrologic forecast system estimates to CNRFC operational forecast system estimates.

The first characteristic specifies dynamic downscaling methods as opposed to statistical downscaling methods to allow the preservation of the synoptic coherence of regional precipitation and temperature changes as weather systems develop and pass over the region of interest. Such coherence is important to allow good reproduction of the short term co-variability of precipitation and temperature that determines reservoir inflow variability, especially in an environment with significant seasonal snow cover. The second characteristic is imposed by the operational nature of the demonstration project and is a prerequisite for operational utilization of the forecast products by the collaborating forecast and management agencies in Northern California. Similarly, the last characteristic allows the incorporation of corrections based on data and forecaster experience and assures compatibility with operational procedures.

Fig. 2 shows a schematic representation of the RTFS components and interactions.

There are two main sources of real-time, ensemble, large-scale forecast information for the INFORM RTFS, both originating at NOAA NCEP. The first is the Global Ensemble Forecast System (GEFS) (e.g., Hamill et al., 2011, 2013) and the second is the Climate Forecast System (CFS) (Saha et al., 2006, 2014). The first is used for 0–16 day forecasts with a spatial resolution of approximately 100 km, and the second is used for forecasts up to 41.5 days with a spatial resolution of approximately 100 km (version 2, see discussion below). There were two versions of CFS output used as described below. We discuss next each of the RTFS pathways that emanate from these sources (see Fig. 2).

The mesoscale Weather Research and Forecasting (WRF) model was used to dynamically downscale the 0–16 day 20-member ensemble predictions from the GEFS operational system on a 10 × 10 km^2 spatial and a 6-hourly temporal scale (see following section for more details on implementation). However, test simulations revealed that running WRF with CFS ensemble input beyond this forecast lead time (i.e., 16 days) is impractical. In addition, use of WRF for longer lead times assures divergence of the WRF land-surface component states (i.e., soil water estimates) from those of the operational hydrologic models used by the CNRFC because of different model structure and physics (e.g., the former uses energy fluxes to estimate the potential rates of evapotranspiration, while the latter uses climatological estimates based on observed data), which also prevents direct assimilation of operational hydrologic model states into WRF.

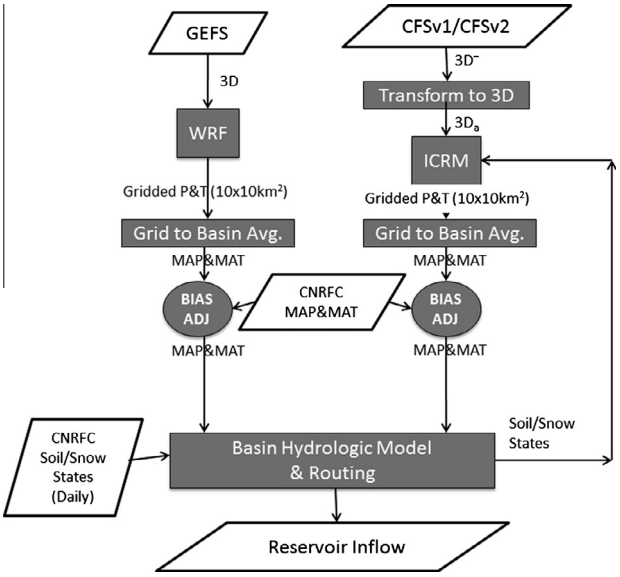

Fig. 2. Conceptual design of the INFORM RTFS for evaluation. CNRFC MAP and MAT refers to observational data used in deriving forecast bias adjustments.

Thus, WRF is used for the short-term forecasts (0–16 days) that support short-range objectives (e.g., flood control, hydroelectric power concerns) at the reservoir sites. A computationally-efficient intermediate complexity regional model (ICRM) is utilized to provide dynamic downscaling for the CFS forecasts of three dimensional (3-D) atmospheric variables to produce surface precipitation and temperature fields out to 41.5 days maximum lead time. In earlier studies, the ICRM forced by the 3-D (profile) ensemble forecasts from the CFS predicted the occurrence of heavy rainfall for lead times greater than 16 days out to 30 days under strong synoptic forcing situations in Northern California (Georgakakos et al., 2010a,b). A description of the ICRM implementation is given in a subsequent section. In this section the links to the other models are highlighted.

The ICRM uses a land surface component with snowpack and soil water modeling that are gridded adaptations of the analogous basin-based operational hydrologic models run at CNRFC. This allows assimilation of operational model states that include any forecaster updates/adjustments once per day to align the initial conditions of the INFORM ICRM integrations with those of the operational CNRFC hydrologic models.

Two different CFS forecast output streams, CFSv1 (or CFS1) and then CFSv2 (or CFS2), became available from NCEP during the evaluation period. In the latter and most recent case, CFS2 provides 16 ensemble members per day (4 members for each of 4 initializations at 00Z, 06Z, 12Z and 18Z) and with the maximum lead time of 41.5 days for INFORM. It is noted that at present the available 3-D information in the CFS1 and CFS2 forecasts is not appropriate for running ICRM directly, so there is a transform module developed to convert the existing 3-D information (indicated as 3D$^-$ in Fig. 2) to ICRM-compatible 3-D information (indicated as 3D$_a$ in Fig. 2).

The WRF and ICRM runs generate 10×10 km^2 gridded surface precipitation and temperature ensemble forecasts for Northern California for the period 0–16 days (GFS 3-D based) and 0–41.5 days (CFS 3-D_a based), respectively. To use these as input to the basin-based hydrologic model of INFORM, requires converting gridded information to mean areal precipitation (MAP) and temperature (MAT) for each of the CNRFC-delineated sub-catchments of interest. This is accomplished by using the available Geographic Information System (GIS) and the model output grid specification (see the box with label "Grid to Basin Avg." in Fig. 2). This conversion facilitates the production of MAP and MAT ensemble forecasts for each of the sub-catchments of the CNRFC/INFORM hydrologic forecast system.

Adjustment of the ensemble MAP and MAT forecasts to comply with the distributional characteristics of the corresponding observations obtained from CNRFC was also made for improved input to the hydrologic models and it is indicated in Fig. 2 (described in a later section). Both the unadjusted and the adjusted MAP and MAT real-time ensemble forecasts and resultant ensemble reservoir inflow forecasts are evaluated in this work.

The INFORM hydrologic model is basin-based and uses tailored river routing schemes for all the rivers of interest. It runs both with the WRF and the ICRM ensemble forecast output mentioned above to produce reservoir inflow ensembles for all the reservoir sites of interest with 6-hourly temporal resolution and with forecast lead times that span the range from 6 h to 41.5 days. The decision to make the hydrologic model basin-based derives from the requirement to align and preserve of the operational CNRFC hydrologic catchment configuration in INFORM, and the requirement for INFORM to maintain use of operational components to the extent possible. In addition, the operational hydrologic forecast models have very good performance in simulations and predictions (e.g., Shamir et al., 2006) and there is no compelling reason to diverge from them in the INFORM system for the scales of interest.

Because of the use of different input and models for obtaining the MAP and MAT forcing, it is expected that the error structure of the ensemble reservoir inflow predictions from WRF and ICRM will differ substantially (this is confirmed in the performance assessments described in the following sections). In addition, it is necessary to merge the WRF-based ensemble reservoir inflow predictions (out to 16 days) with those of the ICRM-based ensemble predictions (out to 41.5 days) and generate the same number of ensemble members throughout the period from 0 to 41.5 days lead time. To achieve this in real time, it is necessary to align the statistical character of the ensemble predictions from these two operational streams shown in Fig. 2. HRC-GWRI (2013) presents the methodology used in INFORM for this. Because in this work we focus on the evaluation and intercomparison of the predictions, we evaluate each forecast stream of Fig. 2 independently. Note that the predictions of reservoir inflows are for unimpaired flows (not including upstream regulation effects), also called full natural flows (FNFs), and such forecasts are evaluated using estimates of FNF based on observed reservoir data such as levels and outflows using water balance analysis for the reservoir.

It is desirable to use the snowpack and soil content states of the hydrologic model component to update the land-surface component of the ICRM (grid based). This was achieved through a feedback from the hydrologic model to the ICRM. This adjustment is made with a lag of −6 h. Note that ICRM runs on a 6-h cycle so the assimilation with a 6-h lag is feasible.

2.1. WRF model implementation

The INFORM RTFS system utilizes version 3.2.1 of the Advanced Research WRF (ARW) dynamical core. The WRF-ARW (called WRF herein) is a state of the art mesoscale model designed for both research and operational applications and is based on the MM5 mesoscale model (Dudhia, 1993). The equation set used by the model is fully compressible, Euler non-hydrostatic with a terrain following, hydrostatic pressure vertical coordinate. A detailed description of the WRF-ARW can be found in Skamarock et al. (2008).

The application of a mesoscale model to an area of interest is typically made by configuring a system of nested grids, the design of which is strongly influenced by the data available for definition of the model's initial and boundary conditions. The GEFS system includes a control forecast and 20 perturbed forecasts that are run out to 16 days on a T190L28 horizontal and vertical resolution. Each of the 20 GEFS ensemble perturbation forecasts is used to drive a separate instance of WRF, with each of these ensemble perturbations being centered and orthogonal at the initial time. Because the resolution of the GEFS ($1^\circ \times 1^\circ$ as supplied by the NCEP) is much coarser than the $10 \times 10^\circ$km^2 resolution of the INFORM WRF, the initial conditions used for each of the WRF runs are supplied by the analysis time of NCEP's North American Model (NAM). These analyses are on a Lambert conformal grid (resolution of approximately 12 km in the horizontal) and are generated by combining model output, i.e. the GFS control run, and observations from various sources (NCEP BUFR data).

The WRF was configured with two nested Lambert conformal grids (Fig. 3) with horizontal resolutions of 30 km and 10 km and corresponding mesh sizes of 30×30 and 52×55 grid points, respectively (the latter labeled "d02" in Fig. 3). A two-way interactive communication occurs between the nested grids. Each grid contains 30 computational atmospheric layers with the finest vertical resolution in the lowest 2 km; the highest computational layer is at approximately 50 hPa (~20 km above sea level) and the lowest layer is approximately 28 m above ground level. Computational time steps of 180 and 60°s are used on the outer and innermost grids, respectively. Surface topography, land use and soil type for the WRF are taken from the United States Geological Survey (USGS) terrestrial data with a spatial resolution of 30°s (approximately 1 km). General terrestrial input files at a resolution of 1° supply the albedo, greenness fraction, slope category, and deep soil temperature. Parameterizations selected for use within WRF include the Kain–Fritsch convective parameterization (Kain, 2004) and the NOAH land surface model (Livneh et al., 2010) along with the WRF Single-Moment 6-class microphysical scheme (e.g., Hong and Lim, 2006).

2.2. ICRM implementation

As mentioned earlier ICRM is an intermediate complexity coupled atmospheric and land surface model used to downscale the NCEP CFS ensemble forecasts for the INFORM system. Over the course of INFORM, NCEP made available two versions of the CFS three-dimensional output. Initially, CFS1 offered a minimal set of upper air variables with four ensemble members produced once daily. Later with CFS2, a more complete set of upper air variables was made available with four ensemble members produced four times a day. To maintain real time forecast capability, the INFORM RTFS was enhanced to work first with CFS1 and then with CFS2. The change-over from CFS1 data to CFS2 data for the RTFS was on the 21st of February 2012.

As noted, CFS1 provided a very limited set of upper-air forecast data (few variables, few levels). In contrast, the orographic precipitation model used in the ICRM atmospheric component requires a complete vertical profile of meteorological conditions [temperature (T), humidity (Q), winds (U), and heights (Z)] on pressure levels from near the surface to ~200 hPa at representative locations upstream (to the west) of the Sierra Nevada. For instance, Oakland, California is well situated for this purpose, and previous work has

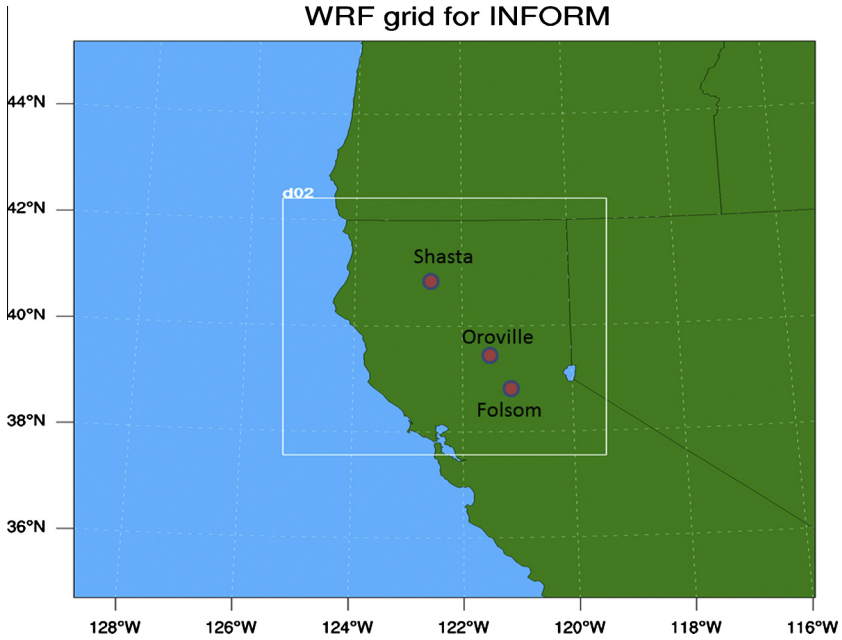

Fig. 3. The computational grids used in the WRF-ARW simulations. The coarse outermost nest takes up the entire map with the boundaries of the inner nest indicated by the white rectangle (labeled "d02").

shown that the use of actual radiosonde soundings from this location to provide boundary conditions gives good results with ICRM (HRC-GWRI, 2007). In contrast to these requirements, CFS1 provides only (a) heights on the 1000, 850, 700, 500 and 200 hPa pressure surfaces, (b) winds at 850 hPa, and (c) integrated column water content (P_{WAT}). Winds on the 700, 500, and 200 hPa pressure surfaces can be derived using the assumption of geostrophy (e.g., Wallace and Hobbs, 2006). Thus, the required variables available or immediately derivable from CFS1 are as shown in Table 1; those required by the INFORM RTFS system are as shown in Table 2.

The basic problem is to develop a statistical model that can produce estimates of the required unknown quantities in Table 2 (those marked "D") when supplied with a single set of the variables in Table 1 (those that are available from CFS1). The necessary models were developed using a technique commonly known as Principle Component ("PC") regression (e.g., Jolliffe, 2002). In PC regression, the predictor and predictand data (in this case, the data available from CFS1 and the data required for the ICRM model, respectively) are expressed in terms of linear combinations (determined by regressions) of orthogonal basis functions known as "Empirical Orthogonal Functions" (EOFs). These paired functions express different modes of variability in "variable space" (e.g., the 16 predictor or 14 predictand variables; these are known as "loadings") and "time" (e.g., the number of observations; these are the "PCs"). For PC regression, the EOF analysis serves to smooth the predictor and predictands, and improves the stability of the regression parameters.

Table 1
Surface and upper air variables available from CFS1 at each grid point. ("X" indicates a CFS1 output variable; "G" indicates that the winds are derived using geostrophy.)

Pressure (hPa)	*Z*	*U*	*V*	P_{WAT}
1000	X	G	G	
850	X	X	X	
700	X	G	G	
500	X	G	G	
200	X	G	G	
Full column				X

Table 2
Surface and upper air variables required for upstream sites for INFORM ICRM. ("X" indicates variables available or easily estimated from CFS1 output; "D" indicates variables that must be statistically derived.)

Pressure (hPa)	*Z*	*T*	*Q*	*U*	*V*
1000	X	D	D	X	X
850	X	D	D	X	X
700	X	D	D	X	X
500	X	D	D	X	X
400	X	D	D	D	D
200	X	D	D	X	X

The CFS2 ensemble forecast output is substantially more extensive than that from CFS1 (Table 3). In this case, generation of ICRM input for each time and ensemble member requires hypsometric estimation for *T* and *Z* to 925 and 250 mbar levels from the neighboring higher and lower levels, for which the values exist. Estimates of specific humidity for the 1000, 250 and 200 hPa levels were obtained by setting the first equal to the 925 hPa value and the latter two to zero.

