AGU PUBLICATIONS

Water Resources Research

RESEARCH ARTICLE
10.1002/2014WR015426

Key Points:
- Multimodel ensemble forecasting systems involve several methodological choices
- A robust framework for decision-making is provided
- The utility of this approach is demonstrated for seasonal streamflow forecasts

Correspondence to:
P. A. Mendoza,
pmendoza@ucar.edu

Citation:
Mendoza, P. A., B. Rajagopalan, M. P. Clark, G. Cortés, and J. McPhee (2014), A robust multimodel framework for ensemble seasonal hydroclimatic forecasts, *Water Resour. Res.*, *50*, doi:10.1002/2014WR015426.

Received 10 FEB 2014
Accepted 3 JUL 2014
Accepted article online 8 JUL 2014

A robust multimodel framework for ensemble seasonal hydroclimatic forecasts

Pablo A. Mendoza[1,2,3], Balaji Rajagopalan[1,2], Martyn P. Clark[3], Gonzalo Cortés[4], and James McPhee[5,6]

[1]Department of Civil, Environmental, and Architectural Engineering, University of Colorado at Boulder, Boulder, Colorado, USA, [2]Cooperative Institute for Research in Environmental Sciences (CIRES), University of Colorado at Boulder, Boulder, Colorado, USA, [3]Research Applications Laboratory, National Center for Atmospheric Research (NCAR), Boulder, Colorado, USA, [4]Department of Civil and Environmental Engineering, University of California, Los Angeles, Los Angeles, California, USA, [5]Department of Civil Engineering, Faculty of Physical and Mathematical Sciences, Universidad de Chile, Santiago, Chile, [6]Advanced Mining Technology Center (AMTC), Faculty of Physical and Mathematical Sciences, Universidad de Chile, Santiago, Chile

Abstract We provide a framework for careful analysis of the different methodological choices we make when constructing multimodel ensemble seasonal forecasts of hydroclimatic variables. Specifically, we focus on three common modeling decisions: (i) number of models, (ii) multimodel combination approach, and (iii) lead time for prediction. The analysis scheme includes a multimodel ensemble forecasting algorithm based on nonparametric regression, a set of alternatives for the options previously pointed, and a selection of probabilistic verification methods for ensemble forecast evaluation. The usefulness of this framework is tested through an example application aimed to generate spring/summer streamflow forecasts at multiple locations in Central Chile. Results demonstrate the high impact that subjectivity in decision-making may have on the quality of ensemble seasonal hydroclimatic forecasts. In particular, we note that the probabilistic verification criteria may lead to different choices regarding the number of models or the multimodel combination method. We also illustrate how this objective analysis scheme may lead to results that are extremely relevant for the case study presented here, such as skillful seasonal streamflow predictions for very dry conditions.

1. Introduction

The requirement of seasonal hydroclimate forecasts for water resources management has historically motivated climate scientists, hydrologists, forecasting agencies, and water managers for developing and implementing innovative techniques in order to improve the quality of predictive systems. Several efforts can be found in the literature, including a wide spectrum of hydroclimatic variables such as sea surface temperature [*Barnston et al*., 1994; *Weisheimer et al*., 2009], precipitation [*Sharma*, 2000; *Block and Rajagopalan*, 2007; *Devineni and Sankarasubramanian*, 2010a], streamflow volumes [*Piechota and Chiew*, 1998; *Souza Filho and Lall*, 2003], snowpack accumulation [*McCabe and Dettinger*, 2002], flood quantiles [*Sankarasubramanian and Lall*, 2003], hurricane activity [*Landsea et al*., 1998], and air temperature [*Huang et al*., 1996; *Hwang et al*., 2001], among others.

A logical step after many years of development has been the incorporation of uncertainty estimates. In view of this, ensemble-based techniques have become widely popular in forecasting applications because of their ability to provide probabilistic information. Typically, an ensemble may contain predictions coming from (i) a single model (e.g., by perturbing inputs, state variables, or parameters), (ii) several models, or (iii) a combination of ensembles coming from different models. Approaches (ii) and (iii) are among the well-known multimodel methods, which have been widely explored in several hydrometeorological applications [e.g., *Krishnamurti et al*., 2000; *Rajagopalan et al*., 2002; *Georgakakos et al*., 2004; *Devineni et al*., 2008; *Block et al*., 2009; *Devineni and Sankarasubramanian*, 2010a,b]. A classic argument to support the use of a multimodel approach has been that it allows "compensatory effects" that control the excess of spread coming from individual model errors. However, it should also be regarded that the verification metrics used to compare the single best model with several multimodel configurations might make a big difference when deciding what approach should be used [*Hagedorn et al*., 2005].

A critical issue in multimodel applications is how to mix forecasts from different models. *Raftery et al*. [2005] introduced Bayesian Model Averaging (BMA) with the aim to combine outputs from several models in order to obtain predictions with good probabilistic properties. Since then, BMA has been used in many hydrometeorological applications [e.g., *Ajami et al*., 2007; *Sloughter et al*., 2007; *Duan et al*., 2007; *Vrugt and Robinson*, 2007; *Fraley et al*., 2010; *Schmeits and Kok*, 2010]. Following the same model-weighting principle, *Regonda et al*. [2006] proposed the Generalized Cross Validation (GCV) score to define weights for each model, which are later used to randomly choose one ensemble forecast from one model to generate a final ensemble prediction. A natural implication from this idea is that other model evaluation scores (e.g., correlation coefficient, bias, root mean square error, etc.) could also be used for assigning weights to forecasts coming from different models. Another relevant decision for seasonal forecasting of hydroclimatic variables is the lead time, which is absolutely critical for operational planning purposes.

In summary, setting up a multimodel forecasting system implies a host of subjective decisions regarding the modeling strategy. In the case of ensemble systems, this subjectivity commonly extends to the choice of verification measures (typically based on skill scores), disregarding other properties that also define the quality of probabilistic forecasts [*Wilks*, 2011]. For instance, the use of a single best model could be appropriate in terms of skill, but it might provide a very poor representation of the uncertainty observed in the hydroclimatic variable of interest. Similarly, a specific multimodel blending technique might provide moderate skill, but a very good ability to discriminate the occurrence of a specific event.

The main goal of this paper is to provide an integrated analysis framework for seasonal forecasts of hydroclimatic variables. Specifically, we aim to contribute for a better decision-making strategy focused on: (1) number of models (single best model versus multimodel), (2) weighting approach used to combine forecasts coming from different models, and (3) lead time for seasonal forecasts. We present an example application that aims to generate ensemble seasonal streamflow forecasts in 10 basins located along the Chilean Andean region between 30° and 34° S. In our example, we perform several experiments in order to assess the quality of seasonal forecasts obtained with different options for the choices listed above. The remainder of this paper is organized as follows: the proposed approach is detailed in section 2, the example application is described in section 3, results are provided in section 4, and discussion and conclusions are presented in section 5.

2. Approach

Given the need to forecast a hydroclimatic variable, we propose a methodology based on three main components:

1. A multimodel ensemble forecasting technique.

2. A set of methodological choices to be tested for: (i) number of models, (ii) the multimodel combination approach, and (iii) the lead time for prediction.

3. A selection of probabilistic verification methods for evaluating the modeling decisions listed above.

Because the ultimate goal of an ensemble forecasting system of hydroclimatic variables is to obtain results with good probabilistic properties (e.g., skill, reliability, uncertainty, etc.), these elements interact with each other as part of the experimental setup. We provide a description of the three components in the following subsections.

2.1. Multimodel Ensemble Forecasting Technique

The seasonal forecasting framework builds upon previous work done by *Grantz et al*. [2005], *Regonda et al*. [2006], and *Bracken et al*. [2010]. The main steps of the forecasting methodology are: (1) predictor identification, (2) selection of best models using objective criteria, and (3) multimodel forecasting algorithm, which is based on the combination of ensembles of predictions coming from the models identified in step (2). Because the implementation of step (1) depends on the characteristics of the forecasting application, we focus on the full description of steps (2) and (3).

2.1.1. Model Selection

If y is the hydroclimatic variable of interest (predictand) and $x_1, x_2, \ldots, x_R$ are the predictor variables, they can be put together into a statistical model with the general form:

$$y = f(x_1, x_2, x_3, \ldots, x_R) + \epsilon \quad (1)$$

where ϵ represents the model error, which is commonly assumed to have a normal probability distribution with mean 0 and standard deviation σ. Although it is very common to adopt a linear model for the function f, several case studies have demonstrated that nonparametric regression techniques (e.g., local polynomials) can be much more effective in obtaining skillful predictions [e.g., *Regonda et al.*, 2005; *Prairie et al.*, 2005; *Towler et al.*, 2009; *Bracken et al.*, 2010]. The main idea behind local polynomial models [*Loader*, 1999] is that, given a point x^* where the prediction is to be made, a number $K = \alpha N$ of neighboring points, where N is the length of the data and $\alpha = (0, 1)$, can be selected in order to fit a polynomial of order p, typically chosen to be 1 or 2. Thus, the fitted polynomial can be used to obtain the mean value of the predictand y^* and also an estimate of the error variance σ^2_{le} at that point.

Given a set of predictors, the parameters K and p are determined by minimizing the generalized cross validation score (GCV), which is defined as:

$$GCV(K, p) = \frac{\sum_{i=1}^{N} \frac{e_i^2}{N}}{(1 - q/N)^2} \quad (2)$$

where e_i is the model error at point x_i, N is the length of the data, and q is the number of parameters. The procedure adopted here for defining a set of models for the multimodel framework is:

1. Determine all possible combinations of predictors, excluding all those sets that contain at least two predictors with linear correlation coefficient higher than a specific threshold.

2. For each combination of predictors, fit local polynomial models for several values of K and p, and compute the GCV score using equation (2).

3. Select those values of K and p that minimize GCV.

4. Repeat steps 2 and 3 for all the combinations of predictors identified in step 1. As a result of this, best values for parameters K and p (i.e., a single best local polynomial model) are identified for each set of predictors.

5. Finally, rank the best models obtained for the sets of predictors according to GCV scores.

The ranking defined above will be later used to select the N_{mod} models (i.e., best sets of predictors) to be included in the multimodel forecasting algorithm.

2.1.2. Multimodel Forecasting Algorithm

Given a number N_{mod} of statistical models to be included in the seasonal forecasting framework, ensemble predictions are generated in cross-validation mode for each time step (season) as follows:

1. Identify the best N_{mod} sets of predictors from the GCV-based ranking described in the previous subsection.

2. Fit local polynomial models for the predictand, and use them to compute a prediction and the associated error variance σ^2_{le} for the year of interest. The variance is used to generate an ensemble of forecasts by adding N_{ens} Gaussian random numbers with mean 0 and variance σ^2_{le}, where N_{ens} is the number of ensemble members.

3. Repeat step 2 for all N_{mod} combinations of predictors identified in step 1. After this, an ensemble forecast of size N_{ens} is obtained from each model.

4. The likelihood of predictions coming from different models is not the same. Hence, weights are defined for each one of the N_{mod} models adjusted. These weights are used to create a cumulative distribution function (cdf) from which a model is randomly selected given a random number with distribution U[0,1]. Then, an ensemble member from that model is randomly sampled. The process is repeated N_f times, being N_f the final size of the ensemble forecast.

Note that steps 2–4 simplify when the model combination technique is Bayesian model averaging, as instead of using weights for resampling from each model, the N_f ensemble members of the hydroclimatic variable are directly obtained by sampling the posterior pdf (see section 2.2.2 for further details).

In this framework, several methodological choices have been left open for further testing. For instance, *Regonda et al*. [2006] proposed a GCV-based weighting approach to combine forecasts from different models, but in practice any other score or combination method can be used. Furthermore, the number of models N_{mod} to be included can be modified in order to make sure that ensemble forecasts with good probabilistic properties are obtained. In the next subsection, we describe some configuration choices that we explore in order to set up a robust seasonal forecasting system.

2.2. Methodological Options

2.2.1. Number of Models

Despite the fact that past studies have demonstrated that multimodel approaches tend to outperform the "best" single model [e.g., *Krishnamurti et al*., 2000; *Rajagopalan et al*., 2002; *Georgakakos et al*., 2004], a careful evaluation must be carried out, since the relative performance of single and multimodel techniques will be determined by the verification methods adopted [*Hagedorn et al*., 2005]. For instance, the best single model might be better in terms of skill, but very poor in representing the uncertainty of observations. Therefore, we recommend the evaluation of the best single model obtained through GCV criteria and the posterior comparison with several multimodel configurations in terms of a set of probabilistic forecast properties (see section 2.3 for more detail).

