WATER RESOURCES RESEARCH, VOL. 49, 2729–2746, doi:10.1002/wrcr.20251, 2013

Global analysis of seasonal streamflow predictability using an ensemble prediction system and observations from 6192 small catchments worldwide

Albert I. J. M. van Dijk,[1,2] Jorge L. Peña-Arancibia,[2] Eric F. Wood,[3] Justin Sheffield,[3] and Hylke E. Beck[4]

Received 22 March 2012; accepted 9 April 2013; published 28 May 2013.

[1] Ideally, a seasonal streamflow forecasting system would ingest skilful climate forecasts and propagate these through calibrated hydrological models initialized with observed catchment conditions. At global scale, practical problems exist in each of these aspects. For the first time, we analyzed theoretical and actual skill in bimonthly streamflow forecasts from a global ensemble streamflow prediction (ESP) system. Forecasts were generated six times per year for 1979–2008 by an initialized hydrological model and an ensemble of 1° resolution daily climate estimates for the preceding 30 years. A post-ESP conditional sampling method was applied to 2.6% of forecasts, based on predictive relationships between precipitation and 1 of 21 climate indices prior to the forecast date. Theoretical skill was assessed against a reference run with historic forcing. Actual skill was assessed against streamflow records for 6192 small (<10,000 km^2) catchments worldwide. The results show that initial catchment conditions provide the main source of skill. Post-ESP sampling enhanced skill in equatorial South America and Southeast Asia, particularly in terms of tercile probability skill, due to the persistence and influence of the El Niño Southern Oscillation. Actual skill was on average 54% of theoretical skill but considerably more for selected regions and times of year. The realized fraction of the theoretical skill probably depended primarily on the quality of precipitation estimates. Forecast skill could be predicted as the product of theoretical skill and historic model performance. Increases in seasonal forecast skill are likely to require improvement in the observation of precipitation and initial hydrological conditions.

Citation: van Dijk, A. I. J. M., J. L. Peña-Arancibia, E. F. Wood, J. Sheffield, and H. E. Beck (2013), Global analysis of seasonal streamflow predictability using an ensemble prediction system and observations from 6192 small catchments worldwide, *Water Resour. Res.*, *49*, 2729–2746, doi:10.1002/wrcr.20251.

1. Introduction

1.1. Background

[2] Seasonal streamflow forecasts for months to seasons ahead have many potential uses, even though their skill is inherently limited by the chaotic nature of the atmosphere. Their main use at present is to assist in planning the operation of water reservoirs for hydropower, agricultural and urban water supply, flood mitigation, and environmental flows [*Cherry et al*., 2005; *Hamlet et al*., 2002; *Pagano et al*., 2004]. Forecasts of the likelihood of above- or below-average streamflow levels can also help river water users, environmental water managers, and floodplain communities in decision making, as well as national or international organizations involved in water trading, policy making, regulation, and aid and emergency response [*Chiew et al*., 2003; *Pappenberger et al*., 2011; *Ritchie et al*., 2004; *Sankarasubramanian and Lall*, 2003]. The skill of seasonal forecasts is derived from knowledge of the state of the climate system and of catchment conditions before the forecast period. The state of the climate system, in particular, the pattern of ocean currents and corresponding sea surface temperatures, can provide a basis for forecasting future climate state [*Palmer and Anderson*, 1994]. The predictive value can be exploited using coupled general circulation models (GCMs) or statistical methods based on the state of the climate system (e.g., sea surface temperature and pressure fields). The state of the catchment, in particular, the amount of water stored in the snowpack, soil, and groundwater water system, contributes skill where (some of) this stored water will be released from the catchment or

Additional supporting information may be found in the online version of this article.

[1]Fenner School for Environment & Society, The Australian National University, Canberra, Australian Capital Territory, Australia.
[2]CSIRO Land and Water, Canberra, Australian Capital Territory, Australia.
[3]Department of Civil and Environmental Engineering, Princeton University, Princeton, New Jersey, USA.
[4]Department of Hydrology and Geo-Environmental Sciences, Faculty of Earth and Life Sciences, VU University Amsterdam, Amsterdam, Netherlands.

Corresponding author: A. I. J. M. van Dijk, Fenner School for Environment & Society, The Australian National University, Canberra, ACT 0200, Australia. (albert.vandijk@anu.edu.au)

0043-1397/13/10.1002/wrcr.20251

influences catchment response to precipitation during the forecast period [*Bierkens and van Beek*, 2009; *Koster et al*., 2010].

[3] Seasonal forecasts are currently being issued routinely in several regions. Forecasts of spring streamflow have been issued for several decades in the western United States and Canada based on empirical regression relationships with snow depth observations [*Gobena and Gan*, 2010; *Pagano et al*., 2004]. Since 2010, the Australian Bureau of Meteorology issues monthly experimental probabilistic streamflow forecasts for several stations in southeast Australia derived using Bayesian methods and trained on indices of climate state (e.g., El Niño Southern Oscillation (ENSO)) and antecedent precipitation and streamflow [*Wang et al*., 2009]. Dynamic (i.e., hydrological model-based) forecasting methods have also been developed. In particular, the ensemble streamflow prediction (ESP) system uses antecedent meteorological observations to initialize a hydrological model which is run in forward mode using an ensemble of sampled historic climate time series [*Day*, 1985; *Franz et al*., 2003]. At present, seasonal streamflow forecasts—statistical or dynamic—are not available in most other parts of the world.

[4] Seasonal precipitation forecasts from coupled GCMs are routinely issued by several centers [e.g., *Saha et al*., 2006; *Weisheimer et al*., 2009] and can be propagated through hydrological models to obtain seasonal streamflow forecasts with global coverage. This is not yet done on an operational basis, for several possible reasons. First, the current generation of precipitation forecasts rapidly loses skill beyond the first 2 weeks [*Lavers et al*., 2009; *Yuan et al*., 2011], which is to a considerable extent a fundamental constraint caused by the chaotic evolution of the atmosphere [*Westra and Sharma*, 2010; *Feng et al*., 2011]. Moreover, GCM forecasts are issued at a resolution that is too coarse for most hydrological applications, and the forecasts, of precipitation in particular, are prone to biases. This requires investment in developing bias correction and downscaling methods [*Gobena and Gan*, 2010; *Luo and Wood*, 2008] that may or may not be justified by the added skill in forecasts. There are alternative approaches to GCM downscaling that can still exploit climate state observations to condition seasonal hydrological forecasts to some degree. These include "pre-ESP" methods by which the ensemble of historical time series of meteorological data propagated through the hydrological model is selected from years with similar prior observed or forecasted climate indices, or are adjusted based on forecasted temperature and precipitation [*Bierkens and van Beek*, 2009; *Hamlet and Lettenmaier*, 1999; *Werner et al*., 2004]. "Post-ESP" methods follow a similar logic but sample from the ensemble hydrological forecast traces or assign weightings based on climate indices or climate forecasts from GCMs or statistical methods [*Werner et al*., 2004; *Wilks*, 2008]. Post-ESP sampling methods have been found similarly or more successful than pre-ESP methods and computationally more efficient [*Gobena and Gan*, 2010; *Werner et al*., 2004]. Apart from the skill derived from the predicted climate state, seasonal forecast skill is also derived from knowing the initial catchment state. In fact, several recent analyses have demonstrated that this is likely to be the more important contributor to overall skill, at least for regions with winter snow accumulation [*Bierkens and van Beek*, 2009; *Wood and Lettenmaier*, 2006]. This suggests that for at least some environments, it may be possible to produce seasonal forecasts with useful skill in the absence of skilful climate forecasts.