The atmospheric component of the ICRM is an orographic precipitation model whose formulation is described in HRC-GWRI (2007) and in Georgakakos et al. (2012b). The 10×10 km^2 surface grid of the ICRM is shown in Fig. 4 together with the two upstream CFS upper air input sites near Oakland and Eureka, California.

The 3-D ensemble forecasts of CFS and the resultant ICRM precipitation output have a 12-h resolution. To generate 6-hourly output for the downstream hydrologic models, a uniform temporal distribution of the 12-hourly totals to 6-hourly increments was used based on analysis of the distribution of mean areal

Table 3
Upper air and surface variable availability for CFS2 output. (Numerical values are in hPa.)

T		1000		850	700	500	250	200
q	2 m		925	850	700	500		
u&v	10 m	1000	925	850	700	500	250	200
Z		1000		850	700	500		200

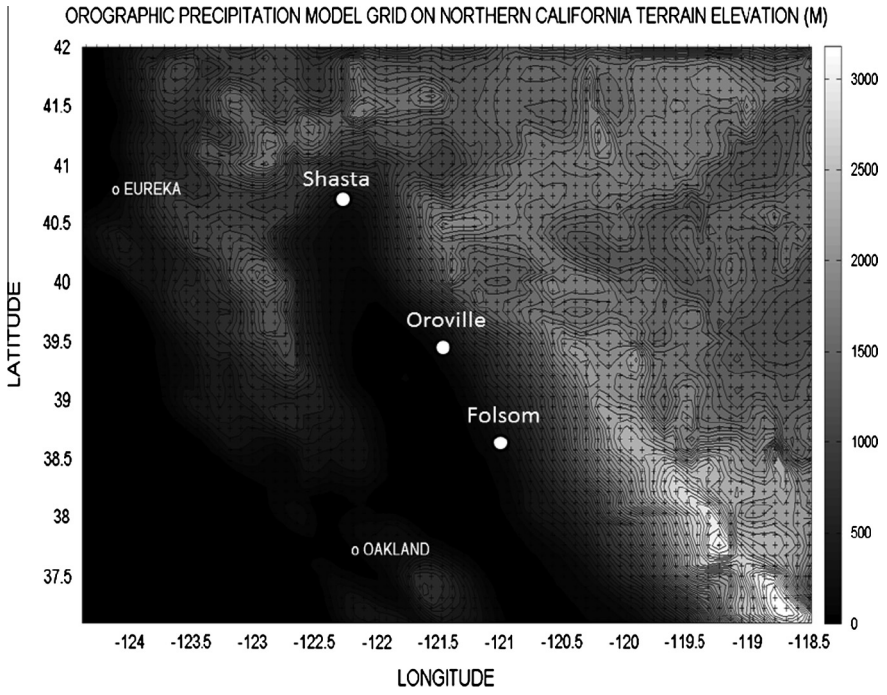

Fig. 4. Surface grid of the ICRM for Northern California. The two vertical profiles at Eureka and Oakland provide CFS three-dimensional ensemble forecast input.

precipitation during the day for the watersheds of interest (HRC-GWRI, 2013).

Downscaling of the CFS ensemble forecasts through the use of ICRM is accomplished as follows: first, the orographic model component of the ICRM is executed with a basic computational interval of 112 s to provide ensemble gridded precipitation estimates at a 10 km horizontal resolution, a 1000 m vertical resolution in mid-troposphere, and a 6-h temporal resolution out to 41.5 days. This component uses CFS ensemble input from two grid points that provide free-stream upstream moist inflow to the ICRM (coastal CFS nodes at Eureka and Oakland in Fig. 4) Next, MAP estimates are obtained for each INFORM subcatchment by area weighting of the gridded model precipitation output within each subcatchment. The MAP estimates are then used to drive the basin hydrologic and routing models described earlier, and to provide cloud cover input for the surface air temperature component of ICRM (see description below).

It is important to note that the orographic model component is a simplified dynamical model that generates precipitation estimates due to the interaction of the moist-wind field and the orographic terrain. The model forecasts are thus suitable for the mountainous terrain of Northern California, where most of the reservoir inflow is generated, while little precipitation is produced in the valley regions. Earlier evaluation (see HRC-GWRI, 2007) showed good performance in estimating the occurrence and amounts of MAP, with some overestimation in higher elevations and the previously noted underestimation in lower elevations.

The ICRM also incorporates a land-surface model used for the estimation of gridded surface air temperature. This formulation is based on surface energy balance considerations for the estimation of the surface (skin) temperature at each grid box (Fig. 4). The surface temperature T_o is computed as the solution of the diagnostic surface energy balance equation, which, for generality that includes melting snow, may be written as (e.g., Pielke, 1984; Liston, 1995):

$$(1-a)Q_{si} + Q_{li} + Q_{lo} + Q_H + Q_E + Q_G = Q_m \quad (1)$$

where Q_{si} is solar radiation reaching the surface, Q_{li} is incoming longwave radiation, Q_{lo} is outgoing longwave radiation, Q_H is sensible turbulent heat flux, Q_E is latent turbulent heat flux, Q_G is conductive energy transport (assumed negligible), Q_m is energy available for melt, and a is surface albedo. The parameterizations of the different heat fluxes as functions of the reference and surface temperatures, reference relative humidity, pressure and wind speed, the presence of precipitating and non-precipitating clouds, the surface soil water saturation level, the presence or absence of snow, and the land surface parameters such as land use type, surface albedo, emissivity and aerodynamic roughness, are discussed in Georgakakos et al. (2012b) and in HRC-GWRI (2013) and are not detailed here. The second to fifth terms of the left-hand side of Eq. (1) (from Q_{li} through Q_E) are functions of surface temperature T_o and thus, given parameterizations of the other terms, T_o may be obtained as the algebraic solution of the aforementioned equation. Appendix A provides the solution method for T_o and the 2-m air temperature, T_a, needed for the snow component model.

To obtain estimates of snow cover and soil water saturation levels at the ground surface and to assure consistency with operational hydrologic forecast models (that will use the surface temperature forcing from the ICRM), gridded versions of the operational snow and soil water models were used with parameter values estimated as described in HRC-GWRI (2013, Appendix B). For real time operation, the relative fraction of water content is estimated over a basin by the basin hydrologic models of the INFORM RTFS and this fraction is distributed over the ICRM grids within the basin to provide initial values for the soil model of ICRM in the feedback process shown in Fig. 2. The snow fraction and snow water equivalent are similarly distributed prior to each ICRM integration.

2.2.1. Cloud influence

The orographic precipitation model provides information pertaining to the development of orographic precipitating clouds. These computed precipitating clouds constitute a minimal cloud condition. However, clouds exist at other times when precipitation

is not occurring. Due to the profound influence that clouds have on surface temperature it is important to account for their development. An important aspect of the computations is the establishment of conditions for the development of clouds through parcel ascent over the mountain barriers in Northern California. Sensitivity studies and available literature indicates that conditions for parcel ascent are that at the 850 mbar level (above the climatological snowline and boundary layer depth) the air is near saturation (relative humidity of 90% or higher), the wind speed is greater than 5 m/s, and the wind is from a westerly direction (this to assure that the moist air will be advected over the mountain barrier rather than be blocked); in addition, and based on Pandey et al. (1999), clouds may only develop when and where the 700 mbar temperature is colder than 6 °C.

Several sensitivity analyses were performed (see HRC-GWRI (2013) for details) using CNRFC mean areal temperature (MAT) data and various conditions for cloud development: minimal cloud condition involving only precipitating clouds, a maximal cloud condition involving unconditional parcel ascent, and a conditional cloud condition involving cloud ascent with the constraints described earlier. The results indicate that the conditional cloud development approach gives best results for the subcatchments with elevations above the mean snow line, while for the lower elevations the max cloud approach offers comparable (and even better results in some cases). The decision to use the conditional cloud approach is based on the fact that the ICRM temperature is most significant for predicting snow pack development and melting and it is the snow pack at higher elevations (e.g., higher than an average snow line of 1,500 m) that is most important for downstream flow predictions.

2.3. INFORM RTFS hydrologic models

The hydrologic models of INFORM RTFS closely follow the operational hydrologic forecast models used by the California Nevada River Forecast Center (CNRFC). The snow and soil water models are adapted (e.g., Anderson, 1973; Georgakakos, 1986; Shamir et al., 2006) from the Community Hydrologic Prediction System (CHPS), formerly the National Weather Service River Forecast System (NWSRFS), and the hydrologic segments within INFORM are based on CNRFC-defined watershed areas for operational forecasting. Thus the hydrologic model components are basin-based and are updated in terms of their sub-basin definitions and parameters to align as previously noted with current CNRFC operations for the five major INFORM reservoir watersheds.

Inputs to the hydrologic model components are the basin MAP and MAT forecasted by the WRF and the ICRM as described earlier. The snow and soil water model components produce estimates of snow depth, snow melt and runoff during the cold season, along with surface and subsurface runoff for each watershed sub-basin throughout the year. This output is input to a channel kinematic routing component (Georgakakos and Bras, 1982), tailored for each major reservoir watershed, to produce streamflow estimates at each sub-watershed and an estimate of total reservoir inflow in each case.

Model parameters for the snow and soil models are derived from the CNRFC operational hydrologic model parameters, which are based on CNRFC calibration of natural flows into the reservoirs. Parameters of the kinematic routing component for each reservoir watershed were estimated using historical streamflow data and the CNRFC-estimated unit hydrograph at the hydrologic segments. The mathematical basis of the hydrologic models of INFORM has been given in HRC-GWRI (2007, Chapter 4) and will not be repeated here.

The simulation performance of the hydrologic models using operational model parameter estimates was compared to unimpaired or "Full Natural Flow" (FNF) estimates obtained from the California Data Exchange Center (CDEC) (see HRC-GWRI (2013) for details). The records used were at least ten years long. Table 4 summarizes the daily simulation performance by reservoir inflow. The overall correlation is lowest for Trinity Reservoir inflows (0.77) and is highest (0.92) for both Folsom and Shasta Reservoir inflows. These high correlation values indicate good agreement between the variability of the simulations and the FNF observations. The bias statistics indicate an over-estimation of the observed flows by approximately 15–50%. In this context, fractional bias is defined as the residual mean over the observations mean.

Fig. 5 presents the comparison of the cumulative distribution of daily flow for the observations (FNF, in black) and simulations (in red) for each of the reservoirs. For all reservoirs, good reproduction of the observed daily flow distribution is found. The largest discrepancies are observed for Trinity and Oroville reservoirs; both showing an over-estimation of the observations over a range from mid- to high-flows. It is noted that estimation errors in MAP and MAT from sparse point observations contribute to the flow simulation errors of mountainous watersheds.

Additional qualitative comparisons of the simulated snow water equivalent (SWE) with available snow sensors were made to assess the performance of the snow model simulations for the period from January 2011 to January 2013. Although direct comparison is not meaningful (point data vs. areal average values), the results indicate that the patterns of accumulation and ablation shown in the snow sensor records are captured by the simulations; especially for the accumulation regions of the upper elevations in the watersheds (see HRC-GWRI (2013) for several examples).

2.4. MAP and MAT bias adjustment

Persistent errors arise in the forecasts of MAP and MAT as a result of imperfect models (WRF and ICRM) and imperfect large scale forcing (GEFS and CFS). Adjustment of the MAP and MAT ensemble forecasts to correct for bias was done for each subcatchment, model (e.g., WRF or ICRM) and each season (excluding the dry summer season): November through February (NDJF) and March through May (MAM). As will be discussed in the next section, in most cases biases in MAP and MAT vary little with lead time. This consistency allows focus on a particular lead time for the estimation of the bias factor, which can then be applied to all lead times. A probabilistic approach was taken to account for the ensemble uncertainty and for the distribution of model forecast errors. The methodology is shown schematically in Fig. 6 that is created with synthetic data for illustration purposes.

Consider the cumulative frequency plot of the forecast ensemble daily mean areal precipitation produced by a given model for a given season and subcatchment, and for a particular lead time (shown is the forecast with lead time of 72 h). The cumulative frequency plot may be divided into deciles (each decile then contains the same number of ensemble forecasts). The mean forecast (denoted by Mean(F)) may then be estimated for each decile. Corresponding to the forecasts of each decile there are observations of

Table 4
Statistical indices comparing the daily simulations and observed FNF.

Reservoir	Correlation coefficient	Fractional bias
Folsom	0.92	0.24
Oroville	0.89	0.53
Shasta	0.92	−0.01
Trinity	0.77	0.44
New Bullards Bar	0.81	0.15
Englebright	0.87	0.21

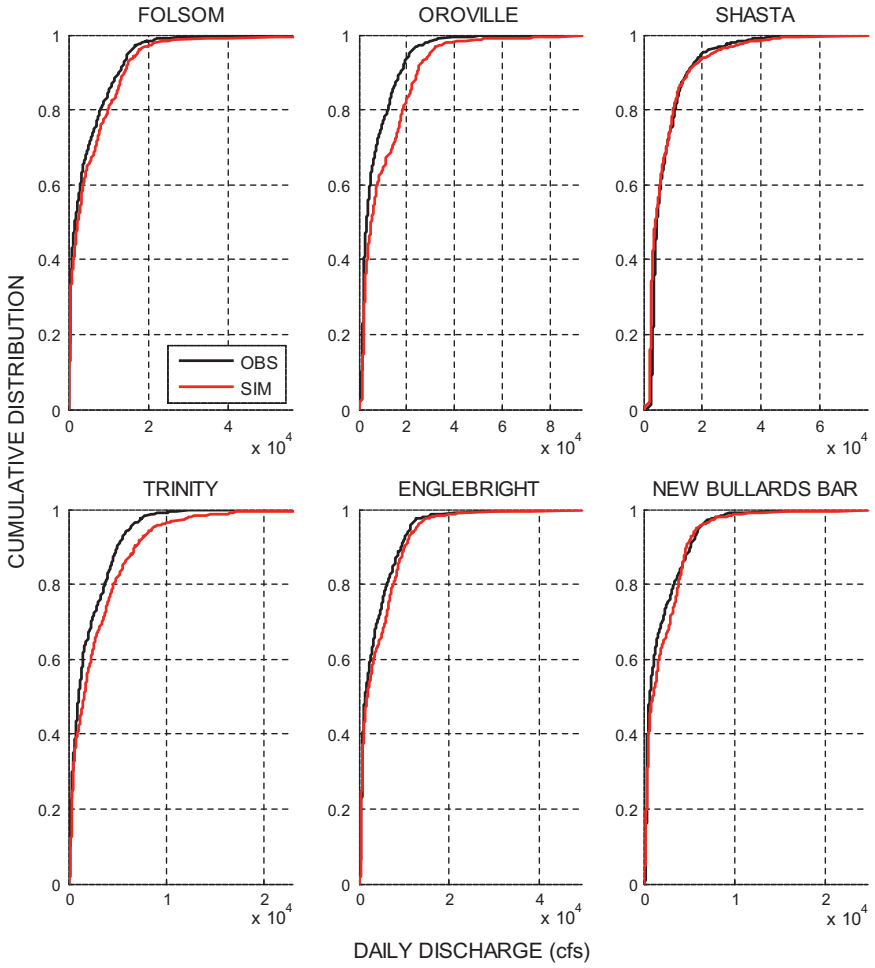

Fig. 5. Cumulative distribution of daily flows. Simulations are in red, observed FNF is in black, with the discharge presented in a log scale. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

mean areal precipitation estimated by CNRFC using station observations. These observations, relevant to a particular forecast decile, span a range and form their own cumulative distribution function as depicted schematically in the inset of Fig. 6 (the better the model forecasts the narrower the range of observations near the forecast decile). Selecting a statistic of the distribution of observations (in the schematic we selected the 70th percentile value $P_{0.7}$ as an example), one may define the bias factor for each decile as the ratio: $P_{0.7}$/Mean(F). Multiplication of the model values in the decile with the computed factor then provides for adjustment of the model ensemble forecasts to reduce their bias with respect to observations. As a result of a small scale sensitivity study in the present implementation of INFORM, the numerator for the computation of the bias factor is $P_{0.3}$ for the deciles that are less than the cumulative frequency of 0.3, and $P_{0.7}$ for the deciles greater than 0.7. The deciles between 0.3 and 0.7 use the mean in the numerator of the bias factor.