2.2.2. Multimodel Combination Approach

Previous studies in several fields have demonstrated the utility of combining forecasts from different models to improve the skill of results [*Reid*, 1968; *Bates and Granger*, 1969; *Clemen*, 1989; *Hagedorn et al*., 2005]. In this seasonal forecasting system, we include four different techniques that combine predictions of leading PCs coming from N_{mod} models: a GCV-based approach [*Regonda et al*., 2006], Bayesian model averaging [*Raftery et al*., 2005], Akaike's information criteria [*Akaike*, 1974] and the root mean square error (RMSE). Note that we do not modify the multimodel selection criteria described above, which is still based on GCV, but we do change the way we blend the N_{mod} models included in the forecasting framework.

Generalized Cross Validation (GCV): the weights for each model are computed as 1/GCV, where GCV is obtained using equation (2). This way, the model with the a smaller GCV value will have more weight relative to the one with higher GCV. These weights are normalized in order to make them sum 1, and then a cdf is created, from which one ensemble from one model is randomly selected following step 4 in section 2.1.2. It is important to note that, in this combination approach, the weights assigned to the models are constant for all years according to *Regonda et al*. [2006]. For the following combination methods (except Bayesian model averaging), weights are computed for each year/model in a cross-validation framework.

Bayesian Model Averaging (BMA): the principle of BMA [*Raftery et al*., 2005] states that given an ensemble forecast with N_{mod} members coming from different models, each ensemble member f_i ($i=1, 2, \ldots, N_{mod}$) is associated with a conditional PDF $h_i(y|f_i)$, which can be interpreted as the PDF of the variable y given f_i, given that f_i is the best forecast in the ensemble. Thus, the BMA predictive model is:

$$p(y|f_1, \ldots, f_{Nens}) = \sum_{i=1}^{Nens} w_i h_i(y|f_i) \tag{3}$$

where the BMA weight w_i is the posterior probability of forecast i being the best one, and is based on forecast i's relative performance in the training period. The weights w_is are probabilities, so they are nonnegative and add up to 1, i.e., $\sum_{i=1}^{N_{mod}} w_i = 1$. For the implementation of this option, the weights for all models are estimated by maximum likelihood in cross-validation mode (i.e., using $N - 1$ years for training) and assuming that the conditional PDFs are approximated by a normal distribution. The likelihood is maximized using the expectation-maximization (EM) algorithm [*Dempster et al*., 1977] which is implemented in the package ensembleBMA (http://cran.r-project.org/web/packages/ensembleBMA/ensembleBMA.pdf) at the public domain statistical software R (http://www.rproject.org/). Prior information (i.e., initial weights) is provided for all years using the GCV-based weights previously estimated. Once the N_{mod} weights have been computed for the year of interest, ensembles of the hydroclimatic variable are obtained by directly sampling from the posterior pdf using random numbers with distribution U[0,1].

Akaike's Information Criterion (AIC): the AIC score [*Akaike*, 1974] is a measure of the relative quality of a statistical model, given a set of training data. The mathematical formulation is:

$$AIC = 2R - 2(llh) \tag{4}$$

where R is the number of parameters and llh is the log likelihood function. Given a particular year, the AIC score is computed for each model, whose parameters are adjusted using the remaining $N - 1$ years of data. Then, the weights for each model are computed as 1/AIC and also normalized to make them sum 1.

Root Mean Square Error (RMSE): the final weighting approach is based on the root mean square error of the desired predictand:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N-1} (y_{mod} - y_{obs})^2}{N-1}} \tag{5}$$

Note that RMSE is computed using $N - 1$ points because model parameters are adjusted for each year. Again, the weights are computed as 1/RMSE for all models and then normalized to make them sum unity. Once all weights are computed, ensembles coming from the models are combined following the process previously described.

2.2.3. Lead Time for Seasonal Forecasts

The choice of forecasting lead time has been widely recognized in the literature as a key factor for the predictability of hydroclimatic variables [e.g., *Barnston et al*., 1994; *Huang et al*., 1996; *Sharma*, 2000; *Hwang et al*., 2001; *Pagano et al*., 2004; *Grantz et al*., 2005; *Regonda et al*., 2006; *Weisheimer et al*., 2009; *Ndiaye et al*., 2011], having strong implications on decision making and operational planning. In the case of ensemble forecasts, the selection of lead time may affect significantly some probabilistic properties such as skill, reliability, or uncertainty. Therefore, we consider that this is a critical modeling choice that should ideally be tested in any seasonal forecasting system.

2.3. Probabilistic Verification Methods

Probabilistic verification techniques have become a powerful tool for providing a description of forecast quality in hydrometeorological applications [e.g., *Hamill*, 2001; *Clark and Slater*, 2006; *Laio and Tamea*, 2007; *Stensrud and Yussouf*, 2007; *Pappenberger et al*., 2009; *Renner et al*., 2009; *Mendoza et al*., 2012]. In this paper, we use three probabilistic verification methods: the Ranked Probability Skill Score (RPSS), discrimination diagrams, and QQ plots. The ranked probability skill score represents the level of improvement of a forecast in comparison to a reference forecast, typically assumed to be the mean climatology. The discrimination diagram is a graphic device that describes the ability of a forecast system to distinguish the occurrence of different events. Finally, the predictive QQ plot allows to assess the ability of an ensemble forecast to properly represent the uncertainty of observations. Appendix A provides a detailed description on the calculation and interpretation of these probabilistic verification measures.

3. Example Application

3.1. Motivation, Study Area, and Predictand

In the Chilean Andean area between 30° and 34° S (Figure 1), the rivers born at the Andes Cordillera are the main source of water for human consumption, irrigation, industry, mining, and energy generation. The northern boundary of this region is characterized by a semiarid climate, while in the southern area the rainy season is somewhat longer, starting in May and continuing into the beginning of Spring (September–October). Nevertheless, approximately 85% of precipitation in this region falls during June, July, and August, when frontal systems stemming from the Antarctic storm track reach lower latitudes [*Rubio-Álvarez and McPhee*, 2010]. Precipitation variability is dominated by variations in the southeastern Pacific anticyclone, and El Niño (La Niña) episodes are associated with above (below) average rainfall in central Chile during winter [*Montecinos and Aceituno*, 2003].

Because of the topographic features and particular meteorological conditions of this zone, most of the surface runoff comes from the water accumulated during winter as snowpack, which melts during the spring/

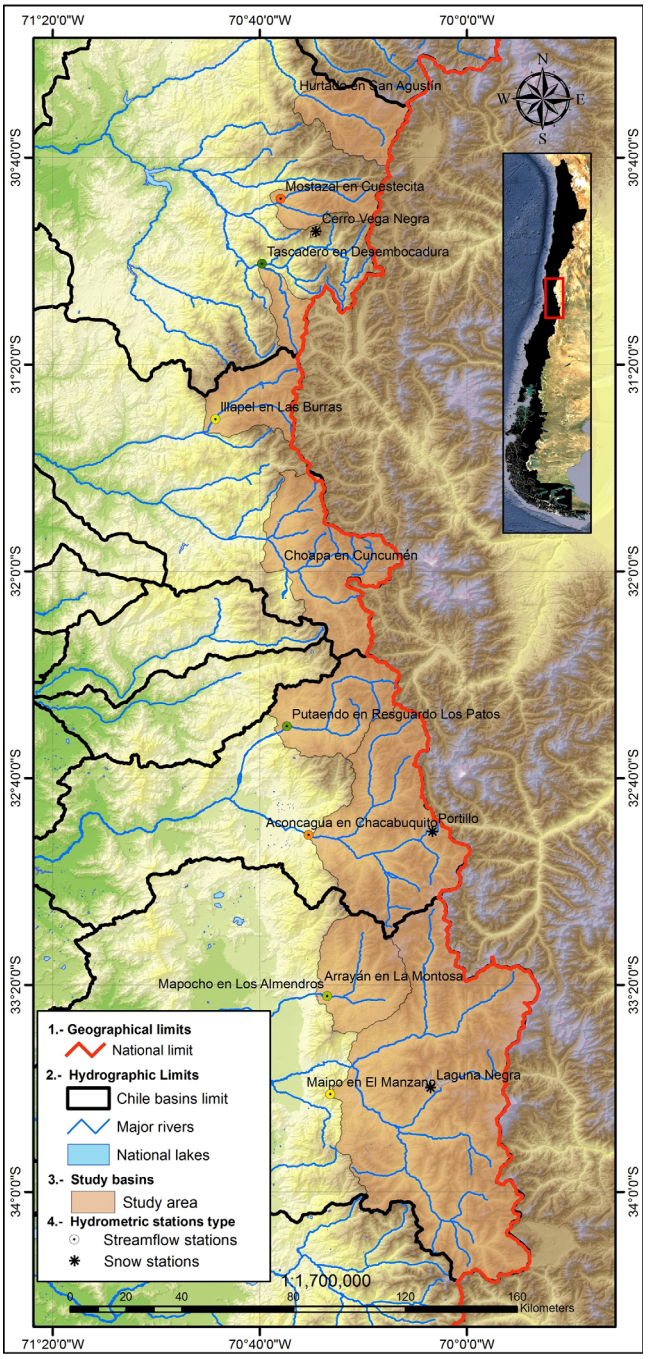

Figure 1. Location of the basins of interest.

summer seasons, and glacier runoff, which becomes relevant at the end of the ablation season [*Ohlanders et al.*, 2013]. Hence, in this area, the Andes cordillera works as a large natural reservoir that provides water for approximately 50% of the total population in Chile.

Therefore, we test our multimodel analysis framework with spring/summer flows (October–March) as predictands. Specifically, we focus our attention on 10 basins that are strategically relevant for water supply and management (Table 1). The main physiographic characteristics of these watersheds are relatively small areas, high slopes, and generally stationary land used patterns resulting from little to no human intervention [*Cortés et al.*, 2011].

3.2. Data

Monthly streamflow records for the period April 1963 to March 2007 are provided by the Ministry of Public Works' Dirección General de Aguas. Annual time series of maximum Snow Water Equivalent (SWE) during August at three snow courses are also available from this agency (Table 1). Finally, large-scale climate and oceanic variables from the $2^{o} \times 2^{o}$ grid of the NCEP-NCAR reanalysis project [*Kalnay et al.*, 1996] are available from NOAA's Climate Diagnostics Center Website (http://www.esrl.noaa.gov/).

3.3. Experimental Setup

3.3.1. Streamflow Analysis

With the aim to better understand the spatial and temporal variability of spring/summer flow in the area of interest, and also to explore the possibility of reducing the dimensionality of the problem (i.e., the number of predictands), we performed principal component analysis (PCA), which is a widely used technique for understanding the predominant modes of variability of hydrometeorological fields. If N is the temporal length of the data, M is the number of sites and $[Z]$ is the $N \times M$ matrix containing the mean spring/summer flow data, PCA allows the following decomposition:

Table 1. Station Data Used in This Study[a]

Station Name	Lat S	Lon W	Elevation (m.a.s.l.)	Area (km^2)	Mean Annual Value	Units
Streamflow stations						
Hurtado en San Agustín	30.46	70.54	2035	656	2.77	m^3/s
Mostazal en Cuestecita	30.80	70.60	1250	353	1.74	m^3/s
Tascadero en Desembocadura	31.01	70.66	1370	238	1.45	m^3/s
Illapel en Las Burras	31.51	70.81	1079	600	2.73	m^3/s
Choapa en Cuncumén	31.97	70.59	1200	1172	10.04	m^3/s
Putaendo en Resguardo Los Patos	32.50	70.58	1218	927	8.40	m^3/s
Aconcagua en Chacabuquito	32.85	70.51	950	2400	34.14	m^3/s
Arrayán en La Montosa	33.33	70.46	970	219	1.65	m^3/s
Mapocho en Los Almendros	33.37	70.45	990	620	6.46	m^3/s
Maipo en El Manzano	33.58	70.67	850	4968	115.63	m^3/s
Snow Stations						
Cerro Vega Negra	30.90	70.52	3600	-	412.7	mm
Portillo	32.84	70.11	3000	-	586.2	mm
Laguna Negra	33.67	70.11	2780	-	517.4	mm

[a]The mean values for streamflow correspond to the period April 1963 to March 2007, while the mean value for Snow Water Equivalent (SWE) is the average computed from the annual time series of maximum SWE in August for the same period.

$$[Z]_{N\times M} = [Y]_{N\times M}[E]^T_{M\times M} \tag{6}$$

$$[Y]_{N\times M} = [Z]_{N\times M}[E]_{M\times M} \tag{7}$$

where the matrix E contains the eigenvectors in each column, and the matrix Y contains principal components (PCs) of the data. Note that the matrix Y is the projection of Z on the orthogonal vectors stored in E. Hence, the PCs are also orthogonal, and commonly only a few of them represent most of the variance in the original data field.