[5] Further impediments to seasonal hydrological forecasting over very large areas are the availability (quantity, quality, and accessibility) of catchment state observations and the accuracy of hydrological models (to estimate initial states in the absence of observations and to forecast their evolution over the forecast period). It is commonly assumed that a hydrological model with carefully calibrated parameters is a prerequisite to produce useable streamflow forecasts [*Shi et al*., 2008]. This assumption appears well-supported when it comes to short-term forecasting, where catchment response time and streamflow retention can have a large impact on the storm hydrograph and thereby on forecast skill. However, several authors questioned the assumption that thorough calibration is a prerequisite to derive seasonal forecasts, for a number of reasons [*Bohn et al*., 2010; *Koster et al*., 2010; *Shi et al*., 2008]. First, at seasonal time scales forecast skill is mostly limited to total streamflow, and therefore, the timing or magnitude of daily streamflow patterns is less important. Second, the interest is often not in absolute forecasts but in the probabilities of streamflow relative to past conditions, e.g., expressed in tercile or above/below median probabilities. Third, postforecast bias correction may be able to compensate suboptimal model calibration [*Shi et al*., 2008]. Avoiding catchment-specific model parameter calibration is attractive, as it is an important obstacle to the generation of forecasts over large areas and many catchments. By contrast, bias correction methods are typically computationally inexpensive.

1.2. Objective

[6] In this study, we evaluate the theoretical and actual skill of a global ensemble streamflow forecasting system. In particular, we wanted to determine (1) whether a comparatively simple global ESP system that uses a computationally efficient post-ESP sampling scheme can provide bimonthly streamflow forecasts with useful skill for at least some regions and seasons and (2) where and when initial conditions and climate forecasts are most likely to lead to useful forecast skill. To our knowledge this is the first time that a global hydrological forecasting system has been developed or evaluated against a large number of small catchments worldwide.

[7] To address these objectives, we performed a retrospective global forecasting experiment using an ESP system with post-ESP sampling. The steps involved in the analysis are illustrated in Figure 2. Briefly, we used gridded meteorological forcing data for the period 1948–2008 [*Sheffield et al*., 2006] and a global hydrological model (the World Wide Water Resources Assessment (W3RA) model, based on the Australian Water Resources Assessment (AWRA) model) [*Van Dijk*, 2010b]. Post-ESP sampling was only done where this was likely to increase skill, based on historic correlations between bimonthly precipitation and 1 out of 21 readily available climate mode indices. Forecasts were generated over a 30 year period (1979–2008), we assume that the skill observed in these

retrospective experiments is a good guide to forecast skill in the (near) future and for readability use the term forecast where retrospective forecast or "hind-cast" might be more accurate. A model run using historical forcing was used as a reference to determine theoretical (or potential) forecast skill. We compared forecast performance with and without post-ESP sampling to estimate the contribution of climate predictability to forecast skill. Finally, we compared theoretical forecast skill to actual skill as measured against streamflow records from several thousand catchments worldwide.

2. Material

2.1. Meteorological Data

[8] Global daily 1° resolution meteorological forcing data for 1948–2008 were available from Princeton University (http://hydrology.princeton.edu). The data sets used included precipitation, downwelling short-wave radiation, and minimum and maximum daily temperature and air pressure. Essentially, these data sets were produced by downscaling observation-based products to finer resolution and shorter time steps using the National Center of Environmental Prediction-National Center for Atmospheric Research reanalysis product and statistical downscaling methods. This ensures that the resulting data have no bias with respect to best available quality observation-based products. Full details on the blending method can be found in *Sheffield et al*. [2006].

2.2. Hydrological Model Structure

[9] The modeling framework used in the experiment is the W3RA system and is based on the landscape hydrology component model of the AWRA system (AWRA-L version 1.0) [*Van Dijk*, 2010b; *Van Dijk and Renzullo*, 2011; *Van Dijk et al*., 2012a]. AWRA-L can be considered a hybrid between a simplified grid-based land surface model and a nonspatial (or so-called "lumped") catchment model applied to individual grid cells. The model was designed to be parsimonious rather than detailed, to support its use where there are few on-ground observations to force and constrain it, as is typical for Australia. Where possible, process equations were selected from the literature and through comparison against observations. The meteorological inputs are gridded daily total precipitation, incoming short-wave radiation, and minimum and maximum temperature, which are converted to daytime effective values [cf. *McVicar and Jupp*, 1999]. Full technical details about the algorithms and default parameters can be found in the model technical documentation [*Van Dijk*, 2010b; http://eos.csiro.au/awra/]. In summary, the configuration considers two hydrological response units (HRUs): deep-rooted tall vegetation ("forest") and shallow-rooted short vegetation ("herbaceous"), each of which occupies a fraction of each grid cell. Vertical processes are described for each HRU individually: (1) the net radiation balance, including incoming and outgoing short-wave and long-wave radiation [*Brutsaert*, 1975] and ground heat flux [*Bastiaanssen et al*., 1998]; (2) partitioning of precipitation between interception evaporation and net precipitation [*Van Dijk and Bruijnzeel*, 2001], and the partitioning of net precipitation between infiltration, infiltration excess surface runoff, and saturation excess runoff [*Van Dijk*, 2010a]; (3) the water balance of three unsaturated soil layers (topsoil, shallow and deep soil layer) including infiltration, drainage (using equations derived from multilayer simulation studies) [see *Peeters et al*., 2013], root water uptake (using a linear ramp function) [*Shuttleworth*, 1992], and soil water evaporation; (4) transpiration, as the lesser of maximum root water uptake and optimum transpiration rate, estimated using the Penman-Monteith equation [*Monteith*, 1965] with aerodynamic conductance estimated from wind speed [*Thom*, 1975] and maximum canopy conductance estimated from model leaf area and remotely sensed greenness [cf. *Yebra et al*., 2013]; (5) groundwater, surface water, and soil evaporation as a linear function of available energy (and for unsaturated soil, relative water content) [*Mutziger et al*., 2005]; (6) vegetation canopy dynamics (leaf biomass, canopy cover, leaf area index, and maximum canopy conductance) that adjust to balance actual and maximum transpiration with a degree of inertia corresponding to vegetation type. In addition, the following integrated catchment processes are described for each grid cell: (7) groundwater dynamics, including recharge from deep drainage, capillary rise (estimated with a linear diffusion equation), evaporation from groundwater saturated areas, and discharge (estimated with a linear reservoir model) [*Peña-Arancibia et al*., 2010; *Van Dijk*, 2009]; and (8) surface water body dynamics, including inflows from runoff and discharge, open water evaporation, and catchment water yield (estimated using a catchment-scale linear routing model) [*Van Dijk*, 2010a].

[10] For the original model version 0.5 [*Van Dijk*, 2010b], prior estimates of all HRU and catchment parameters were derived from the literature or data analysis. For version 1.0, 7 of the 34 parameters for each HRU and 2 catchment parameters were calibrated. The parameters included effective soil parameters determining hydraulic conductivity, water holding capacity and soil evaporation; relating remotely sensed greenness to maximum canopy conductance; and relating groundwater recession and saturated area to catchment characteristics. They were calibrated to optimize agreement with streamflow records from 160 small Australian catchments [*Viney et al*., 2012].