The procedure used is designed to reduce systematic bias in the forecast data while retaining the estimated forecast uncertainty in the system forecasts. Retaining forecast uncertainty reliability is expected to be very important for effective reservoir management (e.g., Georgakakos and Graham, 2008).

3. Evaluation of real-time forecasts

This section presents a selection of verification results comparing INFORM RTFS atmospheric and hydrological model forecasts with observations. The results shown here cover the drainage basins of the Folsom, New Bullards Bar/Englebright, Oroville, Shasta, and Trinity reservoirs (Fig. 1). For the atmospheric model forecasts we compare forecast and observed MAP and MAT, with the latter derived by NOAA/NWS CNRFC from *in situ* measurements. For the hydrologic models, the forecast inflows are compared to full-natural-flow (FNF) estimates of the inflows to the primary reservoirs in each watershed. The next section provides a brief outline of how the observational and forecast data were processed and a description of the statistical measures used for performance assessment. The assessments themselves, for MAP, MAT and INFLOW, follow in separate sections. Evaluation is done first for the unadjusted MAP and MAT forecasts and then for the adjusted MAP and MAT forecasts. It is emphasized that the focus is on the evaluation of the actual real time forecasts from the INFORM RTFS that are of interest for true operational real time reservoir management rather than distilled subsets that would be appropriate for evaluating individual models and model components. Examples of the latter subsets are in Hamill et al. (2013) for the GEFS and in Saha et al. (2013) for the CFSv2.

3.1. Data processing

3.1.1. MAP and MAT data

Observation-based estimates of MAP (mm 6 h^{-1}) and MAT (°C) were obtained from CNRFC for each of the sub-basins within the main watersheds. These data have 6-h temporal resolution and

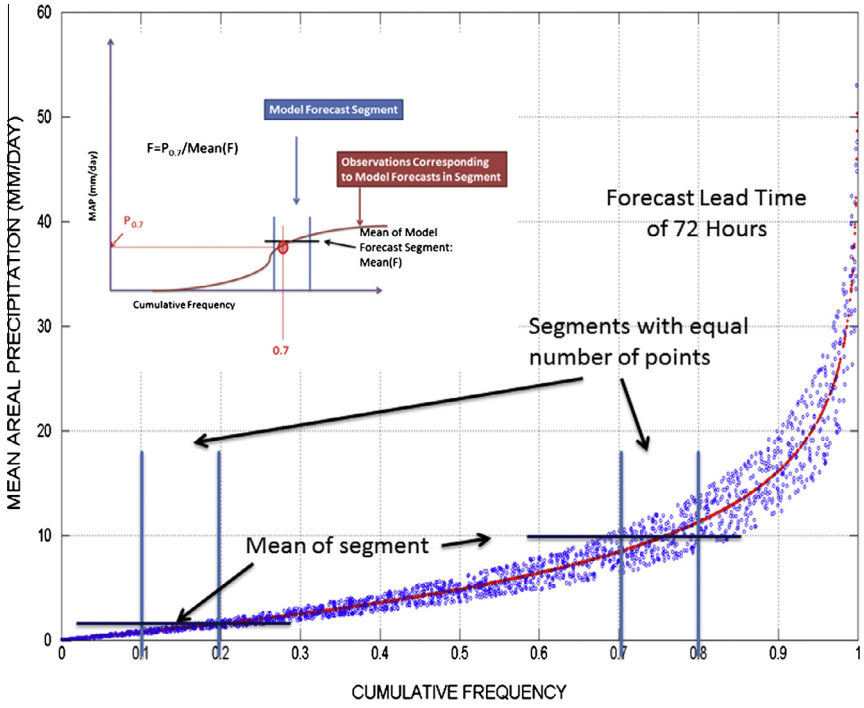

Fig. 6. Schematic for bias factor estimation using a probabilistic approach. The cumulative frequency of the mean areal precipitation forecasts is based on ensemble forecasts with a fixed lead time of 72 h. The inset shows the distributions of observations that fall within a single tercile of the forecast distribution.

cover 00Z on 15 June 2009 through 00Z on 4 November 2012. The data were processed to provide 24-h averages (MAT) and accumulations (MAP) for each 24-h period in the record ending at 00Z or 12Z (the validation time). The observed data times were converted from PST to UTC for the validation. The MAP and MAT comparisons discussed focus on the time periods ending at 00Z (4:00 PM PST), with some discussion of 12Z results as appropriate, as the results do not change much for the different validation times.

Forecast MAP and MAT were available from ensemble simulations with the three atmospheric modeling systems described earlier, designated ICRM-CFS1, WRF, and ICRM-CFS2. These data were available for the constituent sub-basins for times described below. Each forecast system produces output for each 6-h increment through a given forecast run.

To help clarify the following discussion, the ICRM-CFS1 system provided one four member ensemble forecast per day (initialized at 00Z) going out to a lead time of 41.5 days (996 h) and operated until October 2012 when the CFS1 ceased operation. The limitation of just four ensemble members per day precludes the use of daily probabilistic statistics with ICRM-CFS1.

The ICRM-CFS2 system produces a four-member ensemble four times per day (at 00Z, 06Z, 12Z and 18Z) using boundary conditions from the NOAA CFS2 climate forecast system. These forecasts also run out to a lead time of 996 h. This system began running on 21 February 2012 and the final forecast used here was produced on 12 November 2012.

The WRF system is driven by NOAA NCEP GEFS output and produces two 20-member ensemble forecasts per day (initialized at 00Z and 12Z), these running out to a lead time of 384 h (16 days). This system produced forecast data from 7 November 2011 through 2 November 2012.

The output from each forecast system contains gaps when some or all of the ensemble members failed to complete (our analyses required all ensemble members to be present), nevertheless the sample sizes are sufficient to provide statistical guidance concerning performance. Table 5 summarizes the time period covered by observational and forecast system data.

Table 5
Time periods covered by observational data and forecast system output.

Data source	Coverage
Observations (MAP and MAT)	06/15/2009–11/04/2012
Observations (INFLOW)	01/01/2010–01/08/2013
ICRM-CFS1	11/26/2010–10/23/2012
WRF	11/07/2011–11/02/2012
ICRM-CFS2	02/21/2012–11/12/2012

3.1.2. INFLOW data

The INFLOW observational data was provided as the average FNF discharge rate (in cubic feet per second, cfs) for the 24-h period ending at approximately 12Z each day (this contrasts with the 6-hourly observational data for MAT and MAP). The forecast INFLOW data was in the same form described above for MAT and MAP giving inflow for each 6-h lead time increment.

3.1.3. Forecast averaging

For most of the results described here, the forecast data for WRF and ICRM-CFS2 were processed to provide 24-h averages from multiple lead times for comparison with the observed 24-h averages (verifying averaging blocks ended at 12Z for INFLOW data; averaging blocks for MAP and MAT ended at 00Z for the results shown in the following). The use of observed and forecast averages in the processing scheme is exemplified in the diagram of Fig. 7. Forecast initialization time and observation times are shown along the left side, and forecast lead times are shown along the top. The open and filled circles in shaded blue boxes (lower left) indicate the period of time going into one observed 24-h average (average of 4 6-hourly values for MAT and INFLOW; accumulation from 4 6-hourly averages for MAP), with "validation time" indicated by the filled circle (at 12Z on 6 March in this case). The line segments of the filled triangles indicate the "blocks" over which the forecast averages of different lead times were calculated to correspond to the observed 24-h period described earlier (blue shaded region in

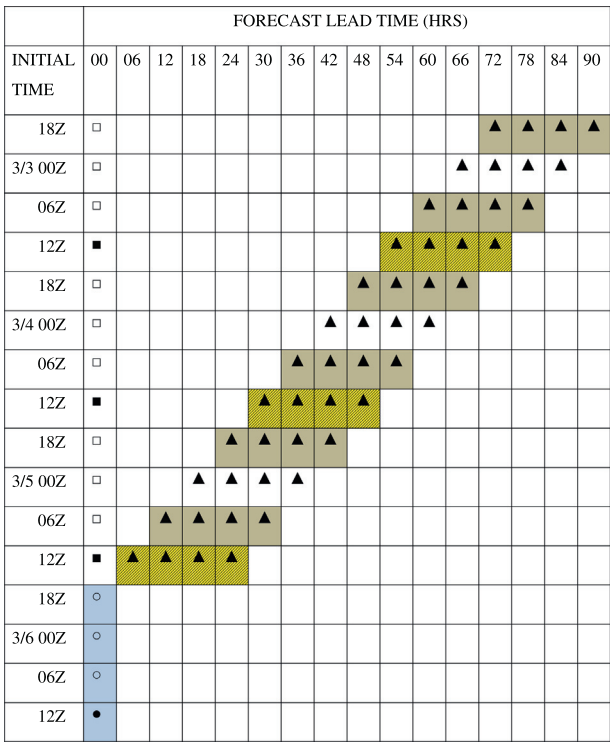

Fig. 7. Diagram showing the "24-h Block Average" verification scheme. (The symbols are explained in the main text.)

lower left). Here, the lowest line of filled triangles in yellow shaded boxes marks the most recent set of forecasts validating at 12Z on 6 March and these belong to the "00-h lead" forecast using the notation in the text. Other line segments with filled triangles signify forecasts with earlier initializations, all validating at 12 Z on 6 March. Solid triangles in yellow shaded boxes signify forecasts initiated at 12Z. The WRF forecast system has initialization times every 12 h; yellow shading and no-shading indicates forecast blocks with initial times used for that system. The ICRM-CFS2 system was initialized every 6 h and uses all the line segments shown with solid triangles. 24-h flow FNF "observations" are only available at ~12Z, so validation statistics are for 12Z only. WRF forecasts went out to 384 h lead time, so the last block having the full complement of forecasts has a nominal lead time of 336 h (14 days). ICRM-CFS2 forecasts ran out to 996 h (41.5 days) but were processed to provide statistics out to a nominal lead time of 720 h (30 days). Note that each forecast is composed of individual ensemble members that are not indicated in Fig. 7. The yellow shaded line segments show the averaging scheme used for the ICRM-CFS1 MAP and MAT forecast output initialized once per day.

The basic idea is to form "daily-ensembles" from all of the forecasts that (a) are valid for the 24-h observational averaging period of interest, and (b) come from separate lead time blocks, each 24 h in length. For each such block, the 24-h average (or accumulation) was calculated for each available ensemble member. The average of these ensemble averages or accumulations provided an overall cross-ensemble average. This procedure provides individual ensemble averages and a cross-ensemble average for each 24-h lead time block going out to a lead time of 720 h for ICRM-CFS2 and 336 h for WRF. Referring to Fig. 7, one can see that what is denoted a "00 h lead" forecast here is actually available ~24 h prior to verification time. With ICRM-CFS2 having four-member ensembles and four initializations per day, there are a total of 16 24-h averages in each daily-ensemble. For WRF, with 20 member ensembles and two initializations per day, each daily-ensemble consists of 40 24-h averages. This processing scheme provides both a scalar value (the daily-ensemble mean) and a set of equiprobable ensemble outcomes for comparison with the observed data. During the period of evaluation and for the WRF processing we used at least 115 cases for MAM and at least 150 cases for NDJF.

An alternative processing scheme was used to incorporate the "single four-member ensemble per day" ICRM-CFS1 data (see Fig. 7). For this method, only the data from the overall cross-ensemble average final forecast in each 24-h block was used (yellow[2] shading in Fig. 7) to compare with the observations.

3.2. Statistical performance measures

These measures compare each observation with a single corresponding multi-ensemble mean of forecasts verifying at the observation time at some defined forecast lead time. In the discussion that follows we use the following nomenclature:

O_T – 24-h average (or accumulation for MAP or inflow) observed at time T ("the verification time").
$F_{T,L}$ – 24-h average (or accumulation for MAP or inflow) cross-ensemble average forecast at lead time L, verifying at time T.
$F_{T,L,E}$ – 24-h average (or accumulation for MAP or inflow) for single ensemble member E at lead time L, verifying at time T.

Symbols with a prime (′) indicate departures from the respective mean of a variable ("anomalies"); "NT" is the length of the record, "NE" is the number of ensemble members; "NL" is the maximum lead time; "T" designates a particular verification time, "E" designates ensemble number and "L" the nominal lead time, a single value (0, 24, 48, etc.) assigned to each "forecast block" (see Fig. 7). Statistics are generated as a function of lead time (1,...,NL), variable (MAP, MAT, INFLOW) and season. Two "seasons" were defined on the basis of the verification time to approximately differentiate between non-melt and melt dominated settings; these are November–February (NDJF, non-melt) and March–May (MAM, melt). No validation was done for the dry season June–October.

For a given variable, lead time (L), and season triplet, corresponding observation-forecast sets were constructed. For example, each observation during NDJF (O_T) was matched (where possible) with the ensemble forecast data for lead L validating at time T, these are $F_{T,L,E}$, $E = 1,2,\ldots,NE$. The cross-ensemble average of the 24-h ensemble means is then

$$F_{T,L} = \mathrm{NE}^{-1}\sum_{\mathrm{NE}}(F_{T,L,E}) \tag{2}$$

where Σ() denotes the arithmetic sum of the quantity in parenthesis. An initial calculation gives the overall seasonal mean for the observations

$$O_{\mathrm{BAR}} = \mathrm{NT}^{-1}\sum_{\mathrm{NT}}(O_T) \tag{3}$$

and for the forecast values verifying during that season at lead time (L)

$$\mathrm{Fbar}_L = \mathrm{NT}^{-1}\sum_{\mathrm{NT}}(F_{T,L}) \tag{4}$$

Some statistics are derived using departures from these mean values and these are indicated with a prime in the discussion below.

The statistical measures defined below (correlation, bias, bias fraction, Brier Skill Score, ROC area) give a single value for each

[2] For interpretation of color in Fig. 7, the reader is referred to the web version of this article.

basin or sub-basin, each season, each forecast lead time (in 24-h blocks) and each variable.

3.2.1. Non-probabilistic measures

Non-probabilistic measures express the "forecast" for a particular verification lead time as a single value, in this case the "grand average" cross-ensemble mean. The three non-probabilistic measures used in the analyses presented in this section (correlation, bias and bias fraction) are described below.

Correlation [R; recall for given basin or sub-basin, variable, and season, there is one value for each lead time (L)].

$$R_L = C_{OF}/(S_O S_F) \tag{5}$$

where

$$C_{OF} = NT^{-1}\sum_{NT}(O'_T F'_{T,L}) \tag{6}$$

$$S_O = \left[NT^{-1}\sum_{NT}(O'_T O'_T)\right]^{1/2} \tag{7}$$

$$S_F = \left[NT^{-1}\sum_{NT}(F'_{T,L} F'_{T,L})\right]^{1/2} \tag{8}$$

Correlation gives a useful measure of linear association between the observations and the forecast "grand means". Correlation does not account for bias, can be sensitive to outliers, and ranges from −1 to 1.

Bias [B]

$$B_L = \text{Fbar}_L - \text{Obar} \tag{9}$$

Bias is used for evaluating systematic errors in MAT.

Bias fraction [BF]

$$BF_L = \text{Fbar}_L \text{Obar}^{-1} \tag{10}$$

Bias fraction is used for evaluating systematic errors in MAP and INFLOW.

3.2.2. Probabilistic measures

In contrast with the non-probabilistic measures, probabilistic measures include information from the individual ensemble means (F) and thus deal directly with questions bearing on the forecast probability of particular outcomes. These measures were not calculated for the ICRM-CFS1 system because its output was limited to four ensemble members per 24-h period; too few to provide robust probabilistic statistics.