3.3.2. Climate Diagnostics for Predictor Identification

Although in this zone the streamflow monitoring network is quite good in terms of both extension and spatial location, the number of meteorological stations and quality of the data is generally poor. Therefore, we need to look for alternative datasets that provide other potential predictors than SWE. Several past studies have demonstrated the link between streamflow and large-scale atmospheric and oceanic variables [e.g., *Piechota and Chiew*, 1998; *Chiew et al*., 2003; *Souza Filho and Lall*, 2003; *Grantz et al*., 2005; *Block and Rajagopalan*, 2007; *Rubio-Álvarez and McPhee*, 2010; *Cortés et al*., 2011; *Urrutia et al*., 2011]. Hence, we develop correlation maps between the leading PCs of spring/summer flows and reanalysis variables from the preceding fall and winter months. The reanalysis variables considered in this case study are geopotential height (GPH), zonal winds (ZW), meridional wind (MW), surface air temperature (SAT), and precipitable water (PW). The correlation maps generated in this step were used to identify areas where these atmospheric/oceanic variables may be highly correlated with the leading PCs. Based on the highest correlation zones, spatial averages of these variables are extracted for future use as potential predictors.

3.3.3. Multimodel Streamflow Forecasting Algorithm

In this example, the GCV-based ranking is constructed considering the leading PCs of streamflow as predictands. Colinearity among predictors was avoided by choosing a linear correlation threshold of 0.7 (step 1 in section 2.1.1), based on preliminary experiments aimed to reduce the standard errors in predictions. Additionally, the multimodel forecasting algorithm described in section 2.1.2 is adapted for this application as follows:

1. Principal component analysis is performed for mean October–March flows at all M sites, using the $N - 1$ remaining years. Only the leading PCs are selected for prediction.

2. Identify the best N_{mod} sets of predictors from the GCV-based ranking.

3. Fit local polynomial models for the leading PCs, and use them to generate an ensemble of predicted leading PCs following step 2 in section 2.1.2. The nonleading PCs are randomly selected (i.e., bootstrapped) from historic values in order to obtain a complete ensemble forecast of all PCs.

4. Repeat step 3 for all N_{mod} combinations of predictors in order to obtain a matrix with ensemble predictions of PCs with size $(N_{mod} * N_{ens}) \times M$.

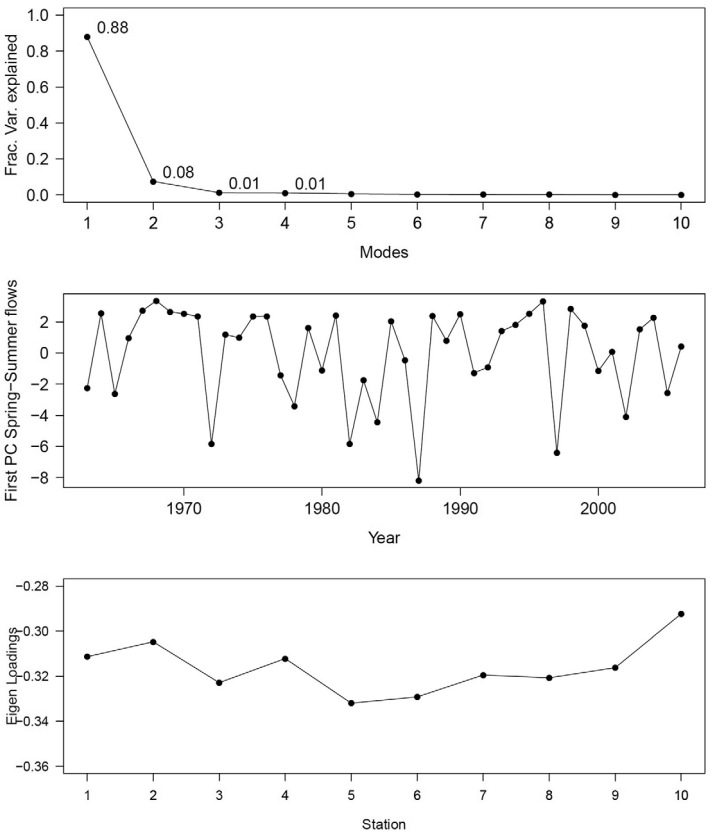

Figure 2. Percentage variance explained by the 10 principal components (PCs) (top), time series of the first PC (middle), and eigenloadings of the first PC at the 10 streamflow locations (bottom).

5. Combine forecasts coming from different models, obtaining a final matrix with ensemble predictions of PCs with size $N_f \times M$, being N_f the final size of the ensemble.

6. Predictions are backtransformed to the original streamflow space using equation (6). After this step, ensemble seasonal streamflow forecasts are obtained at all M locations.

In this algorithm, we use $N_{ens} = N_f = 100$. Note that this forecasting approach differs from *Regonda et al*. [2006] and *Bracken et al*. [2010] in that ensembles from different models here are combined *before* going back to the original streamflow space. We decided to follow this procedure after performing some experiments (not shown here) that demonstrated some improvements in skill and ensemble spread when compared with the multimodel combination framework proposed in those works.

3.3.4. Experiments

In order to provide a careful analysis of the methodological choices listed in section 2.2, we perform the following experiments:

Number of models: With the aim to fairly compare the single best model with other multimodel configurations, we perform a detailed probabilistic and deterministic verification on ensemble streamflow forecasts issued on September 1, including SWE among the set of predictors. In addition, for this first analysis we decide to keep the GCV-based model combination technique.

Model combination approach: Given a number of N_{mod} models, we compare the model combination techniques detailed in section 2.2.2. In this comparison, all streamflow forecasts are issued on September 1 and include SWE in the set of predictors.

Lead time: In Chile, seasonal streamflow forecasts must be provided to water managers at the end of the antecedent snow accumulation season (September 1) in order to supply a reasonable time window for operational planning. In this example, we evaluate the skill of seasonal streamflow forecasts initialized from three different times: September 1 (1 month lead time forecast), August 1 (2 month lead time forecast), and July 1 (3 month lead time forecast). Additionally, we assess the impact of Snow Water Equivalent (SWE) data availability at the end of the month of maximum accumulation (August) on seasonal streamflow forecasts issued on September 1.

4. Results

4.1. Streamflow Characteristics

The results obtained from principal component analysis performed over spring/summer flows are displayed in Figure 2. It is possible to see that the first mode of variability captures 88% of the total variance contained

in the data, a very similar result to what *Regonda et al*. [2006] reported for the Gunnison River Basin, USA. Therefore, we deduce that the first principal component may be used as a representative index of seasonal spring/summer flows in the basins of interest, and the rest of the modes can be treated as "noise." This powerful assumption allows us to reduce substantially the high dimensionality of the original problem, as we can focus on developing statistical models to predict only one variable, instead of seasonal flows at each station location.

The middle panel in Figure 2 shows the time series of the first principal component of mean spring/summer streamflow data, whose temporal evolution has the opposite sign in comparison to the observed seasonal flows in all the basins. It is important to note that the lag-1 year autocorrelation for this variable is equal to −0.143 over the period analyzed, indicating a weak year-to-year dependence (i.e., no persistence). Finally, the bottom panel displays the values of the first column in the matrix E (i.e., the eigenloadings for the first mode of variability). The similarity among these values reaffirms that the hydroclimatic regime in all the basins is practically the same.

4.2. Selection of Predictors

Correlation maps among the first PC of mean spring/summer flows and average values of large-scale atmospheric/oceanic variables for the period May–August are presented in Figure 3. Although these maps were also generated using other time windows during April–September (fall/winter), we found the largest correlations occurring during May–August for the five variables in Figure 3. One may also note that while in some maps (geopotential height, zonal winds, and meridional winds at 1000 mb), there are well-defined areas with both negative and positive correlation, for the cases of surface air temperature at 1000 mb and precipitable water there are only high negative correlation zones.

From the correlation maps, one can note that the first PC of spring/summer flows is positively correlated with geopotential height (1000 mb) in the middle of the South Pacific Ocean, while it is negatively correlated with the same variable close to the Drake passage. This indicates that a decrease in geopotential height over the mid South Pacific and an increase of the same variable over the Drake passage are associated with an increase in spring/summer flows in the Chilean Central Andes (recall that signs of PC1 and actual spring/summer flows are opposite). From Figure 3b, it is inferred that increases in zonal winds over the Pacific Equator and the middle South Pacific will produce increases in spring/summer flows, while a decay in the same variable over Austral Chile will produce the opposite effect. Analysis of meridional winds (Figure 3c) reveals that a decrease in their intensity over the Equator and an increase over Central/South Chile has associated a general decrease in spring/summer flows. The ENSO pattern is reflected in the correlation map with surface air temperature (Figure 3d), which shows that a decrease in this variable will produce a decay in seasonal spring/summer flows. Finally, a negative correlation between the first mode of variability and precipitable water over Central Chile (Figure 3e) simply reflects the link between fall/winter precipitation and spring/summer runoff.

The areas with maximum (positive or negative) correlation between the first mode of mean spring/summer flows are summarized in Table 2. The zones defined in this table were used to extract time series with predictors, computed as the difference of mean values among positive and negative correlation areas. For instance, (GPH-P)-(GPH-N) denote one predictor, computed as the difference between area-averaged geopotential height values over the positively (GPH-P) and the negatively (GPH-N) correlation regions indicated in Table 2. Other large-scale variables preliminarily included in the analysis such as sea level pressure and sea surface temperature were finally removed due to very high spatial correlation with geopotential height and surface air temperature, respectively. Finally, we also included among the set of potential predictors the annual time series with maximum SWE recorded during August, averaged over three snow courses located in the Central Andes.

4.3. Assessment of Methodological Choices

The ranking of models considered in this example application is presented in Table 3. For skill analysis, RPSS is computed for each year, but only the median from all years is reported in the remaining of this paper.

4.3.1. Single Best Model Versus Multimodel Framework

In this section, we examine the impacts of the number of models on the quality of seasonal ensemble forecasts. In order to illustrate such effects in a controlled way, we focus our attention on the GCV-based model

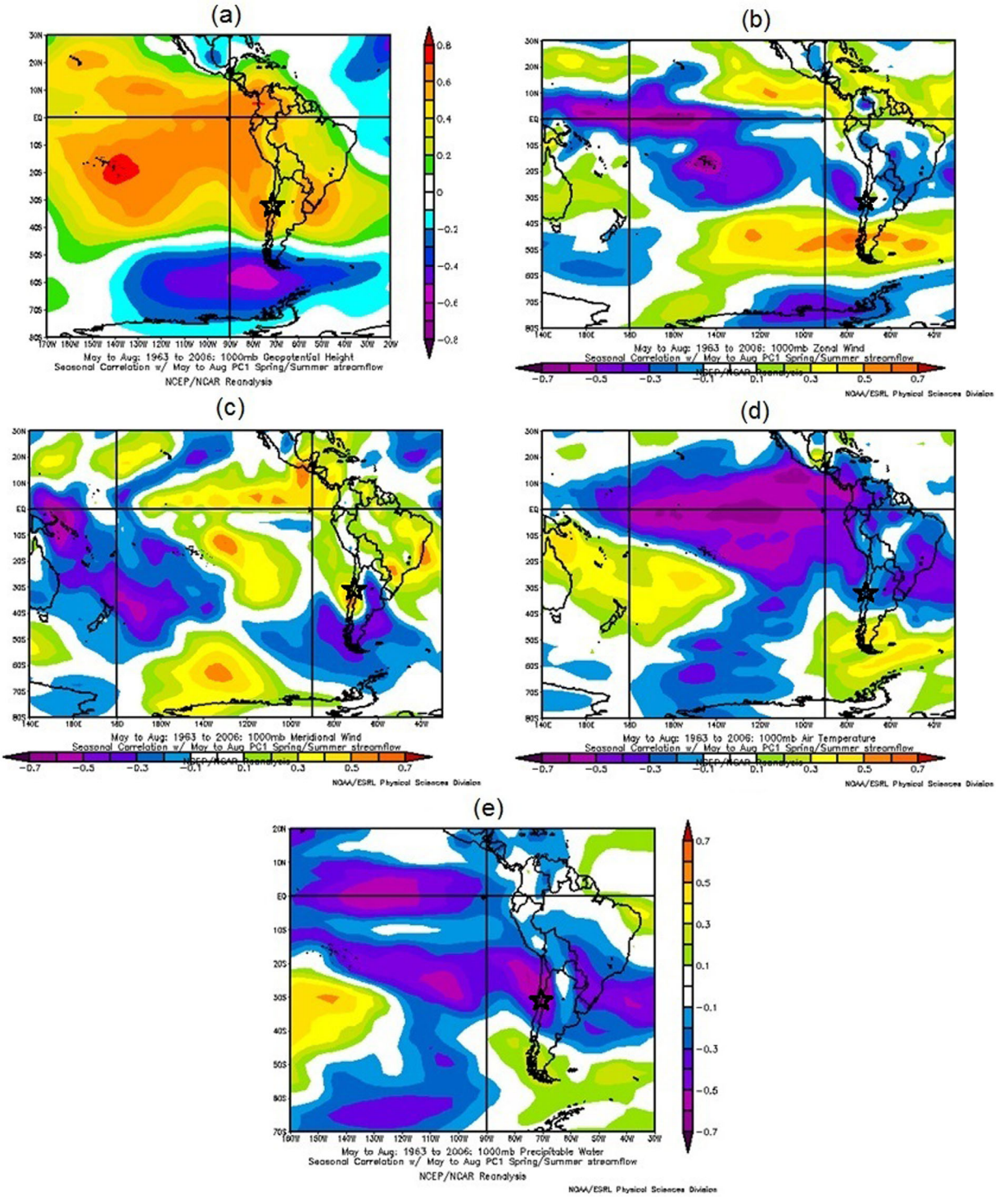

Figure 3. Correlation maps between the first PC of spring/summer flows and May–August large-scale climate variables: (a) geopotential height (1000 mbar), (b) zonal wind (1000 mb), (c) meridional wind (1000 mbar), (d) air temperature (1000 mbar), and (e) precipitable water. Maps were generated from NOAA's Climate Diagnostic Center Website, and the star indicates the location of the study basins.