[11] Some modifications of the AWRA-L version 1.0 model were needed for this application. Global data sets were used to configure the model, including tree cover fraction maps [*Hansen et al*., 2003], an albedo climatology derived from Moderate Resolution Imaging Spectroradiometer white-sky albedo [*Moody et al*., 2005] (http://modis-atmos.gsfc.nasa.gov/ALBEDO/) and a wind speed climatology (1983–1993) from NASA (http://eosweb.larc.nasa.gov/sse/). Australia experiences most major climate types (i.e., seasonal, arid, and humid; tropical, temperate, and cool), but significantly snowpack-affected catchments are few, and the model version used did not have a snow hydrology algorithm. Alternative approaches to including these processes were considered, varying from relatively simple conceptual approaches based on the temperature index or degree day concept [e.g., *Lindström et al*., 1997; *Anderson*, 1996] to complex schemes based on explicit description of the energy balance of multiple snow layers [e.g., *Cherkauer and Lettenmaier*, 1999]. Previous assessments suggest that the increased complexity does not

Table 1. Climate Indices Used in Regression Analysis, Earliest Availability[a] and Source for Data Used in the Analysis

Code	Description and Reference	Since	Source
Nino3.4	El Niño Southern Oscillation index 3.4 [*Kaplan et al*., 1998]	1900	http://climexp.knmi.nl/data/inino5.dat
SOI	Southern Oscillation Index, stand Tahiti-stand Darwin [*Troup*, 1965]	1900[a]	http://www.cpc.ncep.noaa.gov/data/indices/
IOD	Indian Ocean Dipole mode index calculated from the Indian Ocean sea surface temperature gradient [*Saji et al*., 1999]	1871	http://www.jamstec.go.jp/frcgc/research/d1/iod/
PC-NAO	Principle-Component-based NAO index [*Hurrell and Deser*, 2009]	1865	http://www.cgd.ucar.edu/cas/jhurrell/indices.data.html
S-NAO	Station-Based NAO index [*Hurrell and Deser*, 2009]	1899	http://www.cgd.ucar.edu/cas/jhurrell/indices.data.html
NAO, EA, WP, EP/NP, PNA, EA/WR, SCA, TNH, POL, PT	Standardized Northern Hemisphere Teleconnection indices, including North Atlantic Oscillation (NAO), East Atlantic Pattern (EA), West Pacific Pattern (WP), East Pacific/North Pacific Pattern (EP/NP), Pacific/North American Pattern (PNA), East Atlantic/West Russia Pattern (EA/WR), Scandinavia Pattern (SCA), Tropical/Northern Hemisphere Pattern (TNH), Polar/Eurasia Pattern (POL), Pacific Transition Pattern (PT)	1950	http://www.cpc.ncep.noaa.gov/data/teledoc/teleintro.shtml
NP	North Pacific index [*Trenberth and Hurrell*, 1994]	1899	http://www.cgd.ucar.edu/cas/jhurrell/indices.data.html
PDO	Pacific Decadal Oscillation index [*Zhang et al*., 1997]	1900	http://jisao.washington.edu/pdo/PDO.latest
SAM	Southern Hemisphere Annular Mode index [*Nan and Li*, 2003]	1948[a]	http://www.lasg.ac.cn/staff/ljp/data-NAM-SAM-NAO/SAM-AAO.htm
STRI, STRL	Mean Southern Hemisphere Subtropical Ridge Intensity (STRI) and Location (STRL) [*Drosdowsky*, 2005]	1900	B. Timbal, Australian Bureau of Meteorology, personal communication
FSS	Full Sun Sunspot number	1900	http://solarscience.msfc.nasa.gov/greenwch/sunspot_area.txt

[a]Some years are missing.

always lead to improved performance and makes greater demands on input data and calibration [*Ferguson*, 1999]. Therefore, in agreement with the overall parsimonious modeling philosophy, the simple but widely tested snow model used in HBV96 [*Bergström*, 1995; *Lindström et al*., 1997] was implemented.

2.3. Prior Evaluation of the Model

[12] AWRA-L is used operationally in the production of water balance information by the Australian Bureau of Meteorology and has been extensively evaluated in that context [*Band*, 2012; *Frost et al*., 2012; *Stenson et al*., 2012]. AWRA-L has been shown to reproduce observed streamflow with accuracy commensurate to or better than achieved by other rainfall-runoff models using a similar calibration approach [*Zhang et al*., 2011; *Viney et al*., 2012; *Van Dijk et al*., 2012a]. Model simulations also compared favorably with other methods when evaluated against in situ observations of soil moisture (I. Dharssi, Bureau of Meteorology; personal communication) and evapotranspiration measured at flux towers [*Van Dijk and Warren*, 2010; *King et al*., 2012]. Model predictions for Australia have been evaluated against remotely sensed time series of total terrestrial water storage [*Van Dijk et al*., 2011; *Tregoning et al*., 2012; *Forootan et al*., 2012; *Van Dijk et al*., 2013], surface soil wetness [*Van Dijk and Warren*, 2010; *Doubkova et al*., 2012], vegetation greenness [*Van Dijk and Warren*, 2010; *Van Dijk et al*., 2013], and vegetation cover fraction, leaf area, and surface albedo [*Van Dijk and Warren*, 2010]. The model has also been used in drought analysis [*Van Dijk et al*., 2013] and in examining the statistical detectability of land cover impacts on streamflow [*Van Dijk et al*., 2012b].

[13] The global implementation described (W3RA) is used in the experimental Asia-Pacific Water Monitor (http://eos.csiro.au/apwm/) but has not yet been evaluated globally with the same scrutiny as its Australian counterpart [*Peña-Arancibia et al*., 2011]. Because of the addition of the snow algorithm, the Australia-focused calibration, and the different model inputs, we performed additional model evaluation experiments as part of this study (section 3.2).

2.4. Climate Indices

[14] Climate index time series were required for the post-ESP analog sampling. Monthly time series of 21 climate indices were obtained from a range of internet sources (Table 1). The key requirements were that they were readily available as long time series and updated with low latency, so that they could be used in an operational system. No prior judgment about the suitability of each individual index was imposed. Some indices represent the same phenomenon and therefore are likely to be strongly correlated. This was not considered a problem, as only the single best performing index was selected for each grid cell × forecast period combination. The likelihood of high correlations occurring by chance was accounted for in the interpretation (see section 3.2). Some of the time series had one or a few years missing. In these cases, the monthly index values were replaced with the average for that month in the years with data available.

2.5. Streamflow Observations

[15] Streamflow observations for bimonthly or shorter integration periods were required for the years 1979–2008. Streamflow data were from the US Geological Survey, Australian state agencies through the Water Information

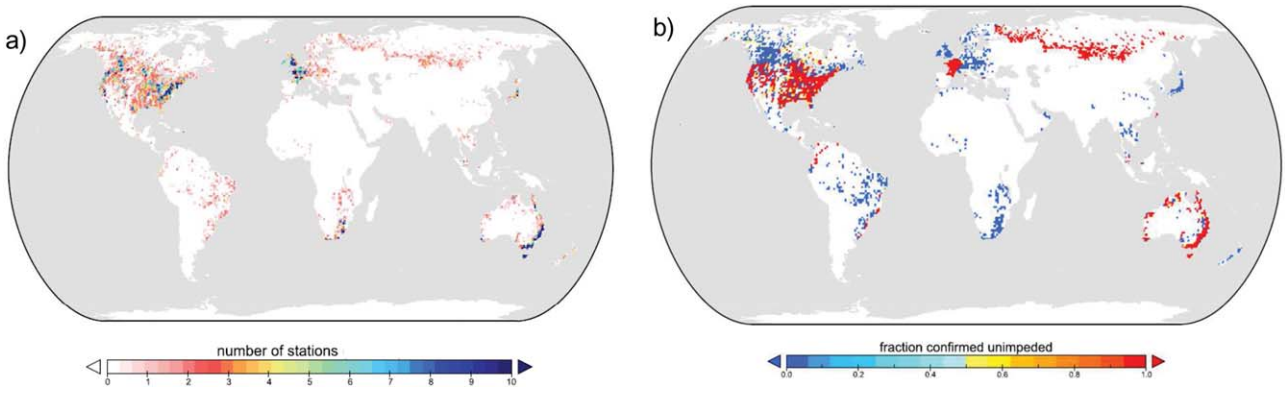

Figure 1. (a) Distribution of the number of stations per 1° grid cell for which streamflow records were suitable for analysis and (b) distribution of suitable catchments that were confirmed to be unimpeded (red) and those unconfirmed (blue).

Research and Development Alliance (WIRADA), the French Ministry of Environment, and a range of additional sources through the Global Runoff Data Centre (GRDC). Ideally, the catchments should be "unimpeded," that is, without significant impoundments or extractive water use. This was confirmed for data from France, Australia, and the United States by others (the latter as a result of the Model Parameter Estimation Experiment) and for northern Eurasia and selected catchments in the humid tropics by us. Some of the catchments may well have experienced some form of land use change; we did not screen for this.