3.2.2.1. Brier skill core. The Brier skill score (BSS) assess the accuracy of forecast probabilities for a pre-defined event to occur (e.g., Hsu and Murphy, 1986). The basic idea is to define a specific "event" of interest, for example, whether measureable rain will fall during the verification time. The ensemble forecasts are grouped according to the forecast probability (P_F) that an "event" will occur; P_F is defined as the fraction of ensemble 24-h forecast means ($F_{T,L,E}$) that exceed the threshold for an event. For example, in the analyses reported here, the groups covered the ten probability of occurrence categories 0–10%, 10–20%,...,90–100%. For each of these categories, the actual frequency of occurrence of events is calculated from the observations verifying with the forecasts in a given bin.

For a given set of probabilistic forecasts and observations, the BSS is defined as follows:

$$BSS = 1 - (BS/BS_{ref}) \tag{11}$$

where BS_{ref} is the Brier Score of a reference forecast. In this case the reference forecast is the climatological frequency of occurrence of events (the "base rate"); it gives a score for always forecasting the probability of the event as the base rate. BS is the "Brier Score" (or "half Brier Score") given by

$$BS = NT^{-1}\sum_{k=1,NT}(y_k - o_k)^2 \tag{12}$$

where NT is the number of forecast-observation pairs available, y_k is the forecast probability of an event (the fraction of ensemble members exceeding the event threshold), and o_k is a binary observation value taking the value of 1.0 if an event occurred, and 0.0 otherwise.

BSS gives a measure of the "improvement over climatology" (BS_{ref} being the climatological score) in terms of the forecast system's actual Brier Score. Note that BSS can range from 1 (perfect) to $-\infty$. Because BSS is an aggregate measure, in our results there is a single BSS value for each lead time, basin, variable, and season, so the results are presented as plots of BSS as a function of lead time for each basin, variable and season.

Note that for the performance assessment period of the CFS2-driven ICRM forecasts, relatively few precipitation events are available for the MAM season to compute the observed climatological frequencies given the model forecast. As such, the BSS of this model and season are subject to a great deal of uncertainty and are not presented here.

3.2.2.2. ROC area. Like the reliability-based measure discussed above, the Relative (or Receiver) Operating Characteristic (ROC) Area measure concerns the accuracy of probabilistic forecasts of a pre-defined "event" (e.g. Green and Swets, 1966; Mason and Graham, 2002; Kharin and Zweirs, 2003), but for these measures the forecast-observations pairs are stratified according to whether or not an event was observed (and not whether an event was forecast as in the BSS score). The contingency Table 6 presents the elements and notation needed to apply the ROC concept to an event of interest (e.g., MAP greater than a certain percent of its observed distribution).

Referring to Table 6, the "hit rate" (HR) is calculated as the number of true positives divided by the total number of events (HITS/NE). Similarly, the "false alarm rate" (FR) is calculated as the number of false positives divided by the number of non-events (FALSE ALARMS/NE′). The ROC curve is then developed by estimating the (HR, FR) pairs for various warning thresholds (e.g., percentages of the observed distribution of MAP).

Just as the Brier Skill Score provided an aggregate (albeit reduced) measure for the reliability diagram, the area under the ROC curve (ROC area, or "AROC") summarizes the information on the ROC curve; AROC above 0.5 (the area under the 45° "guessing" line) is considered skillful. As an aggregate measure, AROC results are displayed as a function of lead time for each basin, variable, and season classification (the computation of the AROC was done using a trapezoidal rule, which is conservative as it underestimates the area below the ROC curve for positive ROC values).

The following sections summarize the demonstration results in terms of forecast performance presenting samples of cases for illustration purposes. The analyses of INFORM system forecast performance produced a very large volume of results. For MAP and MAT, there are results for each of the individual sub-basins com-

Table 6
Forecast-observation contingency table.

	Event observed	Event not observed
Event forecast (warning)	True positive (HITS)	False positive (FALSE ALARMS)
Event not forecast (no warning)	False negative (MISSES)	True negative (CORRECT REJECTIONS)
Total	Total events (NE)	Total non-events (NE′)

prising the five major reservoir catchments, lead times ranging up to 384 h, three atmospheric models, and two seasons. Inflow results are also numerous, but are only for a single variable (main reservoir inflow) in each main catchment. This volume of results precludes showing all but a representative sample to give a general sense of system performance as a function of the variable space coordinates noted above; a complete set of results are available in HRC-GWRI (2013, Appendix E).

The presentation of results begins with consideration of MAP and MAT. First, we present an example of correlation and bias results for March–May MAP and MAT in the Folsom reservoir catchment. Next, we give a short discussion of the potential for very long lead MAP forecasts. This is followed by an example of results displayed on a geographic basis (by subcatchment), in this case for WRF model MAP and MAT correlation and bias for the non-melt (NDJF) and melt (MAM) seasons noted earlier. The presentation of MAP and MAT results closes with samples of the Brier Skill Score (BSS) and ROC area (AROC) results for Oroville reservoir. The second part of the results section presents results for reservoir inflow. These discussions give results in terms bias fraction, correlation, BSS and AROC. The third and final part of the results section gives a sense of the effectiveness of the bias correction procedures.

It should be noted that the performance indices selected measure the performance of the precipitation, temperature and reservoir inflow forecasts with respect to observations or reference values of such quantities over a period of time. They do not directly measure the utility of the forecasts for reservoir management, which is the ultimate performance evaluation for the INFORM system. However, the present type of performance evaluation is a necessary first step because of the significant dependence of the forecast utility for reservoir management on forecast biases and ensemble reliability (e.g., Georgakakos and Graham, 2008; Yao and Georgakakos, 2001).

3.3. MAP and MAT assessments

3.3.1. Biases and correlations

Fig. 8 shows sample results for MAP and MAT correlation and bias for all the available model forecasts, the MAM season and for the Folsom drainage basin sub-catchments. For Folsom, the WRF ensemble MAP forecasts show very high correlations ($\geqslant$0.80) with observations for lead times out to about 5 days (120 h) and good correlations ($\geqslant$0.60) with observations for lead times out to about 8 days (192 h). Significantly lower correlations are shown for the ICRM MAP ensemble forecasts for both the CFS1 and CFS2 forcing, with CFS2 forcing showing slightly better results (correlations $\geqslant$ 0.6 for lead times up to about 4 days). There are non-negligible correlations shown for ICRM-CFS2 forecasts for lead times of about 20 days and beyond. This will be discussed later in this section.

MAP forecast fractional bias is clustered around 2 for the WRF forecasts while it is much greater for the ICRM-CFS1 forecasts for a few subcatchments and much lower than 1 (with one implying no bias) for subcatchments with low sloping terrain (these mostly at low elevations). Fractional bias for ICRM-CFS2 MAP forecasts is less than 2 for all the subcatchments and for most of forecast lead

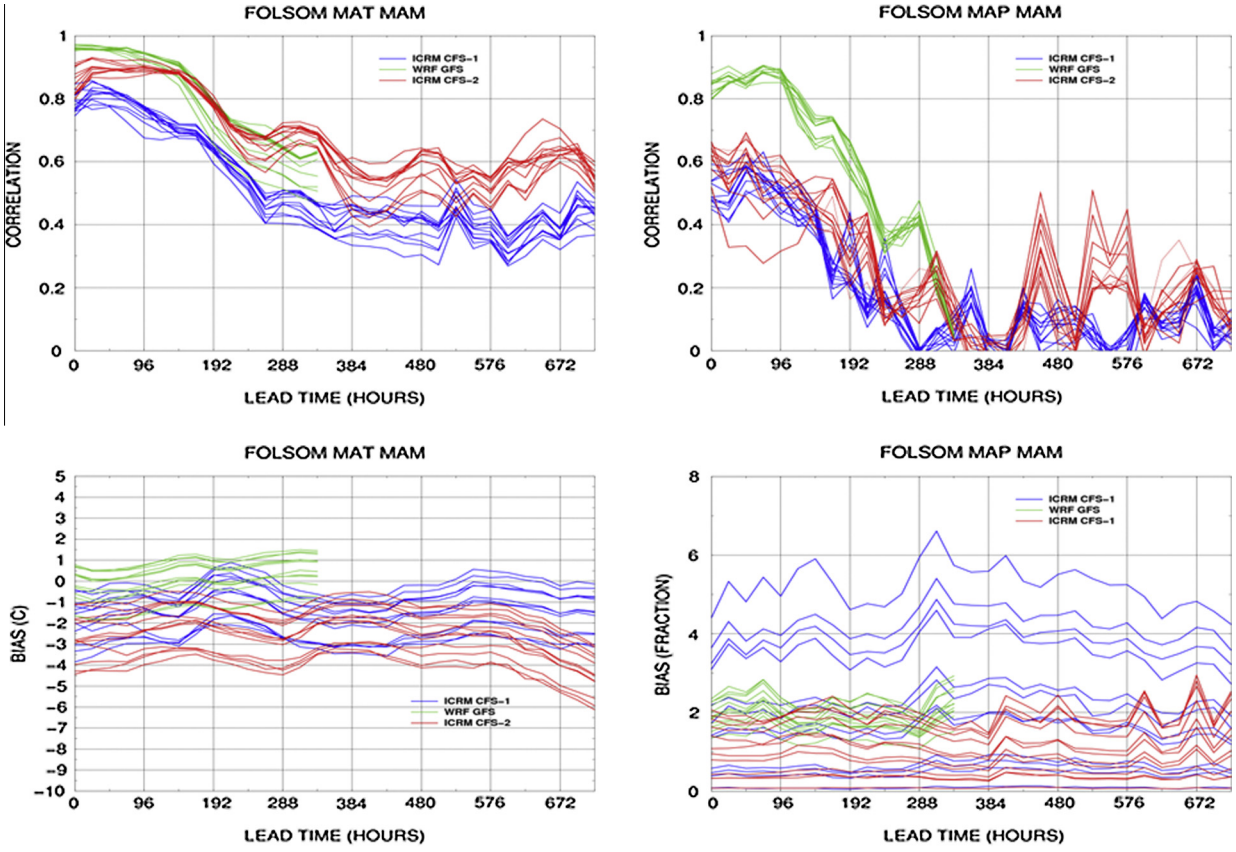

Fig. 8. Bias and correlations of INFORM RTFS forecasts with respect to observations for MAT and MAP in Folsom subcatchments. The figure panels are for bias (fraction for MAP and difference for MAT), and for cross-correlation of daily accumulated precipitation and daily average temperature for the MAM season. Left panels are for MAT and right panels are for MAP. In each panel bias or correlation is shown as a function of forecast lead time, with green lines representing the GFS-WRF forecast results, blue lines representing the CFS1-ICRM forecast results, and red lines representing the CFS2-ICRM results. For each case, multiple lines signify results for various sub-catchments within the Folsom watershed. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

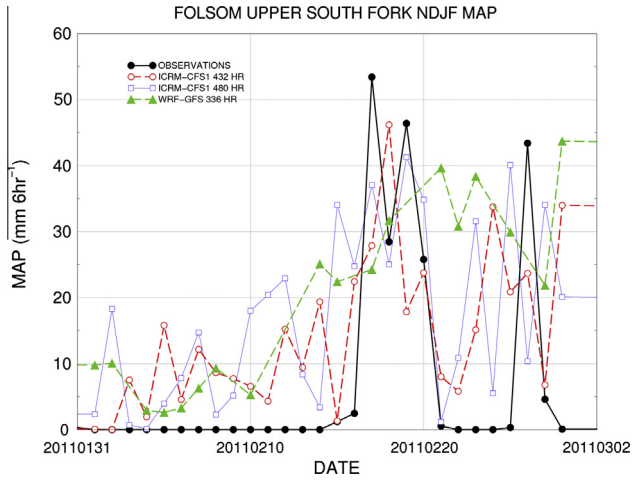

Fig. 9. MAP observations and forecasts for an event in a Folsom subcatchment. The observations are in black line with filled circles, the 432-h-lead ICRM-CFS1 forecast is shown in red while the 480-h-lead ICRM-CFS1 forecast is shown in blue, and the WRF 3336-h-lead forecast is shown with green. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

times (except for some long lead times for a few subcatchments). Like the ICRM-CFS1, these forecasts show <1 bias fraction in subcatchments with low terrain slope. The WRF and ICRM-CFS2 results exhibit weak dependence of bias with lead times, a point discussed later in relation to bias adjustment.

The MAT forecast correlations are high for all cases, with very high values ($\geqslant$0.80) for WRF and ICRM-CFS2 out to about 8 days, while ICRM-CFS2 maintains correlations of 0.6 for several subcatchments out to lead times greater than 25 days. ICRM-CFS1 also exhibits high correlations, but these remain lower than those of both other forecast models for all lead times and catchments. In terms of bias, the WRF model forecasts have lower biases than the other two forecast models that exhibit a cold bias for most subcatchments and lead times. As with MAP it is found that MAT depends only weakly on lead time (up to the maximum lead time of 16 days for GFS and up to about 25 days or so for ICRM). The high correlations for ICRM MAT at long lead times support the potential use of the ensemble forecasts for the prediction of melt out to time scales of several weeks and underscore the potential use of high resolution dynamical ensemble forecasts for the INFORM region and the need for the requisite boundary data from the large-scale long-lead forecast NCEP models.

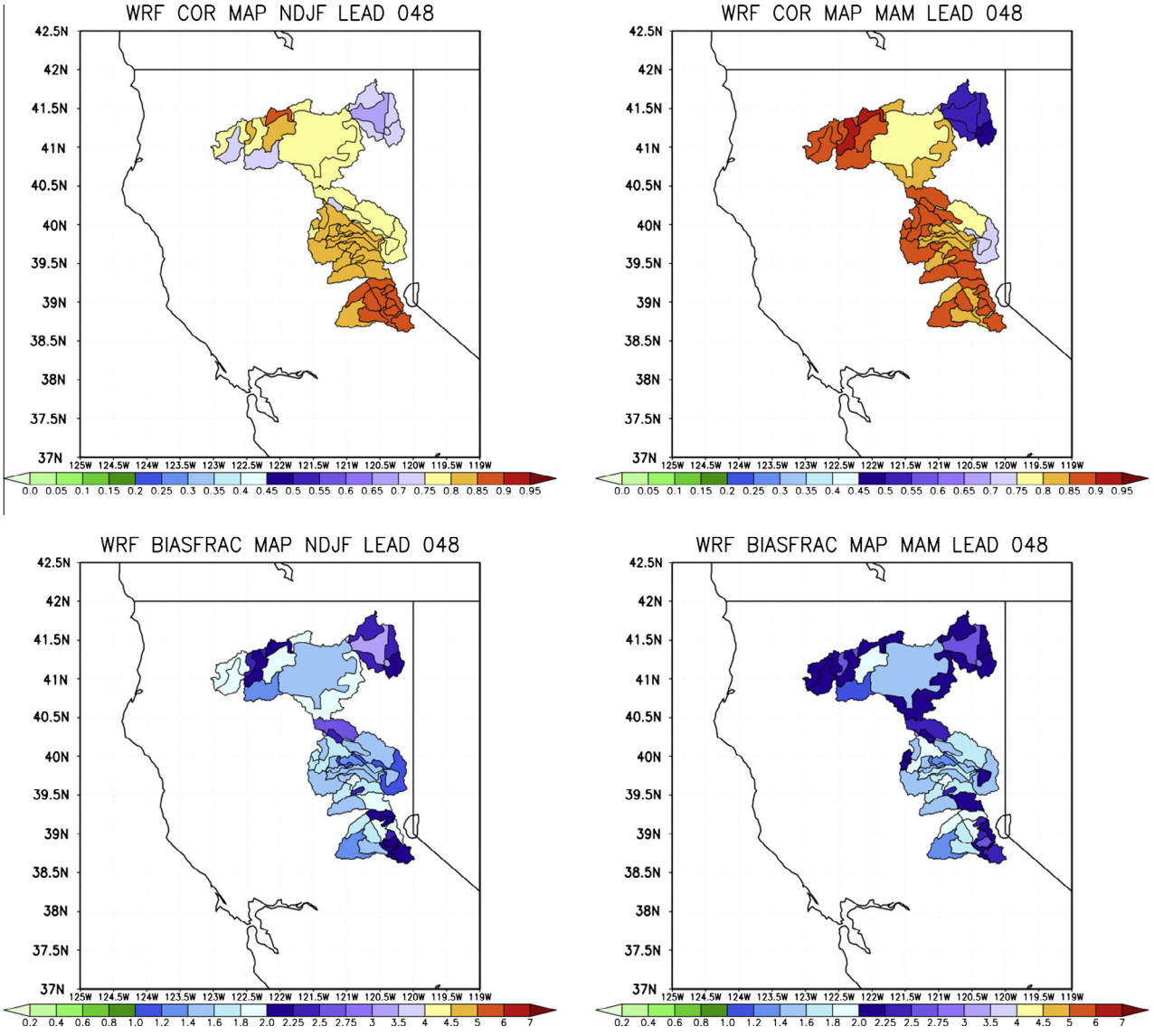

Fig. 10. Subcatchment distribution of MAP correlation and bias fraction for the WRF model.