combination technique. According to the skill results displayed in Figure 4a, scores do not improve further if more than six models are included in the forecasting framework. If we now compare the ability of ensemble forecasts to distinguish the occurrence of dry (below the 33% seasonal streamflow percentile), normal

Table 2. Potential Predictors and Their Regions of Negative and Positive Correlation

Index	Climate Variable	Lead Time	Season	Negative Region (N)	Positive Region (P)
GPH	Geopotential height	1 Sep	May–Aug	−62, −55N:274,286E	−22,-16N:217,228E
ZW1	Zonal winds	1 Sep	May–Aug	−18.5, −14.5N:210,219E	−47,-44.2N:279.5,282.6E
ZW2	Zonal winds	1 Sep	May–Aug	−2,2N:192,208E	−47,-44.2N:279.5,282.6E
MW	Meridional winds	1 Sep	May–Aug	−5, −1N:152.5,156.5E	−37,-34.5N:286.5,289.5E
SAT	Surface air temperature	1 Sep	May–Aug	−4,2N:220,249E	
PW	Precipitable water	1 Sep	May–Aug	−30.5, −27N:290,293E	

(between the 33% and the 66% percentiles) and wet years (above the 66% seasonal streamflow percentile), we do not find sharp differences among predictive pdf's for dry and normal years, regardless of the number of models included (Figure 5). On the other hand, discrimination between wet years and the rest is very clear, especially when $N_{mod} \geq 6$. From the QQ plots presented in Figure 6, it is inferred that all multimodel configurations preserve the ability to correctly represent the uncertainty in observations. Finally, probabilistic verification through reliability diagrams revealed that, as expected, the sample size is not reliable (i.e., only 44 water years x 10 stations = 440 site-years for verification) regardless of the number of models considered; therefore, these results do not provide relevant information to evaluate multimodel configurations and are not shown here.

We also compare bias and linear correlation coefficients including observations and ensemble forecast medians at all locations. For this analysis, we standardize the results using the total basin areas, transforming streamflow values from m^3/s to mm/d. The scatter plots presented in Figure 7 prove that good correlation and bias results can be obtained by combining only three models. In other words, the inclusion of additional models will not improve forecasting results substantially.

From the probabilistic verification and scatter plots described above, it can be summarized that for the example examined here: (1) the inclusion of more than six models does not provide better skill and discrimination results, (2) the framework provides good uncertainty representation regardless of the number of models, and (3) only three models are needed to obtain adequate bias and correlation between ensemble forecast medians and observations. Additionally, we also noted that the spread in forecast ensembles decreased if more models were included, especially for very wet years (e.g., 1987). Over this region, El Niño episodes explain an important part of long-term variability in rainfall and snow accumulation [*Masiokas et al*., 2006], and many of the wettest years observed during the last 40 years correspond to positive ENSO oscillations. With this in mind, it is important for any seasonal streamflow forecasting framework implemented for this area to accurately represent variability observed during the strong positive ENSO episodes.

From the previous analysis, we conclude that, for this particular application, the minimum number of best models that ensures a good performance for the verification measures included in this framework is six.

Table 3. Ranking of Models for September 1[a]

Ranking	No. of Predictors	(GPH-P)-(GPH-N)	(MW-P)-(MW-N)	(ZW1-P)-(ZW1-N)	(ZW2-P)-(ZW2-N)	SAT-N	PW-N	SWE	GCV
1	2	0	1	0	0	0	0	1	0.931
2	3	0	0	1	0	0	1	1	0.966
3	3	1	0	0	0	0	1	1	0.971
4	4	1	1	0	0	0	1	1	1.021
5	3	0	0	0	1	0	1	1	1.038
6	4	0	0	1	0	1	1	1	1.048
7	3	0	1	0	0	0	1	1	1.052
8	2	0	0	1	0	0	0	1	1.068
9	2	0	0	0	0	0	1	1	1.078
10	3	0	0	1	0	1	0	1	1.101
11	3	1	1	0	0	0	0	1	1.152
12	2	0	0	0	1	0	0	1	1.172

[a]The letter P (N) indicates that the variable of interest has been spatially averaged over the positively (negatively) correlated region, following the notation introduced in Table 2. Presence and absence of predictors are indicated by "1" and "0," respectively.

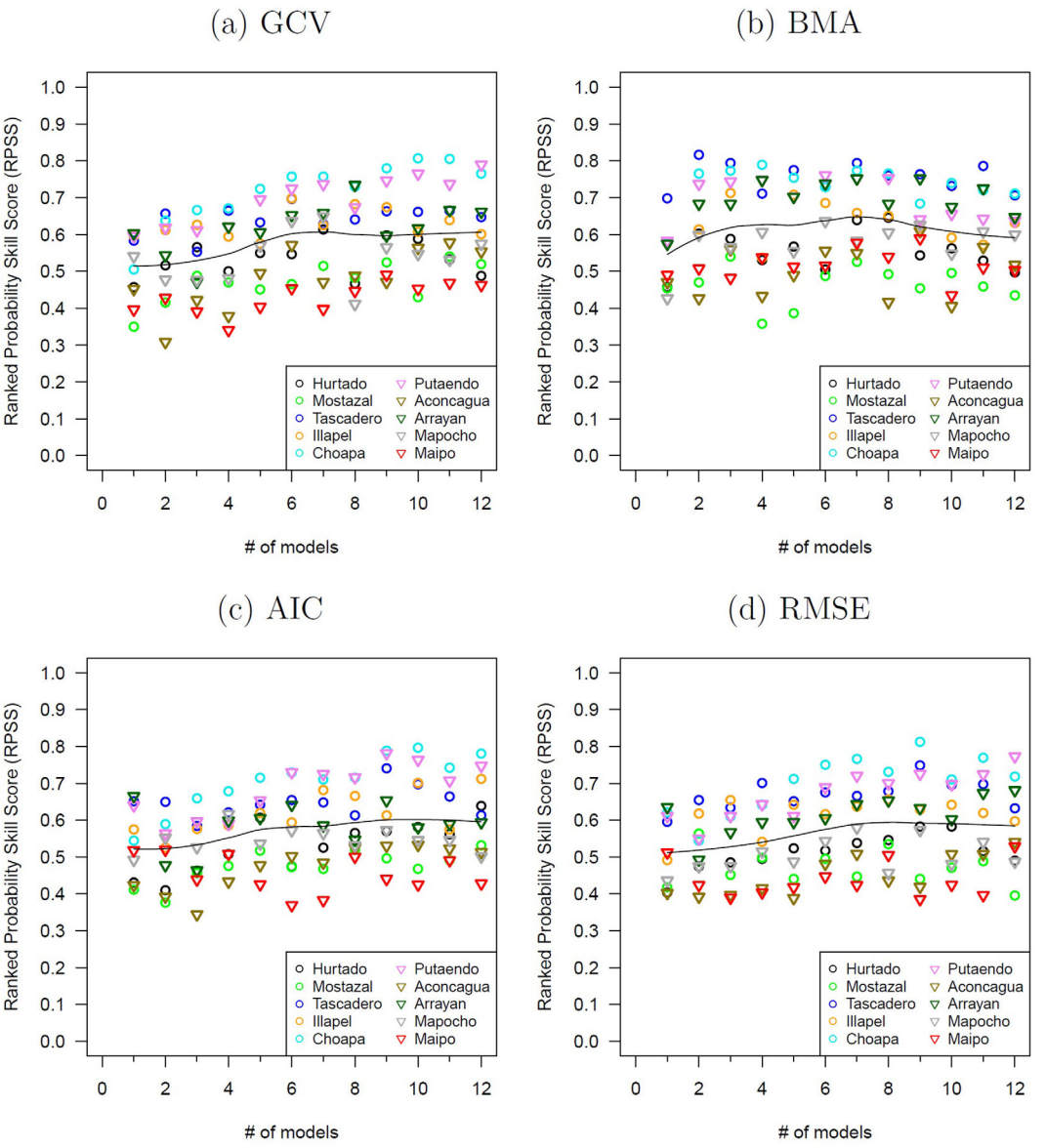

Figure 4. Ranked Probability Skill Score (RPSS) as a function of the number of models included in the multimodel forecasting framework for different model combination approaches: (a) Generalized Cross Validation (GCV) score, (b) Bayesian Model Averaging (BMA), (c) Akaike's Information Criterion (AIC), and (d) Root Mean Square Error (RMSE). The skill score reported for each basin corresponds to the median from all years. All forecasts are issued on September 1 and include SWE in the set of predictors.

4.3.2. Model Combination Approach

Skill score results as a function of the number of models included (using the GCV-based ranking described in section 2.1.1) in the multimodel forecasting framework, using four different model combination approaches, are presented in Figure 4. Ranked Probability Skill Scores are displayed for all basins, representing the northern and southern basins with colored circles and triangles, respectively. The black continuous lines represent a local polynomial smoother that helps to visualize how many models are required to stabilize RPSS values. One can infer that very similar skill values and skill spread across basins are obtained when GCV, AIC, and RMSE are used as weighting criteria. Additionally, better skill scores are obtained when BMA is used for combining seven or less local polynomial models. No significant gain in skill is obtained if more than 6–7 models are included. Furthermore, no relevant difference in skill between northern and southern basins was found.

Is it appropriate to limit the evaluation of methodological choices (in this case, model combination approach) to a single criterion? In the past subsection, we could see that the "best" number of models varies

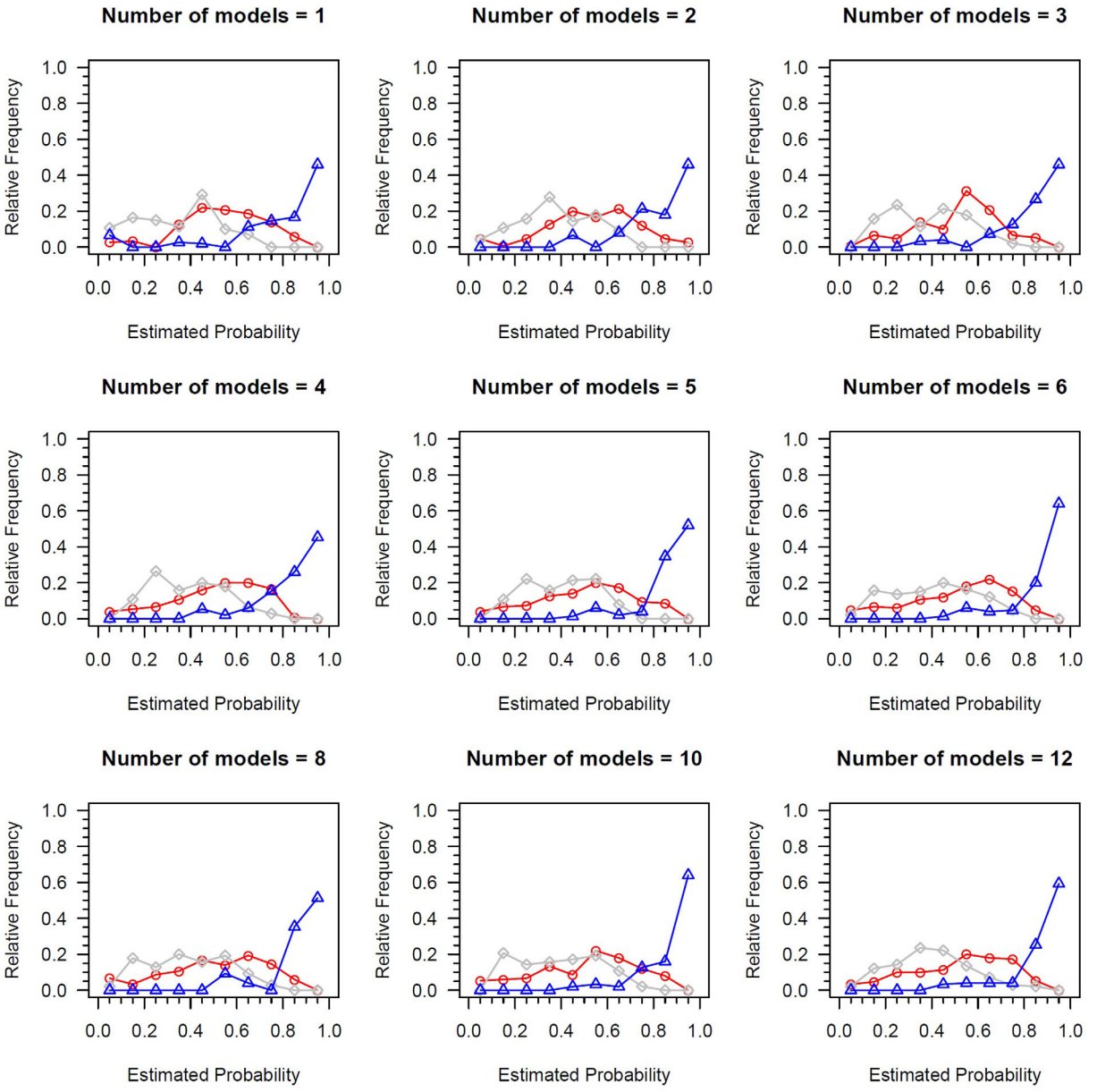

Figure 5. Discrimination diagrams for leave-one-out cross validation spring/summer forecasts generated with different multimodel configurations. The red (circles), gray (diamonds), and blue (triangles) lines represent the pdf of forecasts probabilities for dry years (below the 33% observed quantile), normal years (between the 33% and 66% observed quantiles), and wet years (above the 66% observed quantile), respectively.