[16] The W3RA grid-based forecasts produced should be considered applicable to catchments of sizes up to around the size of a grid cell (~10,000 km^2) but not for rivers draining much larger regions. This is partly because the model has a lumped representation of river routing within but not beyond grid cells, although streamflow integrated over a bimonthly period might not be affected too much by the streamflow timing implications of this. More importantly, however, the model does not represent river hydrological processes in large river systems such as regulation, reservoir operation, diversion and extraction, and floodplain storage and loss dynamics. Forecasts in large river systems are likely to benefit from representation of these processes as well as initialization with measured system storage and streamflow (i.e., "water already in the system"). In the AWRA system these processes are represented in another component model (AWRA-R) [*Van Dijk et al*., 2012a], but W3RA lacks this capability at a global scale. For that reason, only streamflow data for catchments smaller than 10,000 km^2 were used here. For each forecast period, data for each year and station were considered valid if both months had estimates based on daily records that were more than 70% complete. Skill metrics were calculated for stations with 10 or more valid years in the period 1979–2008. This produced a total of 36,636 station records for the six annual forecast dates, originating from 6192 stations of which 3330 were confirmed unimpeded. The remaining 2862 stations could not be confirmed to be unimpeded; however, they are mainly in humid regions and other regions where a major impact of regulation on local catchment streamflow is less likely (Figure 1; see section 5).

3. Methods

3.1. Overall Approach

[17] Before the forecasting experiment, we assessed the adequacy of the W3RA model structure and configuration for the experiment by comparing its performance in explaining observed streamflow records to that of peer models (section 3.2). The steps in our forecast analysis are illustrated in Figure 2, briefly summarized here, and subsequently described in greater detail in the following sections. First, a regression analysis was carried out to determine suitable climate predictor variables for use in the post-ESP sampling scheme (described in section 3.3). Second, we produced probabilistic forecasts for the period 1979–2008 (section 3.4; the core of the forecasting system is shown as the shaded boxes in Figure 2). Third, we retrospectively measured the *theoretical forecast skill* against the synthetic streamflow estimates produced by the reference model run with historic forcing, and the *actual forecast skill* against observed streamflow records for the 6192 catchments (section 3.5).

3.2. Evaluation of the Forecast Model Against Peer Models

[18] Retrospective streamflow estimates from W3RA for the period 1979–2008 were compared to records for 6192 small catchments (cf. section 2.5). We compared performance to that of four hydrological models (CLM, Mosaic, NOAH, and VIC) implemented in the Global Land Data Assimilation System (GLDAS) [*Rodell et al*., 2004], on the basis that GLDAS simulations have been extensively evaluated and widely published (see http://ldas.gsfc.nasa.gov/gldas/GLDASpublications.php) and therefore can be deemed an acceptable benchmark. However, because model performance is the combined result of model code, parameter values, and forcing data, any conclusions are only strictly valid for these specific model configurations. The forcing data are different between GLDAS and W3RA, and the model forcing, code version, and parameter values of the GLDAS models may differ from those used in other published studies. Technical details on the GLDAS model configurations are provided on the NASA Land Data

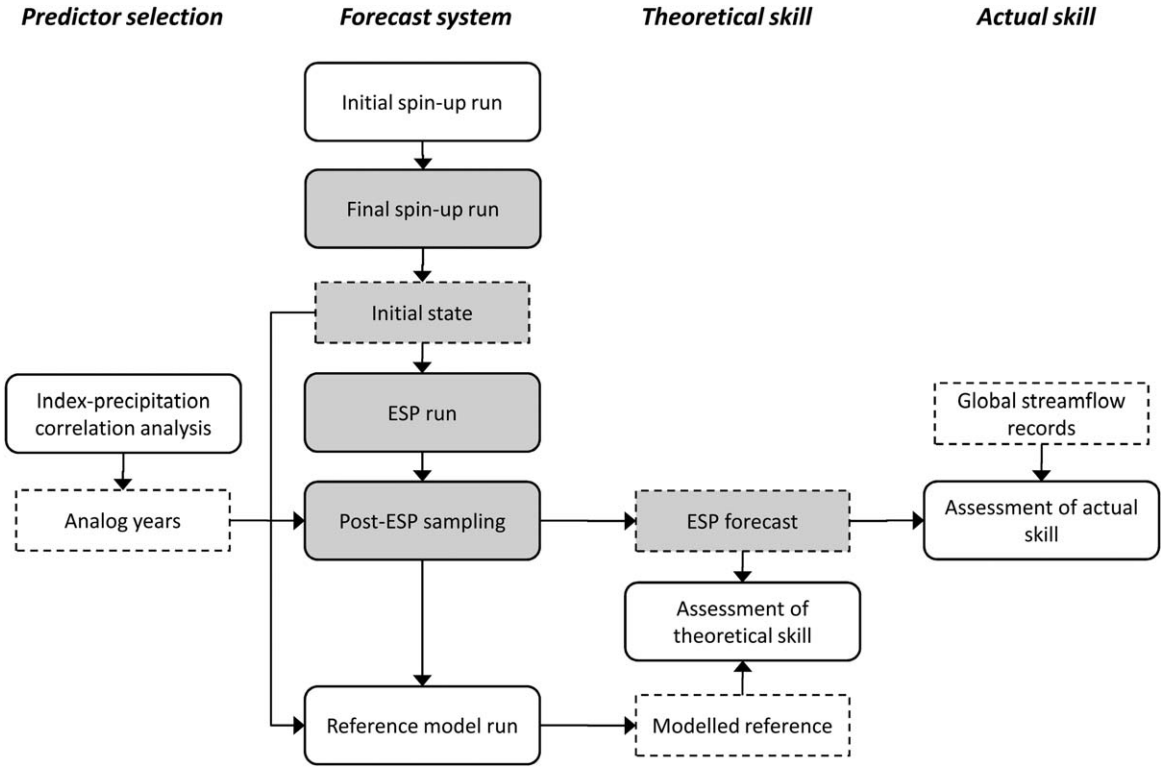

Figure 2. Diagram illustrating the forecasting experiment (shaded shapes) and the associated analyses (clear shapes) carried out as part of this study.

Assimilation Systems web site (http://ldas.gsfc.nasa.gov/gldas/index.php); the data used here are global 1° resolution monthly estimates of surface and subsurface runoff rate (downloaded in June 2012 from http://disc.sci.gsfc.nasa.gov/hydrology/data-holdings; data code "GLDAS_<model name>_10M"). The two streamflow components were summed and converted to mm d^{-1}.

[19] All streamflow stations with more than 10 years of data between 1979 and 2008 were considered for use in this analysis. Most of the streamflow records were reported in m^3 s^{-1}, which were converted to mm d^{-1} using reported catchment areas. For a number of stations, the average streamflow exceeded average precipitation for the corresponding period. Within limits this may be explained by precipitation spatial variability and estimation error, but where mean streamflow exceeded precipitation by more than 2.5 times these were considered likely to have errors in the reported streamflow units or catchment area and were excluded from the analysis.

[20] The agreement between estimated and observed streamflow was calculated for each of the models and compared in terms of mean streamflow and relative bias across stations, as well as the distribution of the parametric (Pearson's r) and ranked (nonparametric) correlation coefficient (Spearman's ρ) between modeled and observed monthly streamflow for all stations.

3.3. Climate Predictor Variable Selection and Post-ESP Sampling Strategy

[21] The post-ESP sampling method used here reflects methods used in the current generation of ESP systems [e.g., *Werner et al*., 2004] but considers a larger set of climate indices. We used regression analysis to determine which of the individual 21 available climate indices (if any) should be used to condition the post-ESP sampling scheme. The squared correlation coefficient (r^2) was calculated between each of the 21 available climate indices (Table 1) reported in the month prior to the forecast period, and precipitation for the 2 month forecasting period. This analysis was performed separately for each grid cell and forecast period for 1950–2008, for which data were available for all indices ($N = 59$). It is noted that the period used to develop the post-ESP scheme includes the forecasting period. This was necessary to obtain a statistically meaningful sample but can potentially lead to overestimation of skill enhancements. Conversely, the nature and strength of predictive relationships can vary over time, and therefore, analyzing relationships over the entire 59 year period can lead to underestimation of skill enhancements. These caveats need to be considered in interpreting the results.