As noted in HRC-GWRI (2013), the results for the rest of the watersheds of interest and for the same (MAM) season, are qualitatively similar with those discussed for the Folsom watershed. In all cases, the WRF model ensemble MAP forecasts have consistently high correlations for lead times less than about 5 days and outperform the ICRM model ensemble forecasts in terms of correlation for these lead times. The WRF MAP results do show a systematic positive bias (bias fraction ~2) for these lead times. Short lead times for ICRM-CFS2 carry MAP-forecast correlations near or above 0.6 except for a few cases where the highest correlation reaches 0.5. For lead times 16–20 days, the MAP-forecast correlation for the ICRM-CFS2 (and the ICRM-CFS1) exhibits anomalously high values that consistently reach or exceed 0.4 in all cases. Biases for ICRM MAP are <1 for several of the subcatchments for all watersheds.

The strength of ICRM-CFS2 is exhibited in the MAT ensemble forecast predictions that maintain high correlation to observations for long lead times during this snow melt season (MAM). For most subcatchments, there is a cold bias for ICRM forecasts. The ICRM-CFS1 model has lower skill for all lead times but maintains correlations $\geqslant 0.4$ even for lead times greater than 15 days during this MAM season. It exhibits less of a cold bias than the ICRM-CFS2 for most subcatchments.

Results have also been obtained for the NDJF season (snow accumulation season for high elevations). There are no ICRM-CFS2 results for this season and the results pertain only to the WRF and ICRM-CFS1. These results are not shown for previty but are briefly summarized below. The MAP results show consistently high correlations of the WRF model forecasts for lead times up to a few days with correlations dropping rapidly at longer lead times and with a bias ratio that is 2 or higher. It is notable that the MAM season exhibits substantially longer persistence of high correlations than the NDJF season for the WRF and to a lesser degree for the ICRM-CFS1 for MAP forecasts. The ICRM-CFS1 shows substantially lower correlations for MAP than WRF, and has a milder slope of decline with lead time. Biases for this system vary significantly among watersheds. In this season and for this ICRM model configuration too, there is high correlation in long lead times suggesting skill in the window between 16 and 20 days. The MAT correlations are somewhat lower in the NDJF season than in the MAM season for both WRF and ICRM-CFS1. Both models indicate broadening of the subcatchment differences in MAT correlation in the NDJF season with biases ranging from −2 °C to +2 °C for WRF and 1 °C or 3 °C for ICRM-CFS1.

The notably high correlations between ICRM MAP forecasts and observations for lead times 15–20 days are a consistent feature in all basins during the MAM season in Fig. 8; this raises the question whether these higher values are due to statistical behavior or represent real skill. Although this cannot be answered unambiguously with the present limited data, one can test whether there is consistent behavior of the models for these long lead times when there is an event. If there is such consistent behavior, then one cannot dismiss the possibility of real forecast skill for these lead times when there is a significant event. An example of a mid-February 2011 event apparently forecast at long lead times is depicted in Fig. 9.

Fig. 9 shows the MAP observations and ICRM-CFS1 forecasts (both lead times of 432 h or 18 days, and 480 h or 20 days) and the WRF forecast with a lead time of 14 days. In all forecast cases, there is evidence of skill in predicting the timing of this mid February 2011 event, with sharper predictions for the ICRM-CFS1 with a lead time of 18 days. This consistency in forecast performance between the same model and different lead times and between different models supports the conjecture that there is skill at these lead times, at least for certain events. Inasmuch as the GFS system does not use dynamical sea surface temperature forecasts, it is conjectured that this example of skill seen in both GFS and CFS1-driven long lead forecasts is due to slow internal atmospheric process. Additional data and analyses are necessary to confirm this conjecture but it is in line with the findings of Georgakakos et al. (2010a,b).

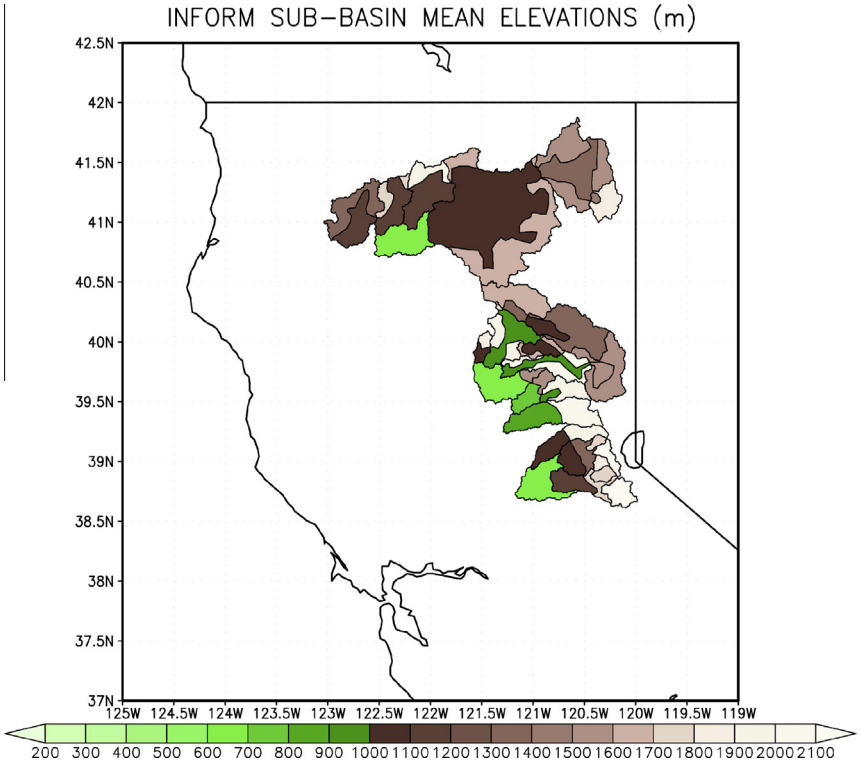

Fig. 11. Subcatchment average elevations.

The geographical distribution of the bias and correlations for the WRF model MAP for a 48-h lead time and for both seasons is shown in Fig. 10. These are representative of most results and a complete set of plots is in HRC-GWRI (2013, Appendix E) for all three models (WRF, ICRM-CFS1, and ICRM-CFS2) and for both MAP and MAT. Fig. 11 shows the subcatchment average elevations for easy reference.

A strikingly non-uniform spatial distribution in WRF MAP correlation between forecasts and observations is apparent in Fig. 10. For the lead time of 48 h, the MAM season shows higher correlations than the NDJF season for most basins apart from the northeastern subcatchments of the Shasta watershed and some of the southern subcatchments of the Folsom watershed, which show lower correlation in MAM. Relatively low correlations are shown for both seasons for large subcatchments of the Shasta watershed and certain subcatchments of the Oroville watershed in higher elevations. Referring to the map of catchment elevations (Fig. 11) shows that all catchments except those in northeastern Shasta drainage have correlations in excess of 0.7. This is important for the higher elevation subcatchments (e.g., high elevations of Yuba and Folsom watersheds and Shasta watersheds of higher elevation) where snow accumulation and melt are significant processes.

High MAP bias fractions are prevalent for most subcatchments for both seasons, with higher biases generally shown for higher elevations (sparse observations over the high terrain may contribute to this tendency). Variability among subcatchments is substantial in both seasons, but with NDJF exhibiting (generally) lower bias fraction for the subcatchments.

The geographically distributed results for the ICRM-CFS1 and ICRM-CFS2 models (not shown) for MAP forecasts show substantial spatial variability in the MAM season, with low elevation basins generally showing lower correlations and high elevation basins showing higher correlations; this is to be expected because ICRM uses an orographic precipitation model that depends on slope for the generation of moist updrafts. For most subcatchments, ICRM-CFS2 performs better with higher MAP correlations and fractional bias closer to 1. For the case of NDJF and ICRM-CFS1, notable increases in correlation are shown in comparison to the MAM season and the same model. Biases are similar between the two seasons except for the high elevation (>1400 m) northeastern subcatchments of Shasta watershed where the bias fraction for the NDJF season is much higher (significant MAP over-estimation) than that of the MAM season.

The MAT forecast bias and correlation subcatchment plots for both seasons and for all three models (not shown) show very high

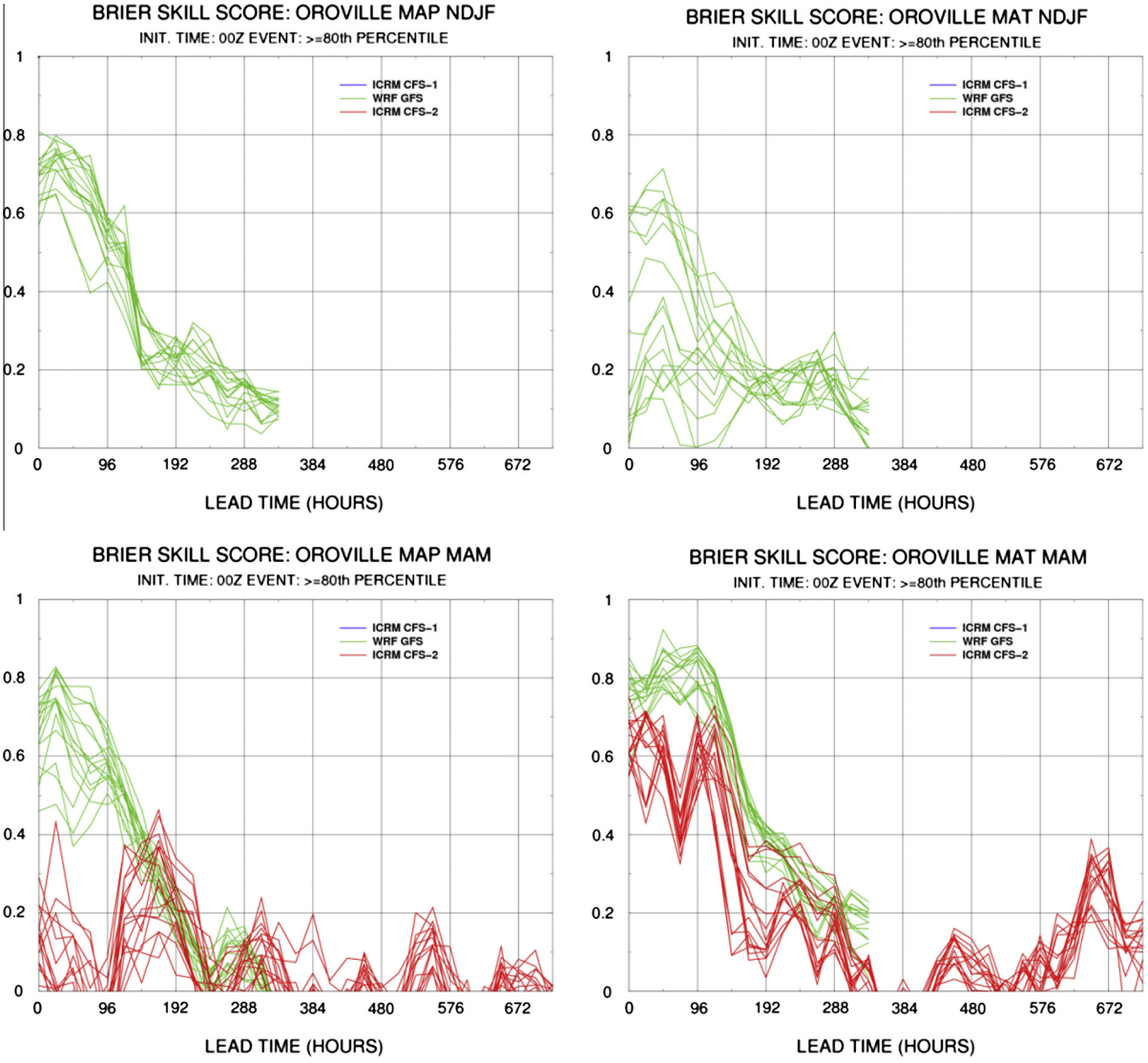

Fig. 12. Brier Skill Score for MAP and MAT in Oroville watershed. Only positive values of the BSS are shown to evaluate the dynamical-model forecasts of INFORM.

correlations over all subcatchments in the MAM season by the WRF while higher spatial variability in correlation is seen for the NDJF season. Bias is generally low for this lead time with cold bias for the higher elevations, more so in NDJF than in MAM. For the MAM season correlations for the ICRM-CFS1 and ICRM-CFS2 are higher in the southern subcatchments than in the northern subcatchments (especially so for the ICRM-CFS1 forecasts). Generally less uniform results for correlations have been obtained with ICRM than with the WRF for the 48-h lead time. Cold bias is exhibited by both versions of ICRM for the MAM season particularly for some lower elevation catchments. For the NDJF season, the ICRM-CFS1 MAT forecasts exhibit lower correlations than for the MAM season, and these NDJF results have a rather uniform warm bias even for the higher-elevation subcatchments, apart from the Trinity watershed where they have a cold bias.

3.3.2. Probabilistic performance measures

The Brier Skill Score (BSS) has been computed for both seasons and for both the models. As noted earlier, the ICRM-CFS2 model results are only available for the MAM season and for the single year 2012. Year 2012 did not contain an adequate number of precipitation events to allow stable statistics for the climatological frequencies conditional on a forecast frequency interval that are necessary for the BSS. Thus, the emphasis of this discussion is on the WRF MAP and MAT results that include two years of operation (2011 and 2012) for both, and on the ICRM-CFS2 MAT results.

Fig. 12 shows the WRF MAP and MAT BSS for multiple lead times and for all the subcatchments of the Oroville watershed and for both seasons NDJF and MAM. Positive BSS is indicated throughout the WRF forecast lead time range (maximum lead time of 16 days), with MAP forecasts for NDJF indicating higher BSS than those in MAM. Opposite seasonal performance tendencies appear for MAT forecasts, with substantial degradation of MAT forecast reliability for the NDJF season. Large differences between subcatchments are shown in MAT reliability for the NDJF season and with lower range of differences in MAP reliability for the MAM season. MAP ensemble forecast BSS values are above 0.2 even out to 8 days indicating some skill, while very good reliability is exhibited by the MAT ensemble forecasts of WRF out to 4 days lead time (BSS near or above 0.8 for all subcatchments of Oroville watershed). It is anticipated that these MAT results derive from the WRF NOAH land surface component formulation, which has only recently been improved (e.g., Wang et al., 2010).