depending on the verification method, and the selection of multimodel combination technique also follows this rule. Indeed, when comparing the time series with cross-validated ensemble seasonal streamflow forecasts at a specific location and number of models $N_{mod} = 6$ (Figure 8), we found that the gain in skill obtained with BMA does not necessarily bring a better match between ensemble median and observations (e.g., years 1987 and 1997). Even more, the ability of BMA-based forecasts to discriminate among dry, normal, and wet years is not superior to that obtained using the GCV-weighting approach, for which better defined and sharper pdf's are obtained (see Figure 9 for a comparison among the four model blending techniques). These results reaffirm the idea that finding the "best" methodological choice is not a trivial problem, as it will depend on the forecasts properties sought by modelers and decision makers.

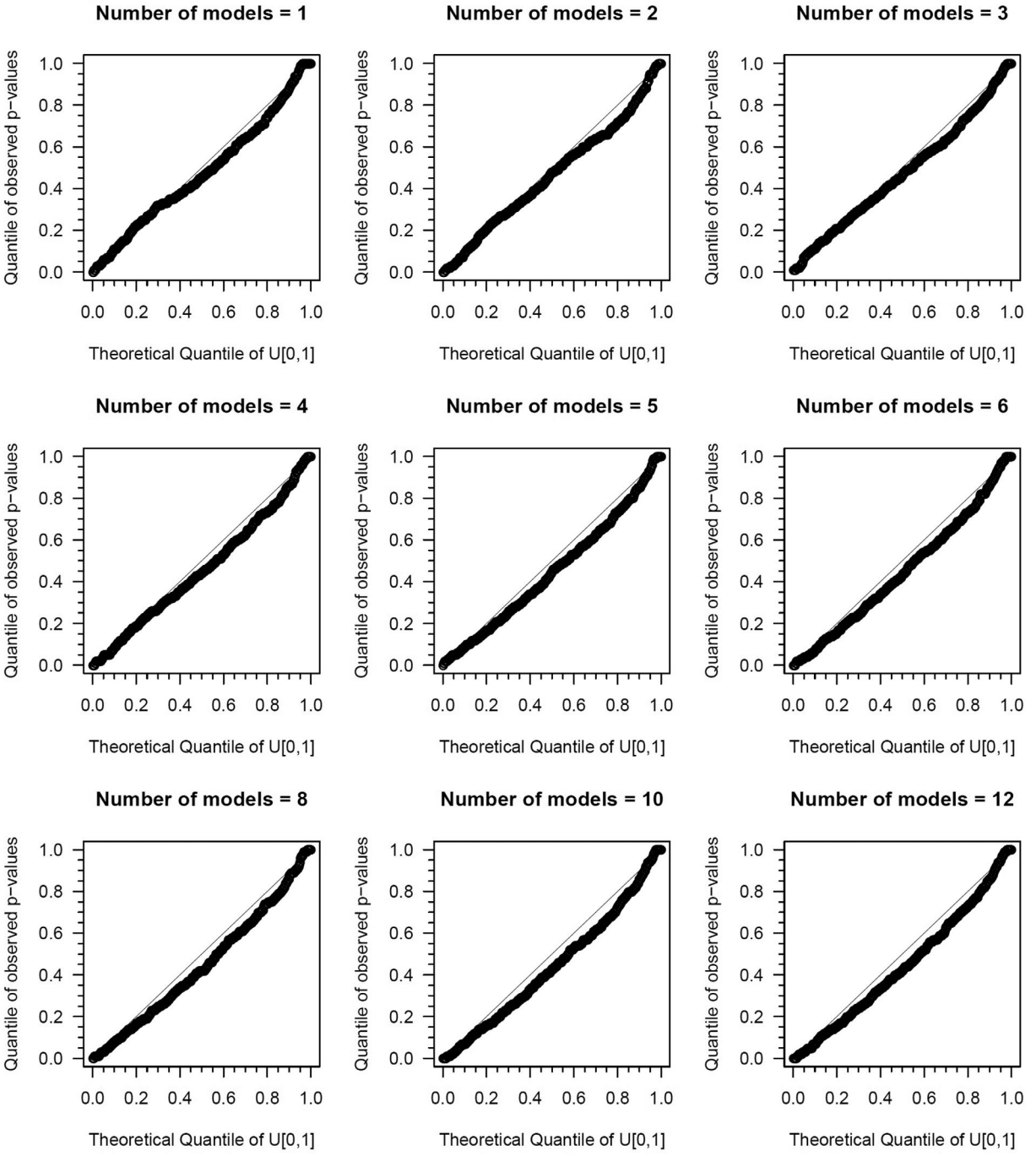

Figure 6. QQ plots for leave-one-out cross validation spring/summer forecasts generated with different numbers of models.

4.3.3. Lead Time

Ranked probability skill scores as a function of the number of models for all basins are presented in Figure 10 for the following cases: (a) forecasts issued on September 1, including SWE in the predictor set, (b) forecasts issued in September 1, excluding SWE, (c) forecasts issued on August 1, and (d) forecasts issued on July 1. GCV-based weighting approach was used to combine models in all cases. From these results, it is clear that SWE contributes considerably to obtain skillful predictions at all station locations, independently

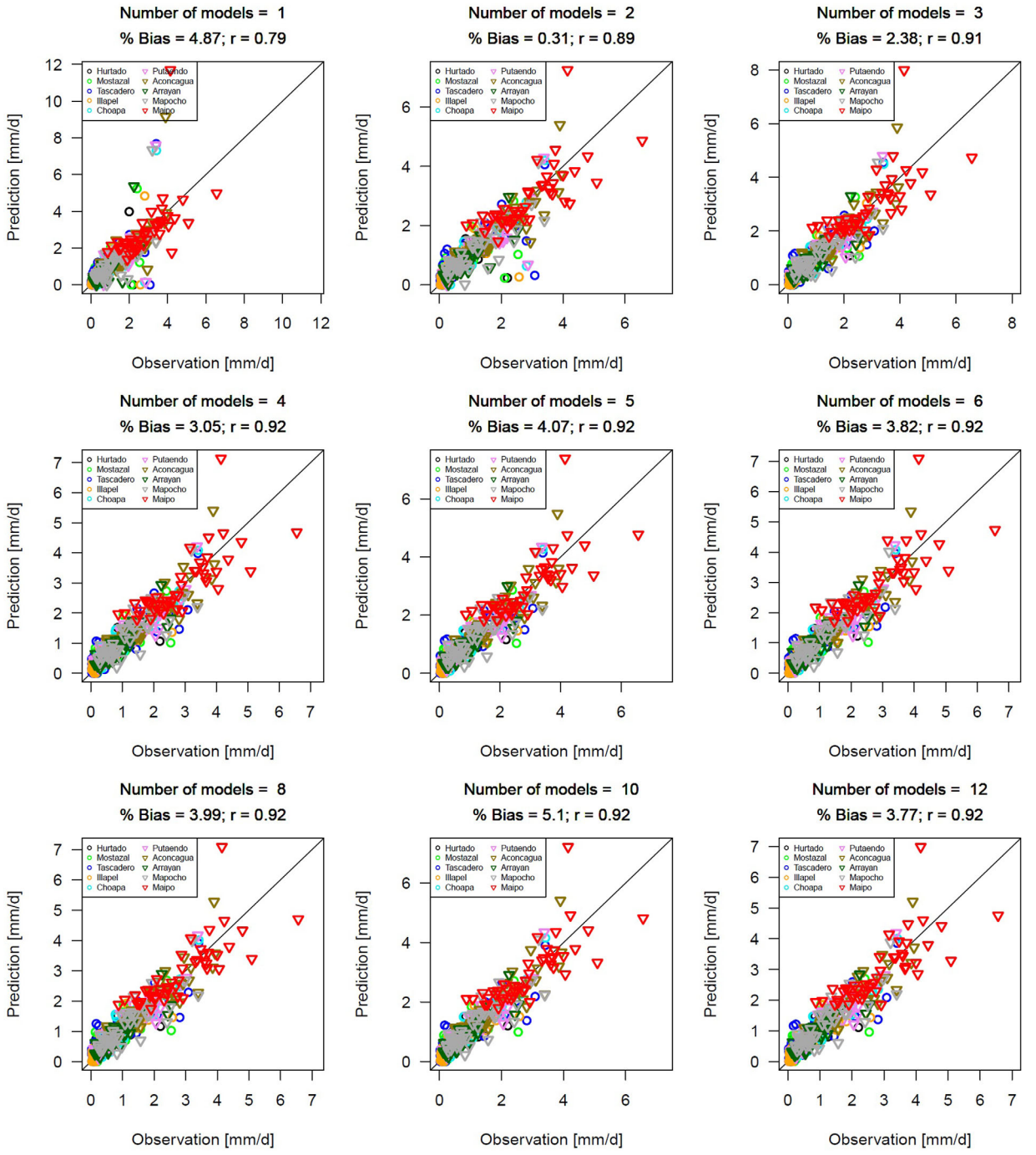

Figure 7. Scatter plots between the 50% seasonal streamflow predictions (ensemble median) issued on September 1 and observed spring/summer flows at all locations obtained with different multimodel configurations.

of the number of models included in the multimodel framework. However, even if SWE data is not available due September 1, skill is still positive in most of the basins (Figure 10b). For the generation of seasonal forecasts on August 1 and July 1, because of the lack of antecedent snowpack information only large-scale variables are included as averages over May–July and May–June, respectively (Figures 10c and 10d). The low skill obtained in both cases demonstrate that hydrometeorological information associated with July and August, when still considerable precipitation amounts may fall over the Central Andes Cordillera, cannot be