[22] The post-ESP sampling procedure consist of reducing the full ensemble of forecasts based on historic analogues to some smaller number that represents those years during which the climate index value (i.e., prior to the date of forecast) was closest to the current index value (i.e., prior to the actual forecast time). In other words, a subset of historical forcing years is selected for which the rainfall predictor had a similar value. Because of the large number of grid cell × period combinations on one hand, and the limited temporal sample size for each of them ($N = 59$) on the other, statistically seemingly significant r^2 values will be calculated by chance (see section 4.1). Furthermore, even where a predictor index is appropriately selected, sampling too few ensemble members can lead to loss of forecast skill and reliability,

for example, because of outliers. The two choices interact, since a high r^2 value can justify sampling a smaller subset of ensemble members. We performed preliminary trials to assess the influence of different choices of correlation threshold and sample size on forecast skill. On this basis, we conservatively chose a correlation threshold of $r^2 = 0.20$ and where this threshold was exceeded sampled 10 out of the 30 member forecast ensemble.

3.4. Forecasting System

[23] Forecasts of global bimonthly streamflow were generated for the period 1979–2008. The forecasting method has four steps (Figure 2):

[24] (i) *Initial spin-up*. The hydrological model was initialized with default states, and the full 61 years of forcing data (1948–2008) were used to spin up initial states. This long spin-up was applied because some initial states (e.g., deep soil water and ground water in arid regions) required some decades to reach dynamic equilibrium.

[25] (ii) *Final spin-up*. Using the states reached at the end of 2008, the model was rerun from 1948 to the forecast date.

[26] (iii) *Ensemble forecast*. An ensemble of analog forcing time series for the 2 month forecast period was derived from the 30 years prior to the forecast year. Each ensemble member was propagated through the hydrological model, starting with the initial model states reached under step (2).

[27] (iv) *Ensemble sampling*. If, for the grid cell and forecast period considered, climate indices for the preceding month had useful predictive skill, post-ESP sampling was applied.

[28] Forecasts were generated for six dates (the first of January, March, May, July, September, and November) and over a 30 year period (1979–2008), creating a total of 180 forecast ensembles for each grid cell. The following example further illustrates the procedure: consider a forecast on 1 March 1979 for a grid cell where Nino3.4 has useful predictive value (i.e., correlation between February Nino3.4 values and March-April precipitation exceeds $r^2 > 0.20$). At step (2) the final spin-up was run with forcing for the period 1 January 1948 to 28 February 1979. At step (3) the model was run with the 1 March to 30 April forcing for each of the 30 years prior to the forecast year (1949–1978). At step (4) the 10 years were identified for the period 1949–1978 for which February Nino34 values were numerically closest to the Nino3.4 value for February 1979. The resulting 10 total streamflow forecasts for March-April make up the forecast ensemble. For grid cells where none of the climate indices showed r^2 greater than 0.20, forcing data for all of the 30 preceding years (1949–1978) were used. As a result, for some grid cells and forecast periods the final forecast ensemble had 10 members; for the remainder it had 30 members.

3.5. Skill Assessment

[29] We compared all forecasts against the corresponding reference model streamflow estimates (Figure 2). The modeled reference was generated by using the same initial state as estimated for the forecast date, but running it subsequently with the historic climate estimates for the actual forecast period. The primary skill metric used for each grid cell and forecast period was the ranked correlation coefficient (Spearman's ρ) between the ensemble median and reference estimate, as well as the parametric correlation coefficient (Pearson's r) between the median forecast and the reference. Two probabilistic skill metrics were also calculated: the Ranked Probability Skill Score (RPSS) [*Wilks*, 1995] and the revised Linear Error in Probability Space score (LEPS) [*Potts et al*., 1996]. The RPSS was chosen for its familiarity among the US seasonal streamflow forecasting community and was applied following *Franz et al*. [2003] (i.e., using 10%, 30%, 70%, and 90% nonexceedance probability thresholds). The LEPS was calculated because of its slightly different characteristics and use in the meteorological community. For all metrics, the 30 year streamflow climatology shows a skill score of zero and a perfect forecast a score of one. It will be shown further on that the different skill metrics are all highly correlated. For some grid cell × period combinations, theoretical skill was very high, even though the mean and variance in model streamflow were extremely small. This occurred particularly in arid regions and can be considered a modeling artifact that creates a misleadingly high estimate of theoretical skill. To avoid this, the calculated skill for grid cell × forecast period combinations with an interannual standard deviation in period-average streamflow less than 0.01 mm d^{-1} was set to zero.

[30] To summarize forecast skill over the six forecast periods, the simple mean and maximum were calculated for each skill metric, as well as the weighted average value, with weightings based on the variance in reference streamflow estimates. For example, the variance-weighted average ranked correlation ($\hat{\rho}$) is given by

$$\hat{\rho} = \sum_{i=1}^{6} [\rho\, var(y)]_i / \sum_{i=1}^{6} [var(y)]_{i,} \quad (1)$$

where $[\mathrm{var}(y)]_i$ represents the variance of streamflow in the model-estimated reference streamflow for forecasting period i ($N = 30$). This procedure was followed for ρ, LEPS, and RPSS. However, rather than calculating the variance-weighted r, the fraction of overall variance in bimonthly streamflow anomalies (i.e., streamflow minus climatology) explained by the median forecast was calculated, which is numerically equal to replacing ρ in equation (1) with r^2. The contribution of climate indices to theoretical skill was calculated as the difference in skill with and without post-ESP sampling, respectively. The ρ score was chosen to compare theoretical skill to the actual skill of the forecasts, as it was considered likely to be least sensitive to the relatively small number of (not necessarily contiguous) years of record available for some stations (N varied from 10 to 30 years). The same calculations were carried out as for theoretical skill, but this time using recorded streamflow. As a measure of (retrospective) model performance, the ρ and r values between the model-estimated reference streamflow and recorded streamflow for each forecast period were also calculated.

4. Results

4.1. Evaluation of the Forecast Model Against Peer Models

[31] The cumulative distribution of the relative bias between mean modeled and observed streamflow for the