The BSS for the MAM season and for the ICRM-CFS2 MAT ensemble forecasts (lower right panel of Fig. 12) shows positive values out to 16 days then increasing again beyond 20 days,

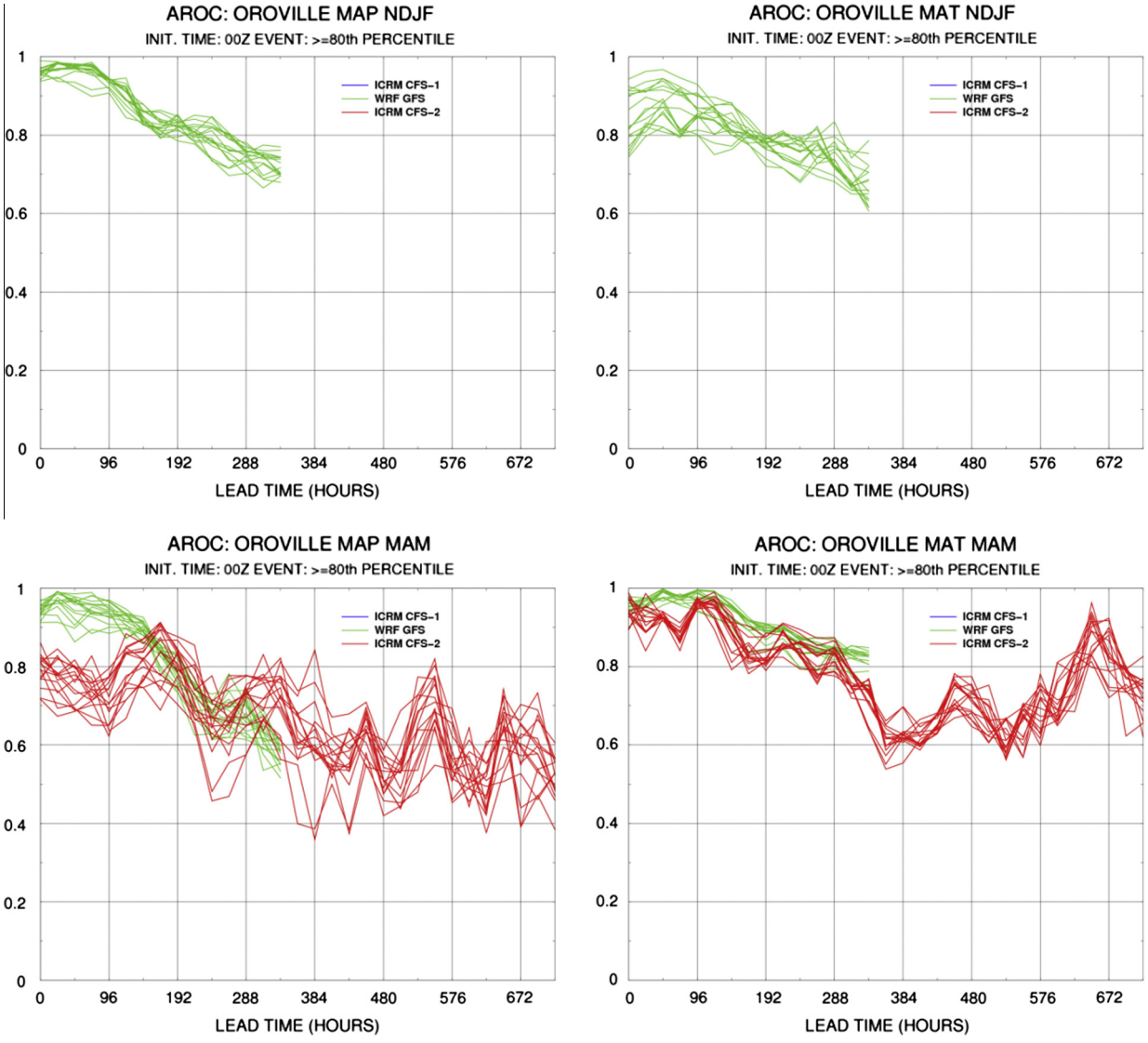

Fig. 13. AROC values of MAP and MAT ensemble forecasts for the Oroville watershed subcatchments.

possibly reflecting seasonal temperature trends. The MAT BSS for ICRM-CFS2 is lower than that of WRF for this season for most lead times and subcatchments, reaching maximum values of about 0.7 for most subcatchments. Overall, Fig. 12 reveals good reliability of the INFORM RTFS ensemble forecasts with respect to climatological forecasts of both MAP and MAT out to about 16 days.

Comments analogous to those made for the Oroville watershed above apply to the other watersheds as well (not shown here but can be found in HRC-GWRI, 2013). Notable exception is the poor reliability performance of the MAT ensemble forecasts for a few catchments of the Folsom watersheds, for short lead times and for the NDJF season. Estimation of the MAT observed values for that season generally carries significant uncertainty, especially for high elevation areas.

Fig. 13 shows the AROC index for MAP and MAT multi-lead forecasts as a function of lead time for all the subcatchments of the Oroville watershed and for the WRF and ICRM-CFS2 models and NDJF and MAM seasons (ICRM-CFS2 only available for the MAM season).

The results of Fig. 13 show skillful forecasts (>0.5) in all cases and for all leads (<16 days) for the WRF model. For the ICRM-CFS2 model, MAP forecasts have good skill for all but two subcatchments for lead times out to 18 days, and for most subcatchments some skill out to 30 days. The MAT ensemble forecasts from the ICRM-CFS2 model have good skill for all subcatchments out to a 30-day lead time. Generally, for MAM and for shorter lead times (<10 days) for MAP the WRF outperforms the ICRM-CFS2, beyond this the skill is comparable for both models. Comments analogous to those made above for the Oroville watershed apply to the other watersheds as well. Generally, the AROC performance index indicates good skill even out to long lead times for the majority of the watersheds and for both MAP and MAT.

3.4. Reservoir inflow assessments

3.4.1. Biases and correlations

The bias and correlation errors of the reservoir inflows (mean daily flow) for the range of lead times out to a month or so have been examined for the forcing from the WRF and the ICRM-CFS2. In the discussion below we refer to the forcing model to distinguish system inflow forecasts (e.g., WRF indicates MAP and MAT forcing from the GEFS-WRF model is used to drive the snow, soil and routing models). Fig. 14 shows the fractional biases for both WRF and ICRM-CFS2 and for the MAM and NDJF seasons. The biases for the MAM season appear to be weakly varying with forecast lead time (except for Trinity inflows for the WRF model), and range from about 1.2 (Yuba, New Bullards Bar) to about 2.0 (Oroville) for WRF, and from about 0.7 (Shasta) to about 1.8 (Oroville) for

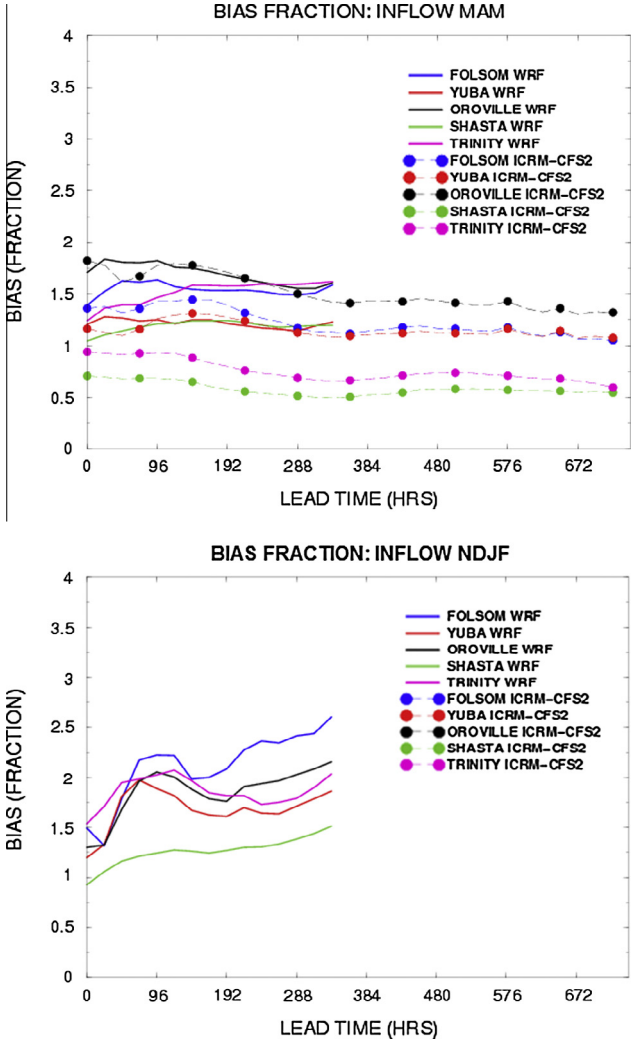

Fig. 14. Bias fraction of reservoir inflows for the MAM season (Upper Panel) and the NDJF season (Lower Panel). No forecasts from ICRM-CFS2 are available for the NDJF season.

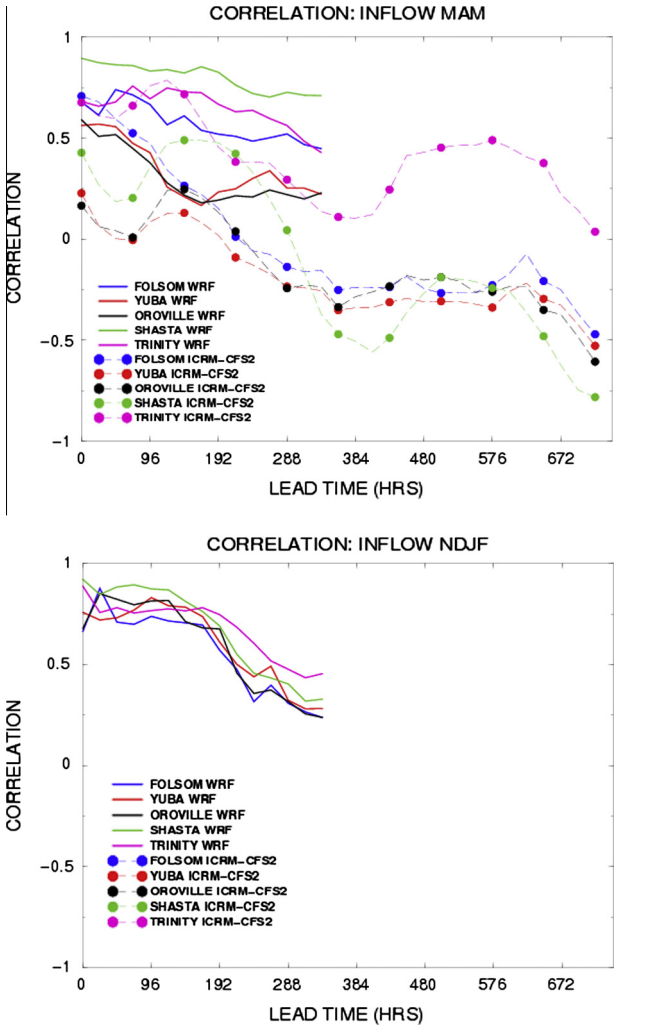

Fig. 15. Correlations of reservoir inflows for the MAM season (Upper Panel) and the NDJF season (Lower Panel).

ICRM-CFS2. The bias difference among inflows to different reservoirs is higher for the ICRM-CFS2 than for the WRF. Lowest bias for WRF is for New Bullards Bar inflows of the Yuba River (1.2) and for ICRM-CFS2 is for Trinity reservoir inflows (0.85), with New Bullards Bar inflows a close second (1.25).

The bias for the NDJF season (for the WRF only) is shown in the lower panel of Fig. 14. There is an increasing trend in the bias fraction for this season for the WRF. Shasta forecast inflow has the lowest bias fractions, near 1 in short lead-times and up to 1.4 at the 20-day lead time, while Folsom forecast inflows have the highest bias fraction for lead times longer than 3 days (2.2 to greater than 2.5). New Bullards Bar inflows on the Yuba River have the highest bias for lead time shorter than 3 days (values in the range 1.5–2). The accumulation season NDJF has higher biases than the melt season MAM.

The correlation for the MAM and NDJF seasons between model forecasts and observations of mean daily FNF inflows is shown in Fig. 15 for both WRF and ICRM-CFS2. These show that for the WRF model and for MAM, Folsom, Trinity and Shasta inflows have the highest correlations, remaining greater than 0.5 even out to 12 days, maintaining values greater than 0.7 out to a 4-day lead time. The correlations for New Bullards Bar forecast inflows on the Yuba River and for Oroville forecast inflows exhibit lower correlations with values greater than 0.5 maintained only up to a 2-day lead time, dropping precipitously after that to levels of about 0.2.

In comparison to WRF, significantly lower correlations in reservoir forecast inflows are exhibited by the ICRM-CFS2 model and for MAM, except for the Trinity basin for which values of 0.5 are shown for lead times out to more than 20 days. For short lead times ($\leqslant$2 days), Trinity and Folsom forecast inflow correlations range from $\sim$0.5 to 0.75. At longer lead times, the Folsom forecast inflow correlations fall substantially. The behavior of New Bullards Bar, Oroville and Shasta inflows shows an initial dip and then an increase to fall again to negligible levels. For Shasta, the increase brings the correlations to the level of 0.5 for a lead time up to 10 days, while for the Oroville and New Bullards Bar inflows the correlations are less than 0.3 in all cases.

For the WRF and for the NDJF season, the correlations remain above 0.5 (Fig. 15, Lower Panel) for lead times out to 10 days or so. On average and for lead times out to 8 days, Shasta reservoir forecast inflows exhibit the highest correlations and Folsom reservoir forecast inflows the lowest.

For the interpretation of the results presented, one must consider the development of errors in the integrated forecast system of INFORM, which consists of the large-scale NCEP model forecasts (GFS and CFS), the downscaling model forecasts (WRF and ICRM), and the snow-soil-routing model inflow forecasts. Thus, the

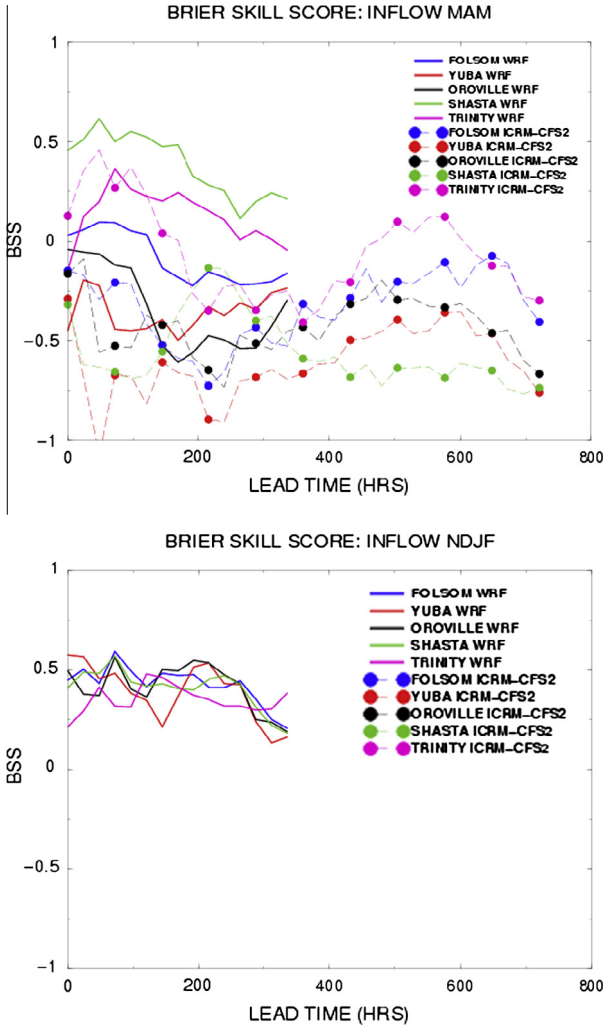

Fig. 16. BSS for reservoir inflows for the MAM season (Upper Panel) and for the NDJF season (Lower Panel).