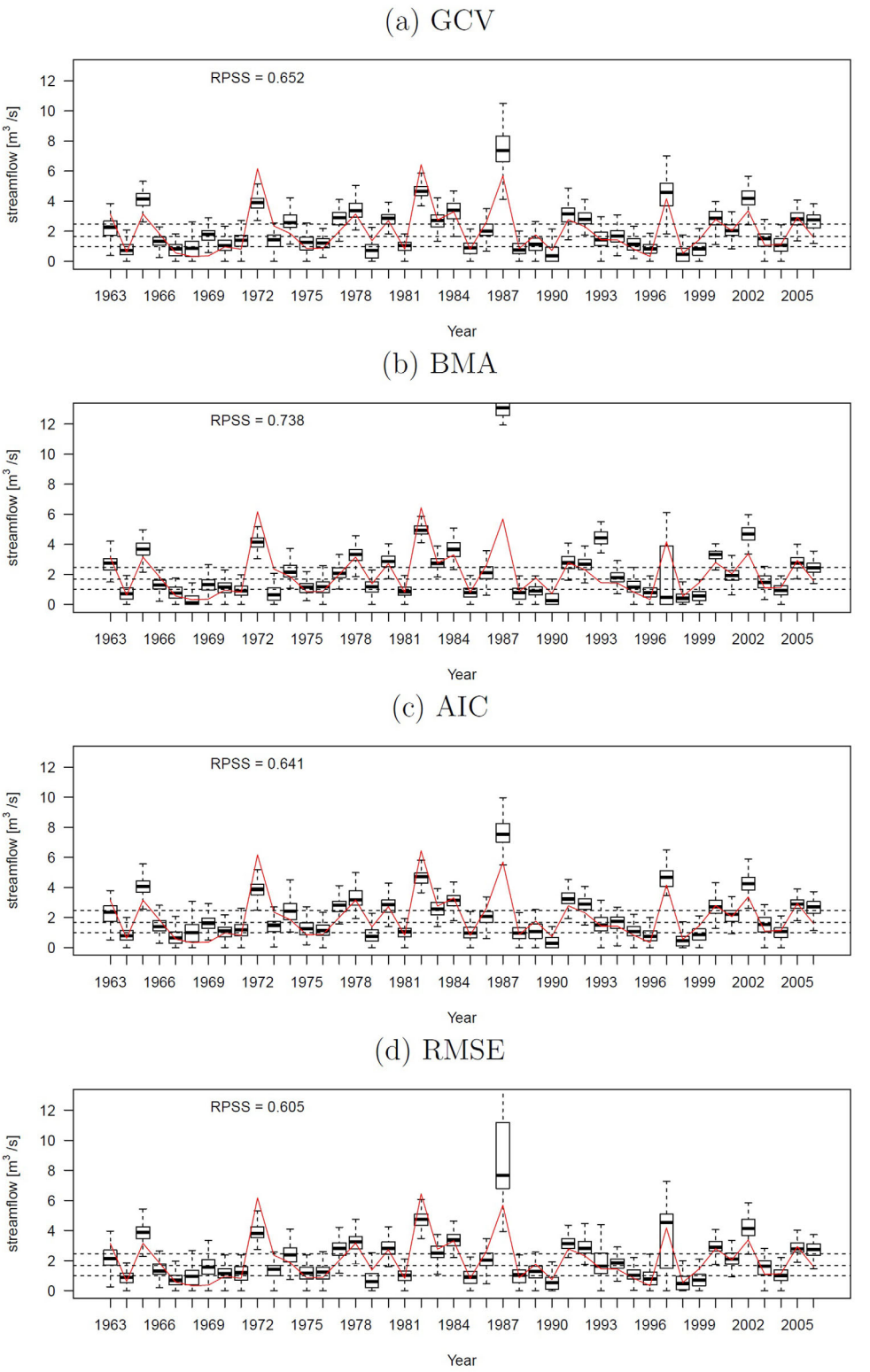

Figure 8. Leave-one-out cross validation ensemble spring/summer forecasts for Arrayán en La Montosa issued on September 1 for different weighting approaches: (a) Generalized Cross Validation (GCV) score, (b) Bayesian Model Averaging (BMA), (c) Akaike's Information Criteria (AIC), and (d) Root Mean Square Error (RMSE). The best six models from Table 3 are combined in all cases.

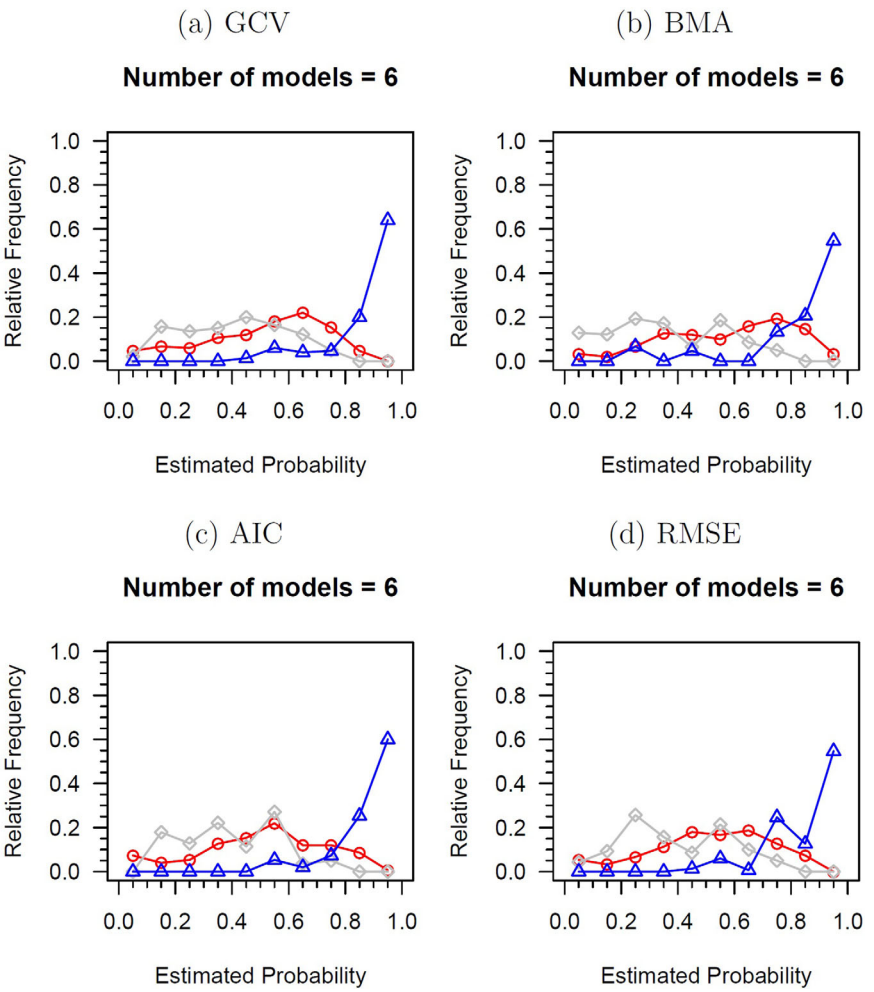

Figure 9. Discrimination diagrams for leave-one-out cross validation spring/summer forecasts generated for different model combination approaches: (a) Generalized Cross Validation (GCV) score, (b) Bayesian Model Averaging (BMA), (c) Akaike's Information Criterion (AIC), and (d) Root Mean Square Error (RMSE). The red (circles), gray (diamonds), and blue (triangles) lines represent the pdf of forecasts probabilities for dry years (below the 33% observed quantile), normal years (between the 33% and 66% observed quantiles), and wet years (above the 66% observed quantile), respectively.

disregarded in order to get good streamflow predictions during the melting season. In view of this, and although positive skill was obtained at some locations, we infer that the system tested here has limited predictability on July 1 and August 1, and therefore the use of ensemble forecasts issued for these lead times should be made with caution. Nevertheless, additional sources of information (i.e., predictors) could be explored in the future with the aim to improve streamflow predictability in this region.

4.4. Performance in Dry Years

In this example application area, evidence of a positive trend in temperature [*Carrasco et al.*, 2005; *Falvey and Garreaud*, 2009] and negative trends in precipitation [*Quintana and Aceituno*, 2012] suggest that the frequency and intensity of drought events may increase during the next years, highlighting the importance of getting skillful flow predictions under very dry conditions. Therefore, we decide to test the multimodel configuration found via objective criteria for those years where the observed values are less than the observed 25th percentile, which was computed separately for each station. Figure 11 contains the time series with ensemble spring/summer streamflow forecasts issued on September 1 including SWE information, six models, and GCV-based multimodel combination. RPSS values reported here are the median from only those years included in each panel (dry years). The inspection of these plots reveals that performance in terms of skill is still good, although it decreases

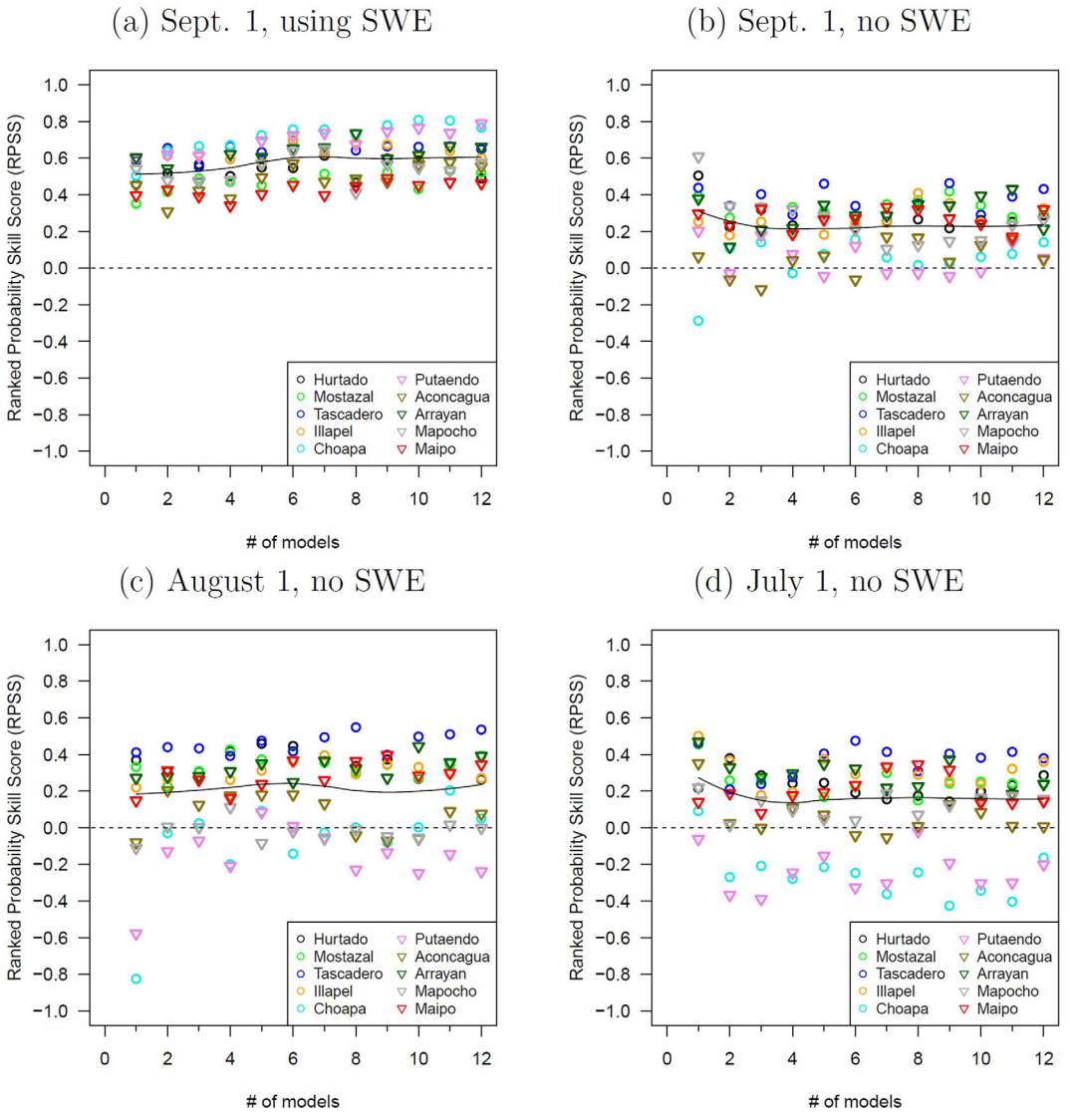

Figure 10. Ranked Probability Skill Score (RPSS) as a function of the number of models included in the multimodel forecasting framework for different cases: (a) forecast issued on September 1, with SWE, (b) forecast issued on September 1, no SWE, (c) forecast issued on August 1, no SWE and (d) forecast issued on July 1, no SWE.

when compared to the entire time series (1963–2006), spanning from 0.265 (Mostazal en Cuestecita) to 0.678 (Aconcagua en Chacabuquito).