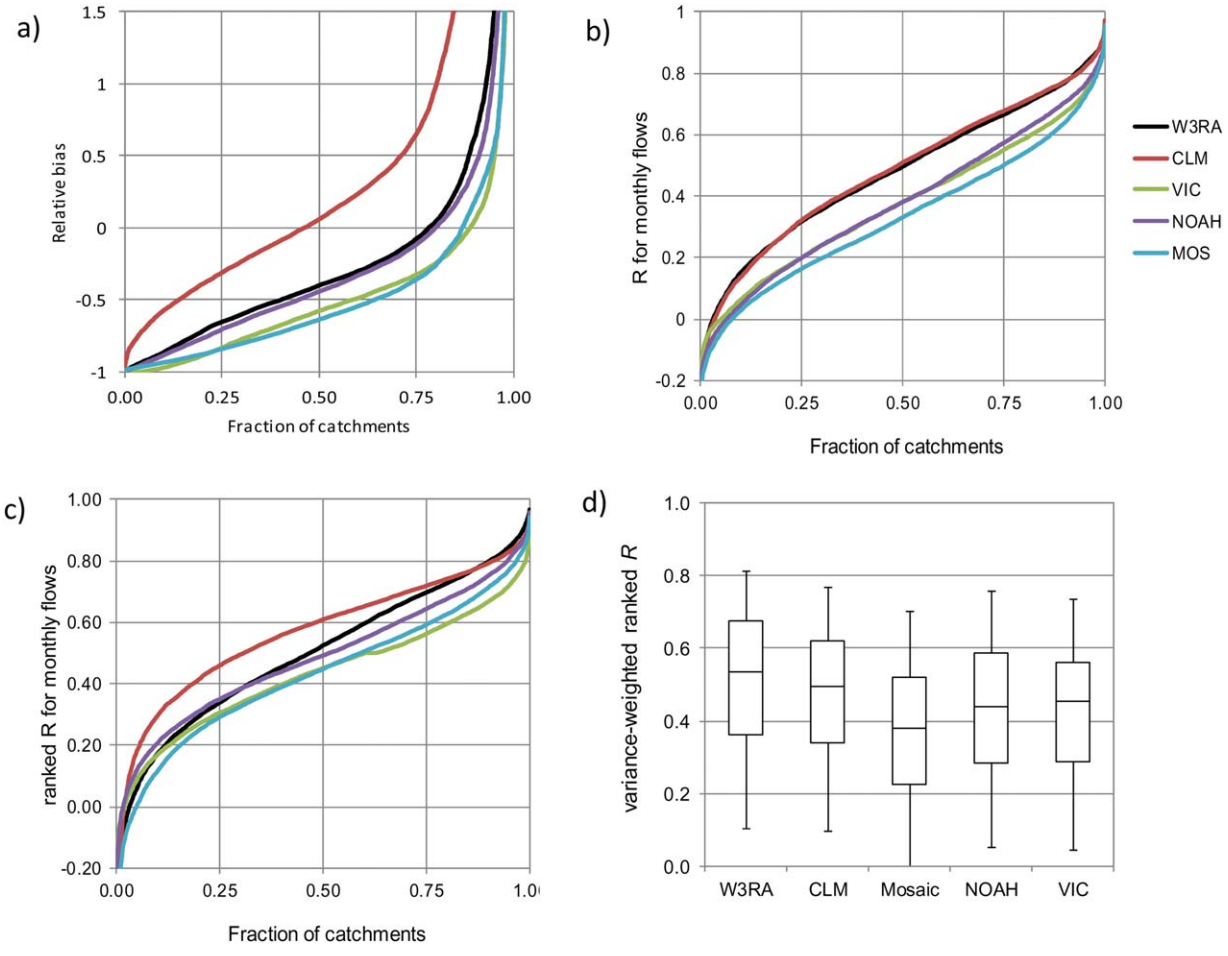

Figure 3. Cumulative distribution functions for (a) relative bias between modeled and observed streamflow, (b) parametric correlation coefficient (Pearson's r) for monthly streamflow, (c) ranked correlation coefficient (Spearman's ρ) for monthly streamflow, and (d) box plots showing the 5th, 25th, 50th, 75th, and 95th percentiles of variance-weighted average $\hat{\rho}$ for bimonthly streamflow.

W3RA and the four GLDAS models is shown in Figure 3a for the global streamflow data set. A linear regression of the form $y = ax$ (where y is the observed and x the model-estimated mean catchment streamflow) produced values for the slope a of 1.43 for W3RA ($r^2 = 0.60$), compared to 1.28 for CLM ($r^2 = 0.51$), 1.77 for Mosaic ($r^2 = 0.37$), 1.56 for NOAH ($r^2 = 0.56$), and 1.77 for VIC ($r^2 = 0.49$; scatter plots provided in the supporting information). W3RA estimates were within $\pm 50\%$ of the recorded values for 48% of catchments. This fraction was higher for CLM (57%) and lower for VIC (37%), NOAH (47%), and Mosaic (30%).

[32] In terms of parametric r for monthly flows, the performance of W3RA was near identical to that of CLM and better than that of the other three models (Figure 3b). In terms of ranked ρ for monthly flows, W3RA performed less than CLM, similar to NOAH, and slightly better than VIC and Mosaic (Figure 3c). In terms of $\hat{\rho}$ (cf. equation (1)) W3RA performed better than any of the GLDAS models (Figure 3d; median 54% versus 37%–49%).

[33] To better understand in what environments W3RA performs significantly better or worse than the benchmark set by the best of the four GLDAS models, $\hat{\rho}$ values for individual catchments were mapped to a global grid (assigning the average $\hat{\rho}$ value where a grid cell had more than one catchment within it). Of all grid cells with at least one catchment ($N = 2722$), W3RA showed the best $\hat{\rho}$ performance for 41% of cells, CLM for 26%, VIC for 18%, NOAH for 11%, and Mosaic for 4% (Figure 4a). W3RA performance is particularly good in cool and temperate regions but appears worse than that of VIC in the drier central regions of the United States and Canada.

4.2. Precipitation Predictor Variables

[34] The fractions of grid cell $\times$ period combinations showing correlation (r^2) greater than 0.20, 0.30, and 0.40 are shown in Figure 5a. The frequency by which these r^2 threshold would be expected to be reached by random chance can be estimated via the two-tailed significance test as 2.7×10^{-4}, 4×10^{-6}, and $<1 \times 10^{-7}$, respectively.

[35] Several indices show greater frequencies of exceedance than these expected values. The most powerful predictor variables at this time scale were those describing the ENSO (Nino3.4 and SOI) and the Indian Ocean Dipole (IOD; Figures 6a and 6b). The spatial correlation patterns associated with these two modes are shown in Figures 6a and 6b. Less powerful predictors were some of the Pacific Ocean indices (PDO, EP/NP, and WP), Southern

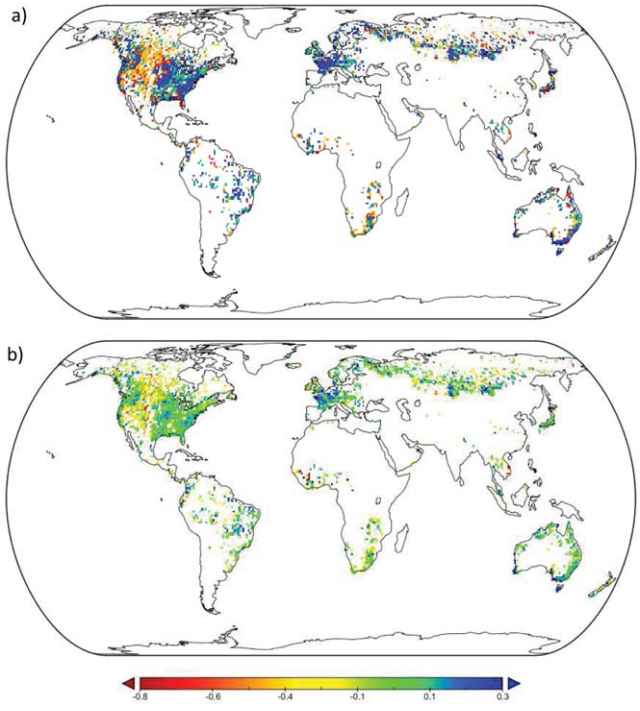

Figure 4. Maps showing (a) the best performing model (in terms of $\hat{\rho}$) for each grid cell with streamflow data (dark blue, W3RA; aqua, CLM; bright green, Mosaic; orange, VIC; red, NOAH) and (b) the difference in performance ($\hat{\rho}$) between W3RA and the best among the four GLDAS models for each grid cell.

Hemisphere Subtropical Ridge indices (STRlat and STRint), and sunspot numbers. The number of forecasts for which $r^2 > 0.20$ for each index is shown in Figure 5b, and the spatial distribution of the number of forecasts for which post-ESP sampling was applied is shown in Figure 6c.

[36] On average, this was the case for only 2.6% of forecasts. The highest overall correlations and the largest number of months with post-ESP sampling were for parts of Indonesia and the Philippines and the north of South America and were associated with ENSO. Climate indices also appeared to potentially add skill in eastern Australia, east Africa, and Uruguay and surrounding regions, associated with ENSO and IOD.

4.3. Theoretical Skill

[37] There was strong relative agreement between the four skill metrics, with linear regression equations of $r = 0.98\rho$ ($r^2 = 0.90$), $\mathrm{RPSS} = 0.70\rho^2$ ($r^2 = 0.92$), and $\mathrm{LEPS} = 0.67\rho^2$ ($r^2 = 0.83$). The ρ scores for individual forecast periods are shown in Figure 7; maps for the other metrics are provided in the supporting information.