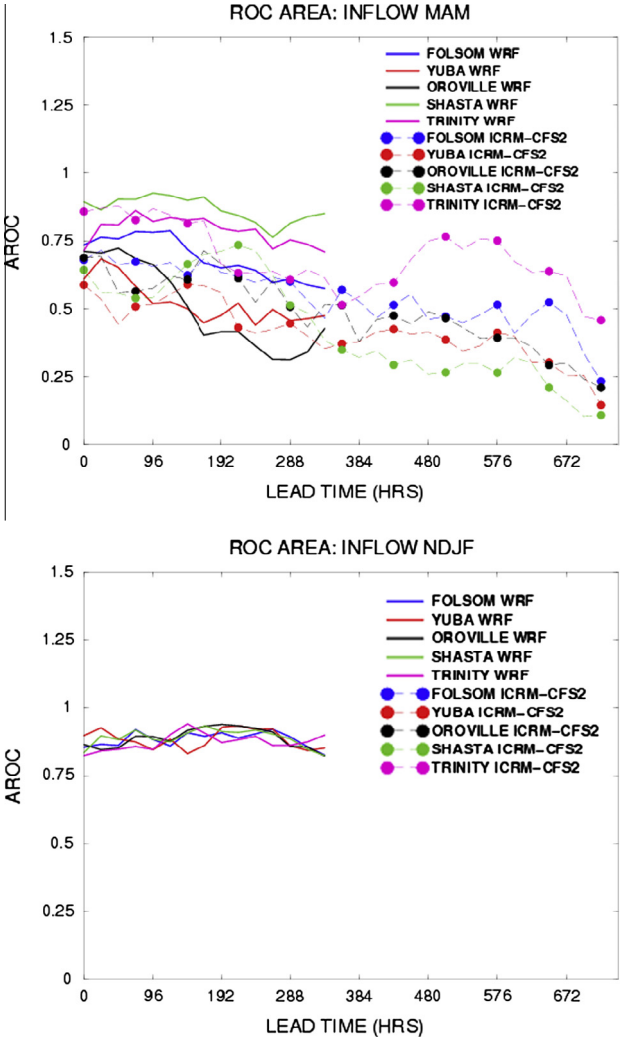

Fig. 17. AROC for reservoir inflows for the MAM season (Upper Panel) and for the NDJF season (Lower Panel).

development of errors in the inflow forecasts is a combination of MAP and MAT errors for the subcatchments of the watershed of interest, initial condition errors in the snow-soil-routing model for these subcatchments and river segments, and the hydrologic model parameter errors.

3.4.2. Probabilistic performance of reservoir inflows

Fig. 16 shows the BSS values for multi-lead reservoir inflow forecasts for the MAM and NDJF seasons from the WRF and ICRM-CFS2 models. The event used is that the flow is greater than 80% of climatological flows. For the NDJF season and WRF, values range from ~0.2 (lowest for the Trinity reservoir) to ~0.6 (Folsom reservoir), with BSS values maintained at about the 0.5 level out to lead times of about 7 days. Folsom inflows exhibit the largest skill fluctuations with lead time for this season. Analogous results for the MAM season for the WRF show low skill for all watershed inflows except Trinity and Folsom for which it is moderate (0.2 to more than 0.5) for lead times out to 5 days or so. Folsom inflow forecasts have positive skill for short lead times, while Yuba and Oroville inflow forecasts essentially show no skill for this season. It is likely that the BSS values are low in part because the number of inflow events available for the validation is low. The BSS score is unstable when the number of events used to compute the score is low. For example, for WRF, just two seasons of events were used to compute the BSS performance index, and for the high threshold of 80% the number of events for come deciles was very low.

The AROC (area below the relative operating characteristic curve) values for the MAM and NDJF seasons are shown in Fig. 17, respectively, for all reservoir inflows. Significant skill is shown for AROC throughout the range of lead times for WRF driven inflow forecasts and for the NDJF season (values above 0.8 for all cases). Lower skills are exhibited for the MAM season but with good performance for Folsom, Trinity and Shasta inflows for all lead times out to 16 days, and for Yuba (New Bullards Bar) and Oroville inflow forecasts for lead times out to at least 4 days. Greater differences among watersheds exist in the MAM season than in the NDJF season, presumably because of the presence of snow in the upper regions of watersheds and dependence of inflow forecasts on both MAP and MAT forecasts in those regions. The ICRM-CFS2 has AROC greater than 0.5 only for short lead times for all reservoir inflow forecasts, and for those of the Trinity reservoir exhibiting skill out to more than 20 days.

3.5. MAP and MAT bias adjustment impacts

Data preparation procedures were essentially identical for MAP and MAT and are summarized in Appendix B. The procedures produce separate forecast calibration factors (CMAP and CMAT) for MAP and MAT, specific to a given season and sub-catchment by comparing forecast values and observations.

The results for the correlation and bias performance indices are exemplified in Fig. 18 for the two models, the MAM melting season

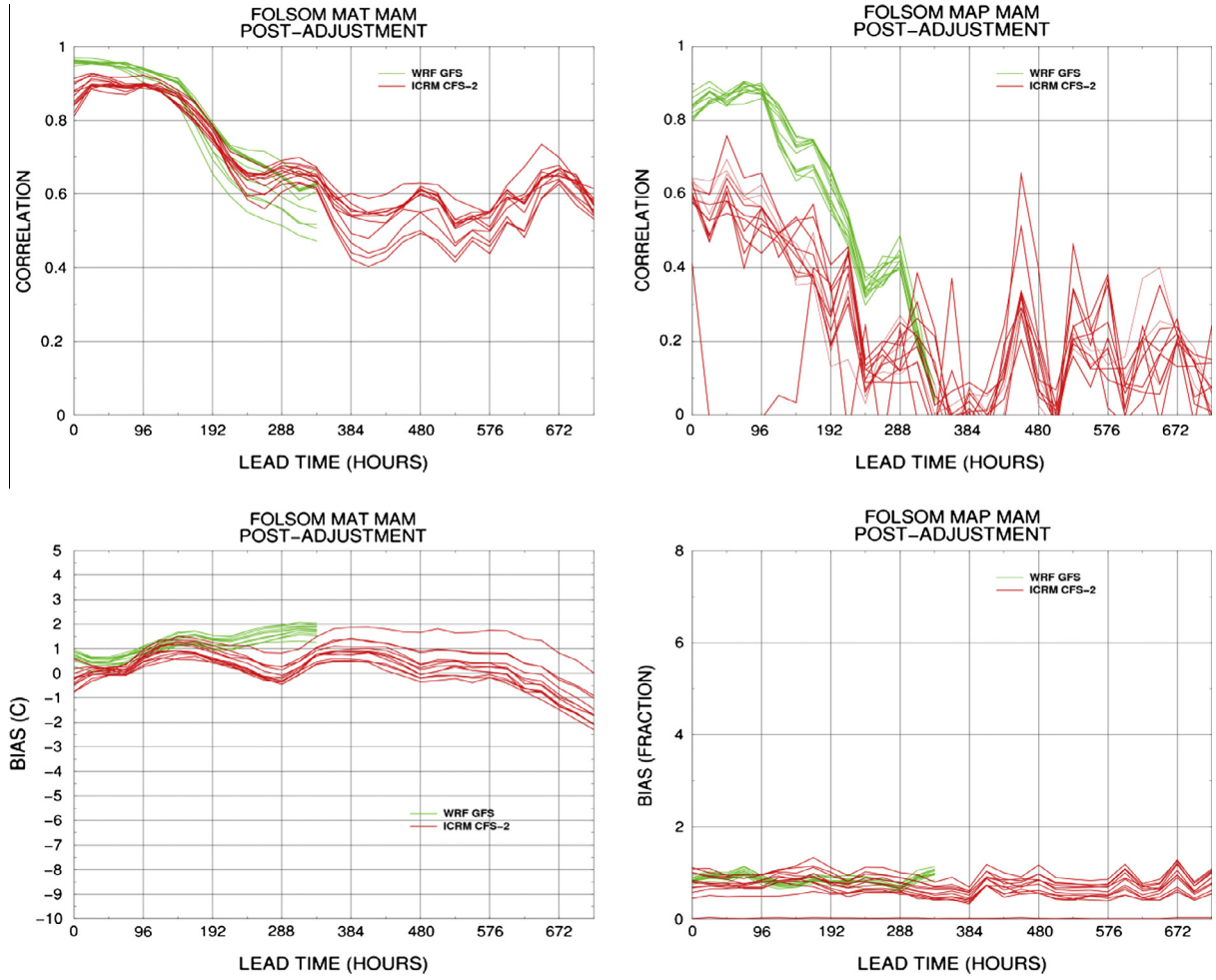

Fig. 18. Bias and correlations of INFORM RTFS forecasts with respect to observations for MAT and MAP in Folsom subcatchments – post bias adjustment.

and for the Folsom Reservoir contributing subcatchments. In all cases the post-adjustment results show substantially lower bias than the original results (Fig. 8) and correlations improved or stayed the same in all but for one subcatchment for MAP and for shorter lead times. Detailed results are included in HRC-GWRI (2013, Appendix F) for all the cases and lead to a similar conclusion.

Results pertaining to reliability for the post bias-adjusted results have also been obtained for all cases available in order to examine the impact of the bias adjustment methodology. No significant changes were observed with respect to the original results (without bias adjustment as shown for example in Figs. 12 and 13). The overall conclusion for the validation of the MAT and MAP bias adjustment procedure is that the probabilistic methodology used provides useful results as it corrects the bias well for most lead times and leaves essentially unaffected the performance of the models with respect to cross-correlations and probabilistic measures.

The question is whether such bias adjustments are useful for the reservoir inflow forecasts. Note that these inflow forecasts depend on the evolution of both MAP and MAT, and on their concurrent variability, and that adjustments to those will change not only liquid precipitation and evapotranspiration but snow accumulation and melt as well. The latter affects reservoir inflows during the spring melt season.

Figs. 19 and 20 present the reservoir inflow performance metrics of bias fraction and cross-correlation with observations for all reservoirs, all seasons and both cases of hydrologic-model forcing (WRF and ICRM-CFS2). The bias results indicate substantial improvement with respect to the original reservoir inflow metrics of Figs. 14 and 15 for most basins, though the Trinity reservoir inflow forecasts maintain high bias for NDJF and for the WRF at all lead times.

The cross-correlations in Fig. 20 show slightly modified results than those originally obtained (without bias adjustment in Fig. 15) except from the ICRM-CFS2 forced hydrologic model forecasts for MAM that have improved (higher correlation values) substantially. The short lead time forecasts of the WRF-forced hydrologic model using the adjusted MAP and MAT exhibit lower correlations with the observations than those using the unadjusted forecast data. This result manifests the sensitivity of the reservoir inflow forecasts to simultaneous changes on the MAP and MAT forcing.

Probabilistic assessments were made for the reservoir inflow forecast ensembles after bias adjustment of the MAP and MAT as well. The results are basically consistent with those discussed above and indicate moderate improvements with respect to the original results. Thus, the most important impact of adjusting the MAP and MAT ensemble forecasts for bias is the improvement in the bias of the ensemble reservoir inflow forecasts.

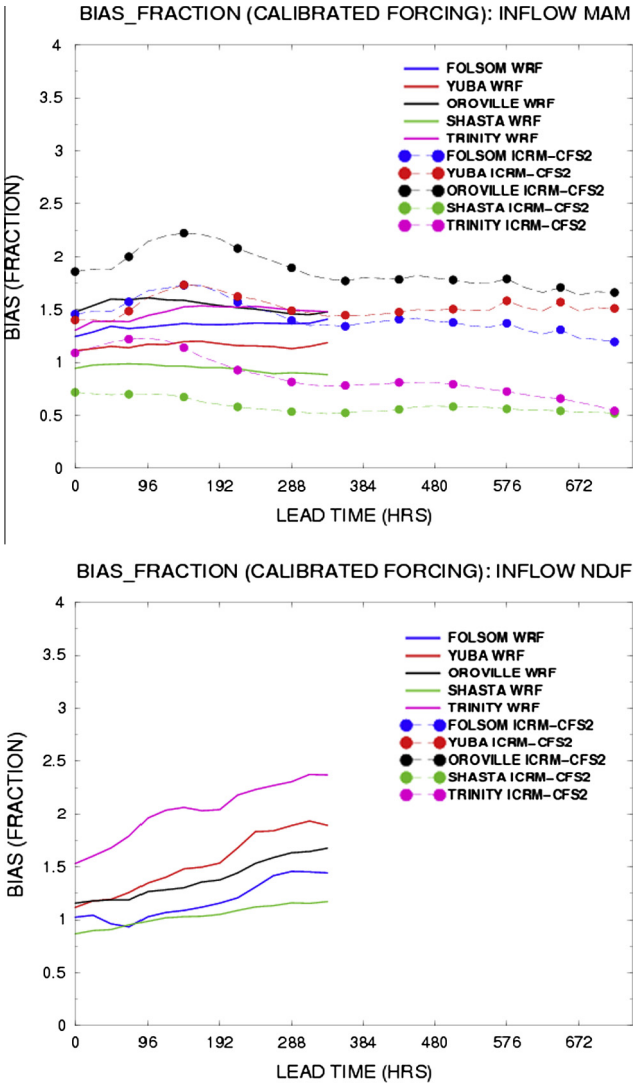

Fig. 19. Bias fraction of reservoir inflows for the MAM (Upper Panel) and NDJF (Lower Panel) seasons – post bias adjustment.

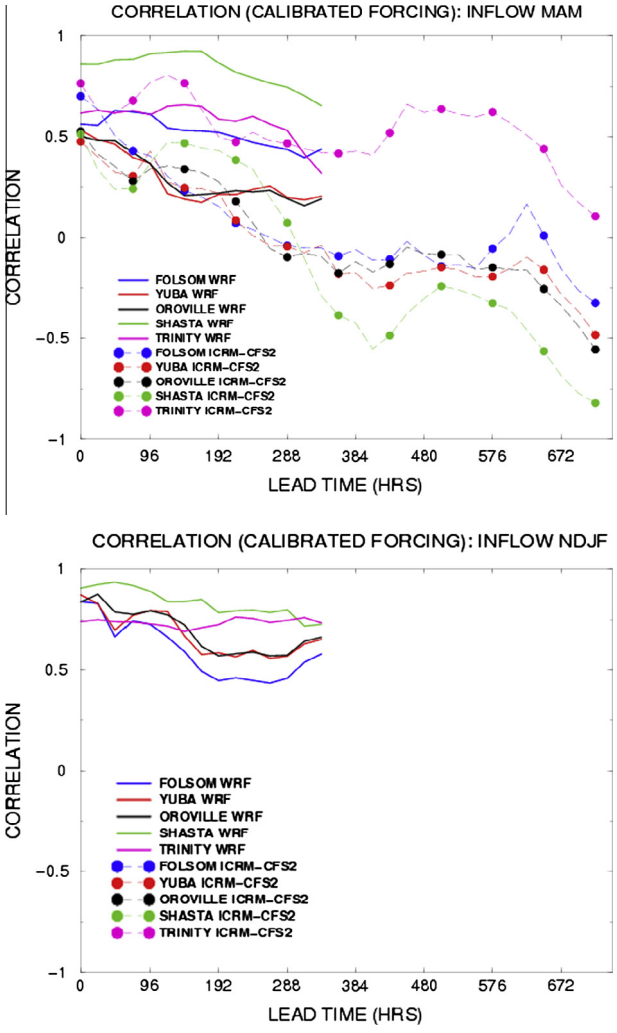

Fig. 20. Correlations of reservoir inflows for the MAM and NDJF seasons – post bias adjustment.

4. Conclusions and recommendations

The main conclusions from these analyses are discussed below:

(1) Impressive skill was found for MAP and MAT from the WRF forced by GEFS ensemble forecasts and to a lesser degree for the ICRM forced by CFS2 ensemble forecasts for the tributary subcatchments of the INFORM domain in Northern California. The results show that these models can be used to provide useful forecasts for the INFORM subcatchments in real time out to lead times of two weeks both for the snow accumulation (NDJF) and the melt (MAM) periods.
(2) Good forecast skill was found for reservoir ensemble full natural inflows generated by the WRF-GEFS forcing and the hydrologic models of INFORM with lead times out to several days (4–6 days).
(3) Steady performance-index values throughout the range of lead times provide clear evidence that bias corrections for the WRF model for MAP and MAT will be effective and provide improved flow forecasts. Improved reservoir inflow forecasts resulted for most reservoir inflows after bias correction of the WRF-forecast MAT and MAP.
(4) Useful MAT forecast skill is exhibited by the ICRM-CFS2 model for lead times longer than a week, and in some cases out to four weeks in the MAM season. Additionally, there is evidence suggesting long-lead skill for the MAP forecasts of the ICRM-CFS2 (and ICRM-CFS1) for large events in some mountainous watersheds. Further work and longer data sets are required to demonstrate that this skill is real and can result in useful forecast information.
(5) Bias correction of the ICRM-CFS2 produced MAP and MAT forecasts with a fixed bias correction factor improved the performance statistics of these forecast variables substantially for almost all lead times. Reservoir inflow benefits on the other hand were realized in the first forecast week for most basins, except for the Trinity basin for which correlations of forecasts to observations remain above 0.5 out to four weeks lead time with bias fractions that are in the range from 0.7 to 1.3.