5. Discussion and Conclusions

Over the last decade, several studies have reported the benefits of using multimodel methods in hydroclimate forecasting applications. However, a common deficiency that can be found in the literature is the lack of detailed evaluation of key methodological choices involved in the development of multimodel seasonal forecasting systems. Therefore, we propose a framework for the assessment of relevant decisions via objective criteria, focusing on: (i) number of models, (ii) multimodel combination approach, and (iii) lead time for prediction. Our methodology is based on three elements: a multimodel ensemble forecasting algorithm based on nonparametric regression, a set of options for the modeling decisions previously listed, and a suite of probabilistic verification techniques for the evaluation of ensemble forecasts. We also provide an example application in order to illustrate the utility of our methodology. The case study presented here is aimed

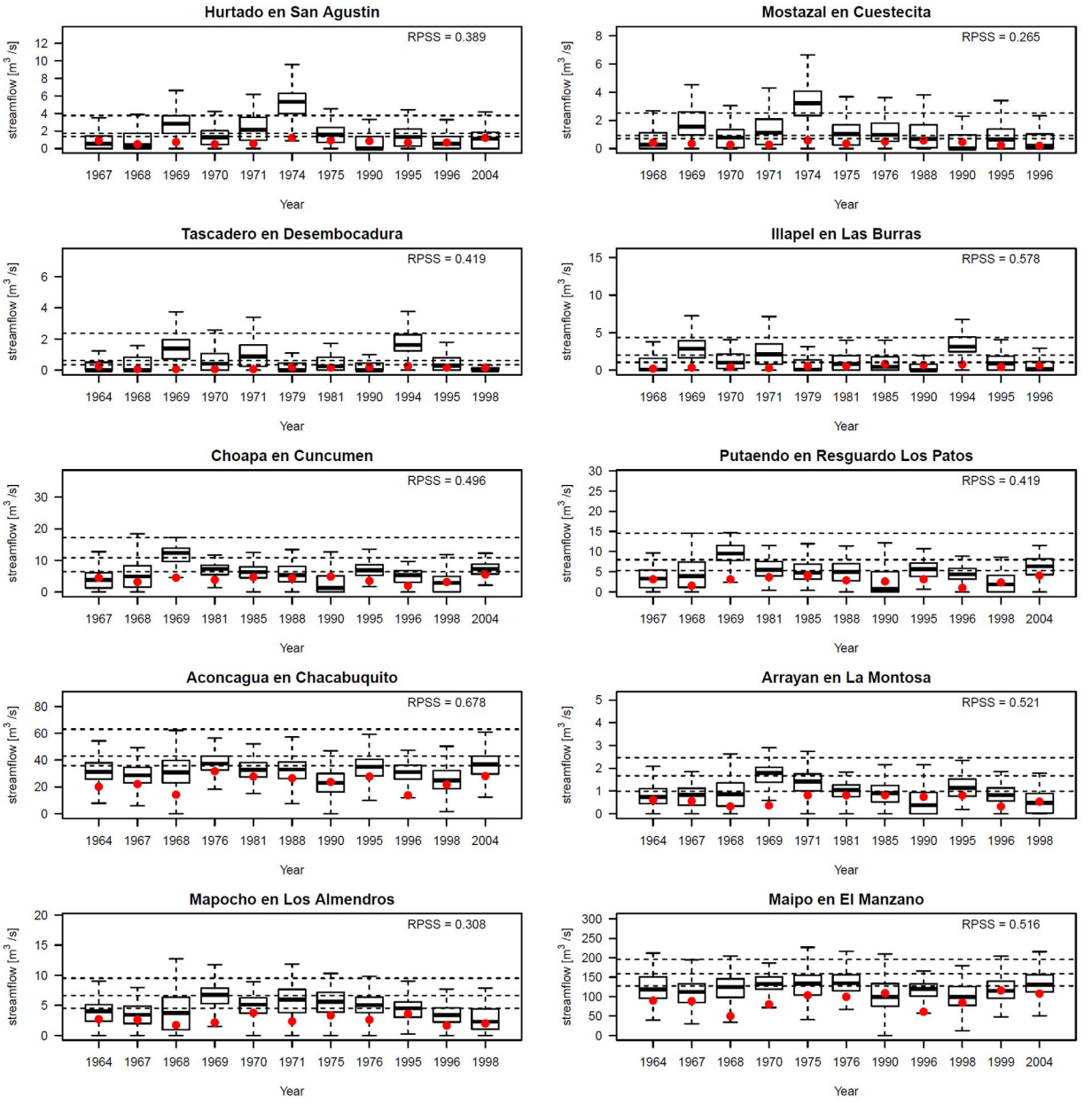

Figure 11. Leave-one-out cross validated ensemble spring/summer forecasts at all sites for very dry years (i.e., observed spring/summer flow has a probability of exceedance larger than 75%). Forecasts are issued on September 1 using the best six models, GCV-weighting approach, and including SWE information.

to generate seasonal forecasts of spring/summer streamflow at 10 basins located in the Chilean semiarid Andean region between 30° and 34° S.

The comparison of the best single model with other multimodel configurations (i.e., number of models) showed that while the best single model is enough for streamflow uncertainty representation, three models are needed to obtain a good correlation and bias among the median of ensemble seasonal forecasts and observations, and six models assure a good skill and a sharper discrimination among dry, normal, and wet years. Regarding the choice of the multimodel combination method, it was obtained that although Bayesian model averaging provided better skill scores than the rest of the methods, especially when six or less models were included, a worse performance was obtained in terms of discrimination when compared with GCV-based model weighting technique. These results demonstrate that the two multimodel decisions included in this framework (number of models and multimodel combination) are strongly dependent on the forecast evaluation criteria.

A comparison of different forecast lead times revealed that the availability of SWE data is critical to correctly reproduce interannual streamflow variability and extremes, even in a region with a very limited monitoring network. Furthermore, hydroclimatological information contained in large-scale predictors for July and August is critical to obtain skillful forecasts for the melting season, as considerable precipitation amounts can still fall during these months. Hence, from the objective assessment of methodological choices, it was obtained that, in the example application examined here, the use of six models with GCV-based model weighting and lead time of 1 month can provide seasonal predictions with good probabilistic properties at all sites. The multimodel seasonal forecasting configuration obtained also proved its usefulness for providing skillful predictions under very dry conditions.

It is noteworthy that since the proposed methodology is based on data-driven models, multidecade time series are ideally required in order to develop robust statistical relationships. Moreover, despite the extension of the datasets used in our example application (44 years) was long enough to construct forecasts with good probabilistic properties, longer datasets—and therefore longer training periods—would certainly help to build more reliable forecasting systems.

Although limited number of forecasting options have been tested in this paper for each methodological choice, the forecaster could naturally explore other alternatives (e.g., different model-weighting techniques, additional lead time forecasts, etc.), keeping in mind that the choice of verification criteria may impact decision-making. Furthermore, ensemble forecasts obtained using a framework like this could be wisely combined with predictions from existing deterministic models. One option could be the "enlargement" of the best multimodel ensemble found by adding the best deterministic forecast as a new individual member [*Rodwell*, 2006]. The weight of the deterministic forecast in the new ensemble may be computed using a cross-validation strategy, i.e., defining a training period to calibrate the weights via a statistical post processing technique (e.g., Bayesian model averaging), and then apply these weights for the year of interest. Alternatively, the spread of the best multimodel ensemble can be used to derive a probability distribution around the best deterministic forecast. For instance, *Blanc* [2009] found that a simple statistical model can be used to predict the uncertainty of a deterministic 2 m temperature forecast, using the spread of ensemble temperature forecasts as the only predictor.

The results obtained in this study demonstrate the impact that decision-making may have on the quality of hydroclimatic ensemble forecasts. Consequently, the selection of one or more probabilistic verification methods should be properly justified, because the "best" configuration option is closely tied to the evaluation metrics used. In view of this, we strongly encourage forecasters to perform a careful analysis of the configuration choices adopted in order to develop more robust forecasting systems.

Appendix A: Probabilistic Verification Methods

A1. Ranked Probability Skill Score

The Ranked Probability Skill Score (RPSS) measures the accuracy of multicategory probability forecasts relative to a climatological forecast. If k is the number of mutually exclusive categories for seasonal flow, $p=(p_1, p_2, \ldots, p_k)$ is the probabilistic forecast (obtained as the number of ensemble members within each category divided by the ensemble size) and $d=(d_1, d_2, \ldots, d_k)$ is the observation vector, such that d_j equals 1 if the observation falls in the j-th category and 0 otherwise, the Ranked Probability Score is defined as:

$$RPS_f = \sum_{i=1}^{k} \left[\left(\sum_{j=1}^{i} p_j - \sum_{j=1}^{i} d_j \right)^2 \right] \tag{A1}$$

In this study, we use $k = 3$, with the categories defined by the tercile boundaries at each observation location (i.e., 33% and 66% percentiles obtained from the historical record). The climatological forecast for each category is 1/3. If one wants to compare with other sites or data sets, it is recommended to use the Ranked Probability Skill Score (RPSS):

$$RPSS = 1 - \frac{RPS_f}{RPS_{clim}} \tag{A2}$$

RPSS values range from positive 1 (perfect forecast) to negative infinity. Negative RPSS values indicate that the forecast is less skillful than mean climatology, positive values indicate the opposite, and null values indicate that forecasts are equally skillful when compared with climatology.

A2. Discrimination Diagram

The discrimination diagram [*Wilks*, 2011] shows the ability of the forecast system to clearly distinguish situations leading to the occurrence of an event of interest from those leading to the nonoccurrence of the event. Thus, given a polychotomous predictand ($J = 3$), the discrimination diagram consists on the superimposed plots of the three conditional distributions $p(y_i|o_j), j=1, 2, 3$ as functions of forecast probabilities y_i. A perfect discrimination between the three events will be given by no overlap between their likelihoods. Finally, a good discrimination will depend on the separation of means of conditional distributions, and on the variance within conditional distributions.

A3. Predictive QQ Plot

In order to assess how well ensemble forecasts represent the uncertainty in observations, we use the predictive QQ plot [*Laio and Tamea*, 2007; *Thyer et al*., 2009]. Let y_t be the variable representing runoff at an instant t, F_t the cumulative distribution function for that variable, and $\tilde{y}_t$ the corresponding observation [*Thyer et al*., 2009]. A probabilistic forecast of y_t will be correct if the observed probability density function $p_t(\tilde{y}_t)$ coincides with the true distribution of y_t, $f_t(y_t)$. Even if $f_t(y_t)$ is not known (the distribution changes with t and there is only one observed value available), it is possible to construct a simple test of hypothesis [*Laio and Tamea*, 2007]:

$$H_0 : p_t(y_t) = f_t(\tilde{y}_t) \tag{A3}$$

This test is based on the evaluation of the cumulative distribution function (built from the set of forecasts at every time step) for the observation, i.e., find $z_t = P_t(\tilde{y}_t)$ [*Laio and Tamea*, 2007]. Under the hypothesis H_0, the quantile found above should be, like $F_t(\tilde{y}_t)$, a realization from a uniform distribution on U[0,1] [*Thyer et al*., 2009]. Once z_t quantiles are computed, these values must be ranked from lowest to highest, getting the positions R_t of these quantiles within that system. Finally, the predictive QQ plot is obtained by plotting the values of z_t in terms of R_t/N (theoretical quantile U[0,1]) where N is the number of events or time steps analyzed. If the curve obtained matches the 1:1 line, the observation is equally likely to be any ensemble member. The reader is referred to *Thyer et al*. [2009] for further details in the interpretation of QQ plots.

Acknowledgments
The authors wish to thank Dirección General de Aguas for providing the streamflow and snowpack data, Yohann Videla for his assistance with GIS datasets, and James McCreight, Katrina Grantz, and two anonymous reviewers for their helpful and insightful comments in improving this manuscript. The first author acknowledges support from the Cooperative Institute for Research and Environmental Sciences (CIRES), Bureau of Reclamation (USBR), and the U.S. Army Corps of Engineers (USACE).

References

Ajami, N. K., Q. Duan, and S. Sorooshian (2007), An integrated hydrologic Bayesian multimodel combination framework: Confronting input, parameter, and model structural uncertainty in hydrologic prediction, *Water Resour. Res.*, *43*, W01403, doi:10.1029/2005WR004745.

Akaike, H. (1974), A New Look at Statistical Model Identification, *IEEE Trans. Autom. Control*, *19*, 716–723.

Barnston, A., et al. (1994), Long-lead seasonal forecasts-Where do we stand?, *Bull. Am. Meteorol. Soc.*, *75*(11), 2097–2114.

Bates, J., and C. Granger (1969), The combination of forecasts, *Oper. Res. Q.*, *20*(4), 451–468.

Blanc, P. (2009), Ensemble-based uncertainty prediction for deterministic 2 m temperature forecasts, *Veröffentlichungen der MeteoSchweiz*, *82*, master thesis, 90 pp., University of Bern, Zurich, Switzerland.

Block, P., and B. Rajagopalan (2007), Interannual variability and ensemble forecast of upper blue Nile Basin Kiremt season precipitation, *J. Hydrometeor.*, *8*(3), 327–343, doi:10.1175/JHM580.1.

Block, P. J., F. A. Souza Filho, L. Sun, and H. -H. Kwon (2009), A streamflow forecasting framework using multiple climate and hydrological models, *J. Am. Water Resour. Assoc.*, *45*(4), 828–843, doi:10.1111/j.1752-1688.2009.00327.x.

Bracken, C., B. Rajagopalan, and J. Prairie (2010), A multisite seasonal ensemble streamflow forecasting technique, *Water Resour. Res.*, *46*, W03532, doi:10.1029/2009WR007965.

Carrasco, J. F., G. Casassa, and J. Quintana (2005), Changes of the 0C isotherm and the equilibrium line altitude in central Chile during the last quarter of the 20th century/Changements de l'isotherme 0C et de la ligne d'équilibre des neiges dans le Chili central durant le dernier quart du 20ème siècle, *Hydrol. Sci. J.*, *50*(6), 933–948, doi:10.1623/hysj.2005.50.6.933.