[38] Significance level maps would potentially be misleading [cf. *Mason*, 2008], but the ρ corresponding to different significance levels may be considered in interpretation: $\rho > 0.24$ for $p = 0.1$, $\rho > 0.30$ for $p = 0.05$, and $\rho > 0.42$ for $p = 0.01$ (one-tailed test, $N = 30$). Theoretical skill is generally significant in cold, temperate, and humid tropical zones and in seasonally wet regions during and after the wet season. Skill was insignificant for arid regions and in seasonally wet regions during the dry season. The highest skills ($\rho > 0.8$) were found for the winter and snowmelt period in cold regions, the transition from wet to dry season in seasonally wet regions, and the Indonesian region. The global distribution of summary metrics of overall theoretical skill over the six periods is also similar for all metrics and shows the greatest mean theoretical skill for boreal regions and the tropical monsoon regions (Figure 8).

[39] The contribution of post-ESP sampling to theoretical skill was generally small (Figure 9). All skill metrics were enhanced in equatorial Asia (Indonesia, the Philippines) and South America (parts of Colombia, northeast Brazil, the Guyanas), the pampas region (Uruguay and central Argentina), and a smaller region around Mumbai, India. For some metrics and dates, there appear to be meaningful contributions for eastern Australia, East Africa, southwest United States, and Spain (Figure 9b). Contributions in other regions are much less, and in some cases the reduction of the forecast ensemble led to an apparent deterioration in theoretical skill. For the forecasts with post-ESP sampling, median ρ increased from 0.68 to 0.71 (median increase +0.02). Median r increased from 0.67 to 0.71 (median increase +0.02). The median probabilistic skill scores increased slightly more: from 0.23 to 0.26 (median increase +0.01) for RPSS and from 0.25 to 0.31 (median increase +0.04) for LEPS. Hence consideration of climate appears to have a slightly more beneficial effect on probabilistic forecast skill (+11%–26%) than on deterministic forecast skill (+4%–6%).

4.4. Actual Forecast Skill

[40] The spatial distribution of actual skill and retrospective model performance given the estimated historic forcing (both estimated from ρ) show some clear spatial patterns (Figure 10). The following observations can be made:

[41] (i) The highest actual forecast skill ($\rho > 0.4$) was calculated for stations in the northwestern and southeastern United States, the Upper Mississippi basin, western France, northwest Russia and south-central Siberia, northern South America, and southeast Australia (Figure 10a). These regions show typical theoretical skill levels of $\rho > 0.5$ (Figure 10b) of which more than 70% was realized (Figure 10c). This coincides with relatively high retrospective model performance ($\rho > 0.7$; Figure 10d).

[42] (ii) Low actual forecast skill ($\rho < 0.4$) despite high theoretical skill ($\rho > 0.8$) occurred in the more humid parts of Canada, Scandinavia, the Alps, and remaining parts of northern Russia (Figures 10a and 10b), where the fraction of realized skill was correspondingly low (<30%; Figure 10c).

[43] (iii) The spatial pattern in the fraction of realized skill (Figure 10c) qualitatively agrees with the spatial pattern in retrospective model performance, that is, the agreement between the reference model estimates and actual observed streamflow (Figure 10d).

[44] (iv) Dry regions (e.g., in central north America, Siberia, the Middle East, inland Australia, and Southern Africa) sometimes show actual skill that exceeds theoretical skill (Figure 10c), even though model performance is poor ($\rho < 0.4$; Figure 10d). Because theoretical skill is very low, the actual skill is still very low in these cases however, with a few isolated, possibly coincidental exceptions (e.g., in Israel and Saudi Arabia; Figures 10a and 10b).

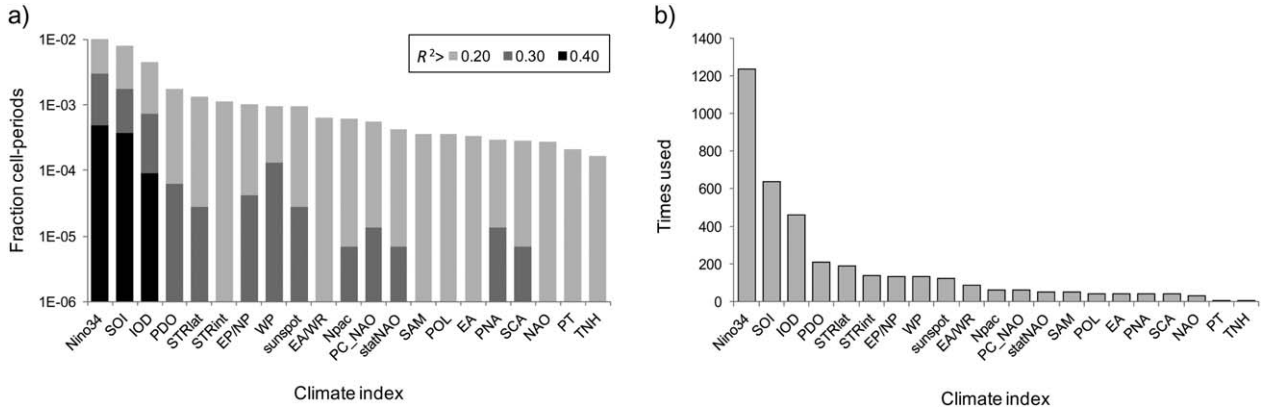

Figure 5. (a) Frequency of exceedance of different r^2 thresholds for the 21 climate mode indices tested for each grid cell × forecast period combination; and (b) number of grid cell × forecast period combinations for which each climate mode index was used in post-ESP sampling (see Table 1 for meaning of acronyms).

[45] (v) Regions with low realized skill levels (<30%) occur in all humid and subhumid climate zones. This includes regions with both relatively low theoretical skill ($\rho < 0.5$; e.g., southern Brazil, southern Africa, and Ireland) and relatively high theoretical skill (Amazonia, Southeast Asia, Canadian Rocky Mountains, eastern Canada, Scandinavia, central Europe, and Britain; Figures 10b and 10c). Generally, these patterns again agree with patterns in retrospective model performance (Figure 10d).

[46] The overall relationship between theoretical and actual skill, and between retrospective performance and realized skill, is shown in Figures 11a and 11b, respectively.

[47] Actual skill increases with theoretical skill, but the fraction of theoretical skill that is realized decreases with increasing theoretical skill (Figure 11a). Average actual skill slightly exceeds theoretical skill for very low skill values. It falls away rapidly for very high theoretical skill values; the associated stations are at polar latitudes (Canada and Scandinavia) and in the humid tropics (Colombia, Amazonia, and Borneo). The median theoretical and actual forecast skill (ρ) across all stations and forecast periods were 0.54 (intersextile range, ISR 0.23–0.80) and 0.30 (ISR 0.01–0.54), respectively. The median ratio between them (the fraction of theoretical skill realized) was 0.58 (ISR 0.03–1.21). The median realized skill was higher when only using the catchments confirmed to be unimpeded (0.71, ISR 0.24–1.34) when compared to catchments not confirmed to be unimpeded (0.40, ISR −0.17–0.93; Figure 11a). There are probably other reasons underlying this difference, however (see section 5). On average, there is a strong and near-proportional relationship between retrospective model performance and realized skill (Figure 11b). Median model performance (ρ) was 0.52 (ISR 0.24–0.74) but slightly greater for catchments confirmed to be unimpeded (0.62, ISR 0.36–0.78) and lower for unconfirmed catchments (0.42, ISR 0.15–0.64).