The implication of the performance evaluation conclusions for reservoir management is that the INFORM real-time forecasts provide skillful quantitative information at least for the first week of forecast lead times and in some cases much beyond that. They thus should be of direct benefit at least to reservoir flood control and energy production objectives. The analysis of the benefits to reservoir management, using metrics pertaining to management objectives, is on-going (see initial evaluations in HRC-GWRI (2013)) and will be reported elsewhere.

An important recommendation is to periodically re-evaluate the performance of the INFORM real-time forecast component (e.g., once 2–3 seasons have past) to provide additional information for reservoir management. The findings also suggest further evaluation of the simultaneous bias correction of the MAP and MAT values and their implications for multi-lead reservoir inflow forecasts during the melt season. This should be done in conjunction of a recalibration of the operational hydrologic models to reduce simulation biases in some of the basins. The use of reforecasts from a frozen version of the operational large scale NCEP models appears fruitful for sharpening the bias correction techniques used.

The current evaluation of the reservoir inflow forecasts was based on unimpaired flows estimated from observations at regulation sites. Recent work has produced and tested successfully methodologies for accounting for upstream regulation effects on ensemble forecasts and allows direct comparison of the adjusted forecasts to streamflow observations (Georgakakos et al., 2012a). Implementation of these methodologies is recommended as part of the INFORM forecast component for improved (and more useful) evaluations of forecast-system performance for reservoir management. Lastly, and based on the results of the present work, the analysis and examination of utility of the long-lead (out to a month) predictability of significant extreme events in Northern California with the WRF and ICRM is warranted.

Appendix A. Solution method for T_o and 2-m air temperature

The numerical solution for the surface temperature T_0 and the 2-meter air temperature T_a over a grid in mountainous terrain is obtained for a single time period with the application of the following steps:

- Heat flux into the soil at the surface is set (assumed) equal to zero.
- Constant wind profiles are assumed in the mixed layer of the boundary layer and a power law profile in the surface layer of the boundary layer (last 100 m near the surface).
- Using conservation of potential temperature and mixing ratio 100 levels of T, p and RH are estimated based on the CFS basic levels (see Tables 2 and 3 for CFS1 and CFS2, respectively) using vertical linear interpolation.
- Using the same conservation assumptions for potential temperature and mixing ratio, we then interpolate along the horizontal from the CFS grid to the ICRM grid using inverse square distance interpolation for all the levels and all the ICRM grid nodes (two upstream grid points are currently used from CFS). Estimates of T, p, and RH are thus obtained in a three dimensional grid for ICRM.
- We apply adiabatic and pseudo-adiabatic adjustment of the T, p, and RH values at the 2 m reference level near the surface (parcel ascent given conditions conducive to parcel ascent). The T_1 values obtained are the initial estimates of the reference air temperature that enters the ICRM computations. A sinusoidal curve is fitted to the temperature solution at this point to interpolate from 12 hourly to 6-hourly values (the CFS input is in 12-hourly intervals).
- With 6-hourly estimates of T_1 at hand we apply the surface energy balance and we solve for the surface temperature T_0 with a 6-h resolution.
- An updated estimate of the air surface temperature is a weighted average of T_1 and T_0: $T_a = T_s\, w_1 + T_1\, w_2$ with w_1 and w_2 being weights that sum to 1.
- The procedure is repeated with T_a used in place of T_1 until there is convergence within a set tolerance.

It is noted that when there is snow on the ground and the surface (skin) temperature solution is higher than 0 °C, the surface temperature solution is constrained to be at 0 °C and the residual heat flux (resulting from unbalanced terms in Eq. (1)) is devoted to melting snow.

Appendix B. Calibration factors for MAP and MAT adjustment

The procedures for the computation of the calibration factors for MAP and MAT adjustment are given below.

B.1. Data preparation

This procedure provides observations paired with complete sets of ensemble 72-h-lead forecasts validating at the observation time.

(1) Collate pairs of observed and 72-h-lead forecast 6-hourly MAP (or MAT). This was done combining all validation times into single data sets (00 and 12 GMT for WRF; 00, 06, 12, and 18 GMT for ICRM-CFS2).
(2) Identify all collated pairs belonging to cases for which all simulation ensemble members are present, discard other cases.
(3) Stratify by season MAM and NDJF.

B.2. Derivation of MAP calibration factors

- For each season NDJF and MAM.
  (a) Sort the model ensemble 6-hourly 72-h forecasts from the available cases into 10 quantiles, keeping each forecast value paired with its corresponding validating observation. Note that each observed value will occur NE (the number of ensemble members) times. Use only cases where simulated ensemble values are >1 mm 6-h^{-1}.
  (b) For each quantile, calculate the mean (F_M) of the forecast values.
  (c) For each quantile, calculate the mean, and 30th and 70th percentile values of the observed values (O_M, O_{30}, O_{70}) *paired with the forecast values in that quantile*.
  (d) For the 3rd to the 8th quantile, the calibration factor (C_{MAP}) is defined as O_M/F_M.
  (e) For 1st and 2nd quantiles, C_{MAP} is defined as O_{30}/F_M.
  (f) For the 9th and 10th quantiles, C_{MAP} is defined as O_{70}/F_M.
- For June–October – set C_{MAP} to 1.0.

B.3. Derivation of MAT calibration factors

- For each season NDJF and MAM.
  (a) Sort the 6-hourly 72-h forecasts from the available cases into 10 quantiles, keeping each simulated value paired with its corresponding validating observation. Note that each observed value will occur NE times.
  (b) For each quantile, calculate the mean (F_M) of the forecast values.
  (c) For each quantile, calculate the mean of the observed values (O_M) paired *with the simulated values in that quantile*.
  (d) The calibration factor (C_{MAT}) is defined as $O_M - F_M$.
- For June–October – C_{MAT} set to 0.0.

B.4. Use of calibration bias factors

- Apply C_{MAP} as a multiplicative adjustment ("scale") to the simulated forecast MAP values for the appropriate season and sub-catchment.
- Apply C_{MAT} as an additive adjustment ("offset") to the simulated forecast MAT values for the appropriate season and sub-catchment.

References

Achleitner, S., Schöber, J., Rinderer, M., Leonhardt, G., Schöberl, F., Kirnbauer, R., Schönlaub, H., 2012. Analyzing the operational performance of the hydrological models in an alpine flood forecasting system. J. Hydrol. 412–413, 90–100.

Anderson, E.A., 1973: National Weather Service river forecast system – snow accumulation and ablation model. NOAA Technical Memorandum NWS HYDRO-17, Office of Hydrology, National Weather Service, NOAA, Silver Spring, MD.

Boucher, M.-A., Tremblay, D., Delorme, L., Perreault, L., Anctil, F., 2012. Hydro-economic assessment of hydrological forecasting systems. J. Hydrol. 416–417, 133–144.

Carpenter, T.M., Georgakakos, K.P., 2001. Assessment of Folsom Lake response to historical and potential future climate scenarios, 1. Forecasting. J. Hydrol. 249, 148–175.

Cloke, H.L., Pappenberger, F., 2009. Ensemble flood forecasting. A review. J. Hydrol. 375, 613–626.

Collischonn, W., Tucci, C.E.M., Clarke, R.T., Chou, S.C., Guilhon, L.G., Cataldi, M., Allasia, D., 2007. Medium-range reservoir inflow predictions based on quantitative precipitation forecasts. J. Hydrol. 344, 112–122.

Dudhia, J., 1993. A nonhydrostatic version of the Penn State/NCAR mesoscale model: validation tests and simulation of an Atlantic cyclone and cloud front. Mon. Weather Rev. 121, 1493–1513.

Georgakakos, K.P., 1986. A generalized stochastic hydrometeorological model for flood and flash flood forecasting, 1. Formulation. Water Resour. Res. 22 (13), 2083–2095.

Georgakakos, K.P., Bras, R., 1982. Real-time, statistically linearized, adaptive flood routing. Water Resour. Res. 18 (3), 513–524.

Georgakakos, K.P., Graham, N.E., 2008. Potential benefits of seasonal inflow prediction uncertainty for reservoir release decisions. J. Appl. Meteorol. Climatol. 47, 1297–1321.

Georgakakos, K.P., Graham, N.E., Georgakakos, A.P., 2000. Can forecasts accrue benefits for reservoir management? The Folsom Lake study. Climate Report 1 (4), 7–10.

Georgakakos, K.P., Graham, N.E., Carpenter, T.M., Georgakakos, A.P., Yao, H., 2005. Integrating climate-hydrology forecasts and multi-objective reservoir management for Northern California. EOS 86 (12), 122–127.

Georgakakos, A.P., Yao, H., Georgakakos, K.P., 2010a. Upstream regulation adjustments to ensemble streamflow prediction. HRC Technical Report No. 7. Hydrologic Research Center, San Diego, California, 76pp. <http://www.hrc-lab.org/projects/projectpdfs/HRC%20Technical%20Report%20No%207.pdf>.

Georgakakos, K.P., Taylor, S.V., Graham, N.E., 2010b. Implications of dynamic downscaling of Climate Forecast System (CFS) information for ensemble streamflow predictions in Northern California. HRC Technical Note 42A. Hydrologic Research Center, San Diego, CA, 19pp. <http://www.hrc-lab.org/projects/projectpdfs/HRC_TN42A_TRACS-20100115.pdf> (15.01.10).

Georgakakos, A.P., Yao, H., Kistenmacher, M., Georgakakos, K.P., Graham, N.E., Cheng, F.-Y., Spencer, C., Shamir, E., 2012a. Value of adaptive water resources management in northern California under climatic variability and change: reservoir management. J. Hydrol. 412–413, 34–46.

Georgakakos, K.P., Graham, N.E., Cheng, F.-Y., Spencer, C., Shamir, E., Georgakakos, A.P., Yao, H., Kistenmacher, M., 2012b. Value of adaptive water resources management in northern California under climatic variability and change: dynamic hydroclimatology. J. Hydrol. 412–413, 47–65.

Green, D.M., Swets, J.A., 1966. Signal Detection and Psychophysics. Peninsula Publishing, Los Altos, OH, pp. 45–52.

Hamill, T.M., Whitaker, J.S., Fiorino, M., Benjamin, S.G., 2011. Global ensemble predictions of 2009s tropical cyclones initialized with an ensemble Kalman filter. Mon. Weather Rev. 139, 668–688.

Hamill, Thomas M., Bates, Gary T., Whitaker, Jeffrey S., Murray, Donald R., Fiorino, Michael., Galarneau, Thomas J., Zhu, Yuejian., Lapenta, William., 2013. NOAA's second-generation global medium-range ensemble reforecast dataset. Bull. Am. Meteor. Soc. 94, 1553–1565.

Hong, S.-Y., Lim, J.-O.J., 2006. The WRF single-moment 6-class microphysics scheme (WSM6). J. Korean Meteor. Soc. 42, 129–151.

HRC-GWRI, 2007. Integrated Forecast and Reservoir Management (INFORM) for Northern California: System Development and Initial Demonstration. California Energy Commission, PIER Energy-Related Environmental Research. CEC-500-2006-109, 244 pp. +9 Appendices. <http://www.energy.ca.gov/pier/project_reports/CEC-500-2006-109.html>.

HRC-GWRI, 2013. Integrated Forecast and Reservoir Management (INFORM): Enhancements and Demonstration Results for Northern California (2008–2012). California Energy Commission, PIER Energy-Related Environmental Research. CEC-500-2014-019, 224 pp. + 6 Appendices (in press). <http://www.energy.ca.gov/publications/displayOneReport.php?pubNum=CEC-500-2014-019>.

Hsu, W.R., Murphy, A.H., 1986. The attributes diagram: a geometrical framework for assessing the quality of probability forecasts. Int. J. Forecasting 2, 285–293.

Janowiak, J.E., Bauer, P., Wang, W., Arkin, P.A., Gottschalck, J., 2010. An evaluation of precipitation forecasts from operational models and reanalyses including precipitation variations associated with MJO activity. Mon. Weather Rev. 138, 4542–4560.

Jolliffe, L.T., 2002. Principal Component Analysis, second ed. Springer-Verlag, New York, p. 167-195, 487pp. + illustrations.

Kain, J.S., 2004. The Kain–Fritsch convective parameterization: an update. J. Appl. Meteor. 43, 170–181. http://dx.doi.org/10.1175/1520-0450(2004)043<0170:TKCPAU>2.0.CO;2.

Kharin, V., Zweirs, F., 2003. On the ROC score of probability forecasts. J. Climate 16, 4145–4150.

Liston, 1995. Local advection of momentum, heat, and moisture during the melt of patchy snow covers. J. Appl. Meteorol. 34, 1705–1715.

Livneh, B., Xia, Y., Mitchell, K.E., Ek, M.B., Lettenmaier, D.P., 2010. Noah LSM snow model diagnostics and enhancements. J. Hydrometeorol. 11, 721–738. http://dx.doi.org/10.1175/2009JHM1174.1.

Mason, S.J., Graham, N.E., 2002. Areas beneath the relative operating characteristics (ROC) and levels (ROL) curves: statistical significance and interpretation. Quart. J. Roy. Meterol. Soc. 128, 2145-1266.

McCollor, D., Stull, R., 2008. Hydrometeorological short-range ensemble forecasts in complex terrain, Part I, Meteorological evaluation. Weather Forecasting 23, 533–556.

Olsson, J., Lindstrom, G., 2008. Evaluation and calibration of operational hydrological ensemble forecasts in Sweden. J. Hydrol. 350, 14–24.
Pandey, G.R., Cayan, D.R., Georgakakos, K.P., 1999. Precipitation structure in the Sierra Nevada of California during winter. J. Geophysic. Res. 104 (D10), 12019–12030.
Pielke, R.A., 1984. Mesoscale Meteorological Modeling. Academic Press, San Diego, CA, 612 pp.
Renner, M., Werner, M.G.F., Rademacher, S., Spokkereef, E., 2009. Verification of ensemble flow forecasts for the Rhine. J. Hydrol. 376, 463–475.
Saha, S.co-authors, 2006. The NCEP climate forecast system. J. Climate 19, 3483–3517.
Saha, S., and co-authors, 2013. The NCEP climate forecast System version 2. J. Climate, (in press). <http://dx.doi.org/10.1175/JCLI-D-12-00823.1>.
Saha, S.co-authors, 2014. The NCEP climate forecast system version 2. J. Climate 27, 2185–2208.
Shamir, E., Carpenter, T.M., Fickenscher, P., Georgakakos, K.P., 2006. Evaluation of the NWS operational hydrologic model for the American river basin. ASCE J. Hydrol. Eng. 11 (5), 392–407.
Skamarock, W.C., Klemp, J.B., Dudhia, J., Gill, D.O., Barker, D.M., Duha, M.G., Huang, X., Wang, W., Powers, J.G., 2008. A description of the Advanced Research WRF Version 3. NCAR Tech. Note 475 STR, 113 pp.
Vannitsem, S., 2008. Dynamical properties of MOS forecasts, analysis of the ECMWF operational forecasting system. Weather Forecasting 23, 1032–1043.
Wallace, J.M., Hobbs, P.V., 2006. Atmospheric Science, An Introductory Survey, second ed. Academic Press, Elsevier, New York, 483 pp.
Wang, Z., Zeng, X., Decker, M., 2010. Improving snow processes in the Noah model. J. Geophys. Res. Atmos. 115 (D20). http://dx.doi.org/10.1029/2009JD013761.
Yao, H., Georgakakos, A.P., 2001. Assessment of Folsom lake response to historical and potential future climate scenarios, 2. Reservoir management. J. Hydrol. 249, 176–196.