Chiew, F., S. Zhou, and T. McMahon (2003), Use of seasonal streamflow forecasts in water resources management, *J. Hydrol.*, *270*(1-2), 135–144, doi:10.1016/S0022-1694(02)00292-5.

Clark, M. P., and A. G. Slater (2006), Probabilistic quantitative precipitation estimation in complex terrain, *J. Hydrometeor.*, *7*, 3–22.

Clemen, R. T. (1989), Combining forecasts: A review and annotated bibliography, *Int. J. Forecast.*, *5*, 559–583.

Cortés, G., X. Vargas, and J. McPhee (2011), Climatic sensitivity of streamflow timing in the extratropical western Andes Cordillera, *J. Hydrol.*, *405*(1-2), 93–109, doi:10.1016/j.jhydrol.2011.05.013.
Dempster, A., N. Laird, and D. Rubin (1977), Maximum likelihood from incomplete data via the EM algorithm, *J. R. Stat. Soc., Ser. B*, *39*(1), 1–38.
Devineni, N., and A. Sankarasubramanian (2010a), Improving the prediction of winter precipitation and temperature over the continental United States: role of the ENSO state in developing multimodel combinations, *Mon. Weather Rev.*, *138*(6), 2447–2468, doi:10.1175/2009MWR3112.1.
Devineni, N., and A. Sankarasubramanian (2010b), Improved categorical winter precipitation forecasts through multimodel combinations of coupled GCMs, *Geophys. Res. Lett.*, *37*, L24704, doi:10.1029/2010GL044989.
Devineni, N., A. Sankarasubramanian, and S. Ghosh (2008), Multimodel ensembles of streamflow forecasts: Role of predictor state in developing optimal combinations, *Water Resour. Res.*, *44*, W09404, doi:10.1029/2006WR005855.
Duan, Q., N. K. Ajami, X. Gao, and S. Sorooshian (2007), Multi-model ensemble hydrologic prediction using Bayesian model averaging, *Adv. Water Resour.*, *30*(5), 1371–1386, doi:10.1016/j.advwatres.2006.11.014.
Falvey, M., and R. D. Garreaud (2009), Regional cooling in a warming world: Recent temperature trends in the southeast Pacific and along the west coast of subtropical South America (19792006), *J. Geophys. Res.*, *114*, D04102, doi:10.1029/2008JD010519.
Fraley, C., A. E. Raftery, and T. Gneiting (2010), Calibrating multimodel forecast ensembles with exchangeable and missing members using bayesian model averaging, *Mon. Weather Rev.*, *138*(1), 190–202, doi:10.1175/2009MWR3046.1.
Georgakakos, K. P., D.-J. Seo, H. Gupta, J. Schaake, and M. B. Butts (2004), Towards the characterization of streamflow simulation uncertainty through multimodel ensembles, *J. Hydrol.*, *298*(1-4), 222–241, doi:10.1016/j.jhydrol.2004.03.037.
Grantz, K., B. Rajagopalan, M. Clark, and E. Zagona (2005), A technique for incorporating large-scale climate information in basin-scale ensemble streamflow forecasts, *Water Resour. Res.*, *41*, W10410, doi:10.1029/2004WR003467.
Hagedorn, R., F. Doblas-Reyes, and T. Palmer (2005), The rationale behind the success of multimodel ensembles in seasonal forecasting I. Basic concept, *Tellus Ser. A*, *57*, 219–233.
Hamill, T. M. (2001), Interpretation of rank histograms for verifying ensemble forecasts, *Mon. Weather Rev.*, *129*, 550–560.
Huang, J., H. van den Dool, and A. Barnston (1996), Long-lead seasonal temperature prediction using optimal climate normals, *J. Clim.*, *9*, 809–817.
Hwang, S.-O., J.-K. E. Schemm, A. G. Barnston, and W.-T. Kwon (2001), Long-lead seasonal forecast skill in far eastern Asia using canonical correlation analysis, *J. Clim.*, *14*(13), 3005–3016, doi:10.1175/1520-0442(2001)014¡3005:LLSFSI¿2.0.CO;2.
Kalnay, E., et al. (1996), The NCEP/NCAR 40-year reanalysis project, *Bull. Am. Meteorol. Soc.*, *77*(3), 437–471.
Krishnamurti, T., C. Kishtawal, Z. Zhang, T. LaRow, D. Bachiochi, and E. Williford (2000), Multimodel ensemble forecasts for weather and seasonal climate, *J. Clim.*, *13*, 4196–4216.
Laio, F., and S. Tamea (2007), Verification tools for probabilistic forecasts of continuous hydrological variables, *Hydrol. Earth Syst. Sci.*, *11*(4), 1267–1277, doi:10.5194/hess-11-1267-2007.
Landsea, C., G. Bell, W. Gray, and S. Goldenberg (1998), The extremely active 1995 Atlantic hurricane season: Environmental conditions and verification of seasonal forecasts, *Mon. Weather Rev.*, *126*, 1174–1193.
Loader, C. (1999), *Local Regression and Likelihood*, Springer, N. Y.
Masiokas, M. H., R. Villalba, B. H. Luckman, C. Le Quesne, and J. C. Aravena (2006), Snowpack variations in the Central Andes of Argentina and Chile, 1951–2005: Large-scale atmospheric influences and implications for water resources in the region, *J. Clim.*, *19*, 6334–6352.
McCabe, G., and M. Dettinger (2002), Primary modes and predictability of year-to-year snowpack variations in the western United States from teleconnections with Pacific Ocean climate, *J. Hydrometeor.*, *3*, 13–25.
Mendoza, P. A., J. McPhee, and X. Vargas (2012), Uncertainty in flood forecasting: A distributed modeling approach in a sparse data catchment, *Water Resour. Res.*, *48*, W09532, doi:10.1029/2011WR011089.
Montecinos, A., and P. Aceituno (2003), Seasonality of the ENSO-related rainfall variability in central Chile and associated circulation anomalies, *J. Clim.*, *16*(1), 281–296.
Ndiaye, O., M. N. Ward, and W. M. Thiaw (2011), Predictability of seasonal Sahel rainfall using GCMs and lead-time improvements through the use of a coupled model, *J. Clim.*, *24*(7), 1931–1949, doi:10.1175/2010JCLI3557.1.
Ohlanders, N., M. Rodriguez, and J. McPhee (2013), Stable water isotope variation in a Central Andean watershed dominated by glacier and snowmelt, *Hydrol. Earth Syst. Sci.*, *17*(3), 1035–1050, doi:10.5194/hess-17-1035-2013.
Pagano, T., D. Garen, and S. Sorooshian (2004), Evaluation of official western US seasonal water supply outlooks, 19222002, *J. Hydrometeor.*, *5*, 896–909.
Pappenberger, F., A. Ghelli, R. Buizza, and K. Bódis (2009), The skill of probabilistic precipitation forecasts under observational uncertainties within the generalized likelihood uncertainty estimation framework for hydrological applications, *J. Hydrometeor.*, *10*(3), 807–819, doi:10.1175/2008JHM956.1.
Piechota, T., and F. Chiew (1998), Seasonal streamflow forecasting in eastern Australia and the El NiñoSouthern oscillation, *Water Resour. Res.*, *34*(11), 3035–3044.
Prairie, J., B. Rajagopalan, T. J. Fulp, and E. A. Zagona (2005), Statistical nonparametric model for natural salt estimation, *J. Environ. Eng.*, *131*(1), 130–138.
Quintana, J., and P. Aceituno (2012), Changes in the rainfall regime along the extratropical west coast of South America (Chile): 30–43 S, *Atmósfera*, *25*(1), 1–22.
Raftery, A. E., T. Gneiting, F. Balabdaoui, and M. Polakowski (2005), Using Bayesian model averaging to calibrate forecast ensembles, *Mon. Weather Rev.*, *133*, 1155–1174.
Rajagopalan, B., U. Lall, and S. Zebiak (2002), Categorical climate forecasts through regularization and optimal combination of multiple GCM ensembles*, *Mon. Weather Rev.*, *130*(7), 1792–1811.
Regonda, S. K., B. Rajagopalan, U. Lall, M. Clark, and Y. Moon (2005), Nonlinear processes in geophysics local polynomial method for ensemble forecast of time series, *Nonlin. Processes Geophys.*, *12*, 397–406.
Regonda, S. K., B. Rajagopalan, M. Clark, and E. Zagona (2006), A multimodel ensemble forecast framework: Application to spring seasonal flows in the Gunnison River Basin, *Water Resour. Res.*, *42*, W09404, doi:10.1029/2005WR004653.
Reid, D. (1968), Combining three estimates of gross domestic product, *Economica*, *35*(140), 431–444.
Renner, M., M. Werner, S. Rademacher, and E. Sprokkereef (2009), Verification of ensemble flow forecasts for the River Rhine, *J. Hydrol.*, *376*(3-4), 463–475, doi:10.1016/j.jhydrol.2009.07.059.
Rodwell, M. (2006), Comparing and combining deterministic and ensemble forecasts: How to predict rainfall occurrence better, *ECMWF Newslett.*, No. 106, 17–23, Bob Riddaway, European Centre for Medium-Range Weather Forecasts (ECMWF), Shinfield Park, Reading, England.

Rubio-Álvarez, E., and J. McPhee (2010), Patterns of spatial and temporal variability in streamflow records in south central Chile in the period 1952–2003, *Water Resour. Res.*, *46*, W05514, doi:10.1029/2009WR007982.

Sankarasubramanian, A., and U. Lall (2003), Flood quantiles in a changing climate: Seasonal forecasts and causal relations, *Water Resour. Res.*, *39*(5), 1134, doi:10.1029/2002WR001593.

Schmeits, M. J., and K. J. Kok (2010), A comparison between raw ensemble output, (modified) Bayesian model averaging, and extended logistic regression using ECMWF ensemble precipitation reforecasts, *Mon. Weather Rev.*, *138*(11), 4199–4211, doi:10.1175/2010MWR3285.1.

Sharma, A. (2000), Seasonal to interannual rainfall probabilistic forecasts for improved water supply management: Part 3 A nonparametric probabilistic forecast model, *J. Hydrol.*, *239*(1-4), 249–258, doi:10.1016/S0022-1694(00)00348-6.

Sloughter, J. M. L., A. E. Raftery, T. Gneiting, and C. Fraley (2007), Probabilistic quantitative precipitation forecasting using Bayesian model averaging, *Mon. Weather Rev.*, *135*(9), 3209–3220, doi:10.1175/MWR3441.1.

Souza Filho, F. A., and U. Lall (2003), Seasonal to interannual ensemble streamflow forecasts for Ceara, Brazil: Applications of a multivariate, semiparametric algorithm, *Water Resour. Res.*, *39*(11), 1307, doi:10.1029/2002WR001373.

Stensrud, D. J., and N. Yussouf (2007), Reliable probabilistic quantitative precipitation forecasts from a short-range ensemble forecasting system, *Weather Forecast.*, *22*(1), 3–17, doi:10.1175/WAF968.1.

Thyer, M., B. Renard, D. Kavetski, G. Kuczera, S. W. Franks, and S. Srikanthan (2009), Critical evaluation of parameter consistency and predictive uncertainty in hydrological modeling: A case study using Bayesian total error analysis, *Water Resour. Res.*, *45*, W00B14, doi:10.1029/2008WR006825.

Towler, E., B. Rajagopalan, and R. S. Summers (2009), Using parametric and nonparametric methods to model total organic, *Environ. Eng. Sci.*, *26*(8), 1299–1308.

Urrutia, R. B., A. Lara, R. Villalba, D. A. Christie, C. Le Quesne, and A. Cuq (2011), Multicentury tree ring reconstruction of annual streamflow for the Maule River watershed in south central Chile, *Water Resour. Res.*, *47*, W06527, doi:10.1029/2010WR009562.

Vrugt, J. A., and B. A. Robinson (2007), Treatment of uncertainty using ensemble methods: Comparison of sequential data assimilation and Bayesian model averaging, *Water Resour. Res.*, *43*, W01411, doi:10.1029/2005WR004838.

Weisheimer, A., F. J. Doblas-Reyes, T. N. Palmer, A. Alessandri, A. Arribas, M. Déqué, N. Keenlyside, M. MacVean, A. Navarra, and P. Rogel (2009), ENSEMBLES: A new multi-model ensemble for seasonal-to-annual predictions: Skill and progress beyond DEMETER in forecasting tropical Pacific SSTs, *Geophys. Res. Lett.*, *36*(21), L21711, doi:10.1029/2009GL040896.

Wilks, D. S. (2011), *Statistical Methods in the Atmospheric Sciences*, 3rd ed., 467 pp., Academic Press, Waltham, Mass.