5. Discussion

5.1. Forecast Model Performance Compared to Peer Models

[48] Overall, W3RA appears to perform similarly well or perhaps slightly better than the four GLDAS models. It was not possible to determine whether the observed performance differences were due to the different hydrological model (version) or forcing data [cf. *Zaitchik et al*., 2010]. A comparative evaluation of W3RA and the four GLDAS models against the global streamflow records showed some regions where the W3RA forecast system could be improved; in particular, the dry interior of the United States and Canada, where VIC performed better. This may be because of better quality forcing for the GLDAS models or because the VIC model was originally developed and calibrated for these types of environments; for example, the degree day approach used in W3RA may not describe the melting of thin snowpacks on open plains well. Further research is needed to determine the main cause and whether it is best addressed by adjusting (calibrating) model parameters or by improving model structure. Arguably, a diverse multimodel, multiforcing ensemble system is a preferable approach to dealing with deficiencies in individual forcing data and models [cf. *Zaitchik et al*., 2010] but would obviously be more challenging to implement and maintain.

5.2. Theoretical Skill

[49] Our results suggest considerable skill in seasonal forecasts, in theory, for many parts of the world for at least some time of the year. We found the greatest overall theoretical skill (Figures 7 and 9) under the following circumstances:

[50] (1) Locations where snowpack accumulates during part of the year [cf. *Koster et al*., 2010; *Wood and Lettenmaier*, 2006]. Such regions include the continental and boreal regions of North America and Eurasia but also the Himalayan and Andean highlands. For some of these regions, theoretical skill was high not only in the snowmelt season but also during other seasons (Figure 7), presumably due to the predictive value of soil wetness conditions.

[51] (2) Locations with a distinct wet and dry season. In this case, estimates of water stored in the soil and groundwater system provide initial conditions for predicting streamflow during the transition months between the wet and dry season. Such regions include the monsoon regions

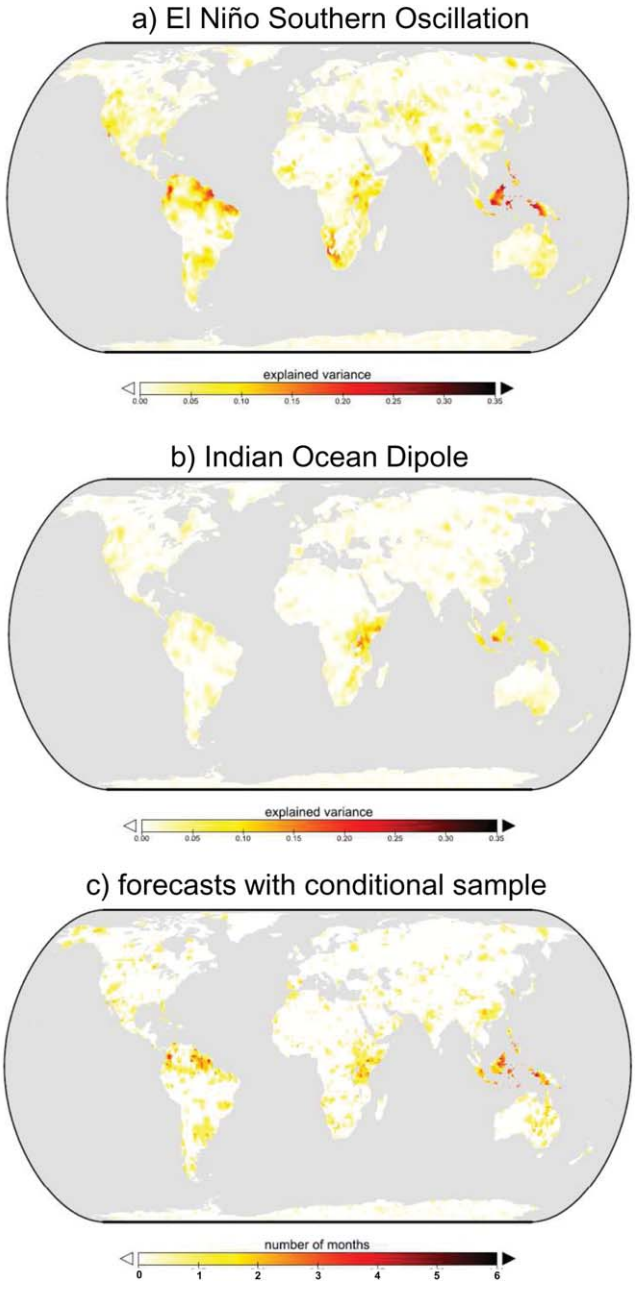

Figure 6. Total variance in bimonthly precipitation anomalies explained by indices of (a) ENSO and (b) the Indian Ocean Dipole. (c) Number of months for which post-ESP sampling was applied.

and seasonal subhumid regions of South America, Africa, southern Asia, and northern Australia.

[52] (3) For catchments with precipitation patterns that are strongly correlated to ENSO, the strength of this relationship appeared to provide a sufficiently strong basis for seasonal forecasting. This was found, in particular, in the equatorial zones of South America and Asia.

[53] (4) For catchments with high precipitation, including the Amazonian and Indonesian regions. This may be explained by the modeled delayed release of anomalously high precipitation prior to the forecast date.

5.3. Contribution of Initial Climate State

[54] Post-ESP sampling was used for 2.6% of forecasts only. Most of the climate indices tested appeared to enhance theoretical skill for a few areas and periods, although often only marginally so. The greatest skill increases occurred in ENSO-affected regions in equatorial South America and Southeast Asia. We cannot conclude that the various other climate modes are not valuable for seasonal forecasting, or that our postsampling procedure provides an upper estimate of the current skill contribution possible by considering climate modes. First, several indices showed minor contributions to overall skill but relatively strong local correlation with precipitation ($r^2 > 0.30$) for certain seasons (e.g., ENSO in the northern Caribbean for January-February; IOD in southern China in September-October; PDO in northeast Australia and nearby Pacific Islands from March-June; NAO in Spain in January-February; data not shown). Where these seasons did not show high streamflow variance compared to other seasons, the higher skill for that period (ρ) did not translate into higher overall skill ($\hat{\rho}$; cf. equation (1)).

[55] Second, our sampling method only considered one climate index for each grid cell and forecast period, whereas we found evidence that more than one index showed theoretical skill in a small number of cases (e.g., Nino34 and IOD for western Indonesia, cf. Figures 6a and 6b). Considering this, the theoretical skill achieved for the Indonesian region seems to bode well for the potential of streamflow forecasting using methods that can consider multiple indices in these areas [see, e.g., *Hamlet and Lettenmaier*, 1999].

[56] Third, we used a very simple post-ESP sampling procedure that made use of (average) climate indices observed during the month prior to the forecast. Therefore, a strong correlation between climate mode index and precipitation will only enhance skill if there is persistence in the climate mode, that is, if there is sufficiently strong autocorrelation between the climate index anomaly in month i and the anomalies in months $i+1$ and $i+2$, respectively. We calculated these correlations for each period and found that with 1 or 2 months lead time, persistence is only moderate to strong ($0.30 < r^2 < 0.98$) for Nino3.4, IOD, sunspot numbers, PDO (all months for both lead times), SOI (9 months), and SAM (10 and 6 months for 1 and 2 months lead time, respectively). By comparison, the maximum autocorrelation among NAO indices was $r^2 = 0.15$ (PC-NAO between January and February) and $r^2 = 0.20$ among any of the other northern hemisphere indices (PNA for January-February). The lack of persistence may explain why we were unable to derive added skill from NAO indices, whereas *Bierkens and van Beek* [2009] found small skill enhancements in some parts of Europe (but deteriorations in other parts); those authors used the NAO forecasts from *Rodwell and Folland* [2002] which are not directly based on autocorrelation.

[57] Fourth, related to this, we did not use seasonal climate mode forecasts. In addition to NAO, these forecasts are also available for ENSO [*Van Oldenborgh et al.*, 2005]. At present, the skill of most statistical and dynamic methods appears commensurate and generally fairly low when it comes to seasonal precipitation forecasts [*Bierkens and van Beek*, 2009; *Rodwell and Doblas-Reyes*, 2006], but this may change in future. Similarly, we did not examine

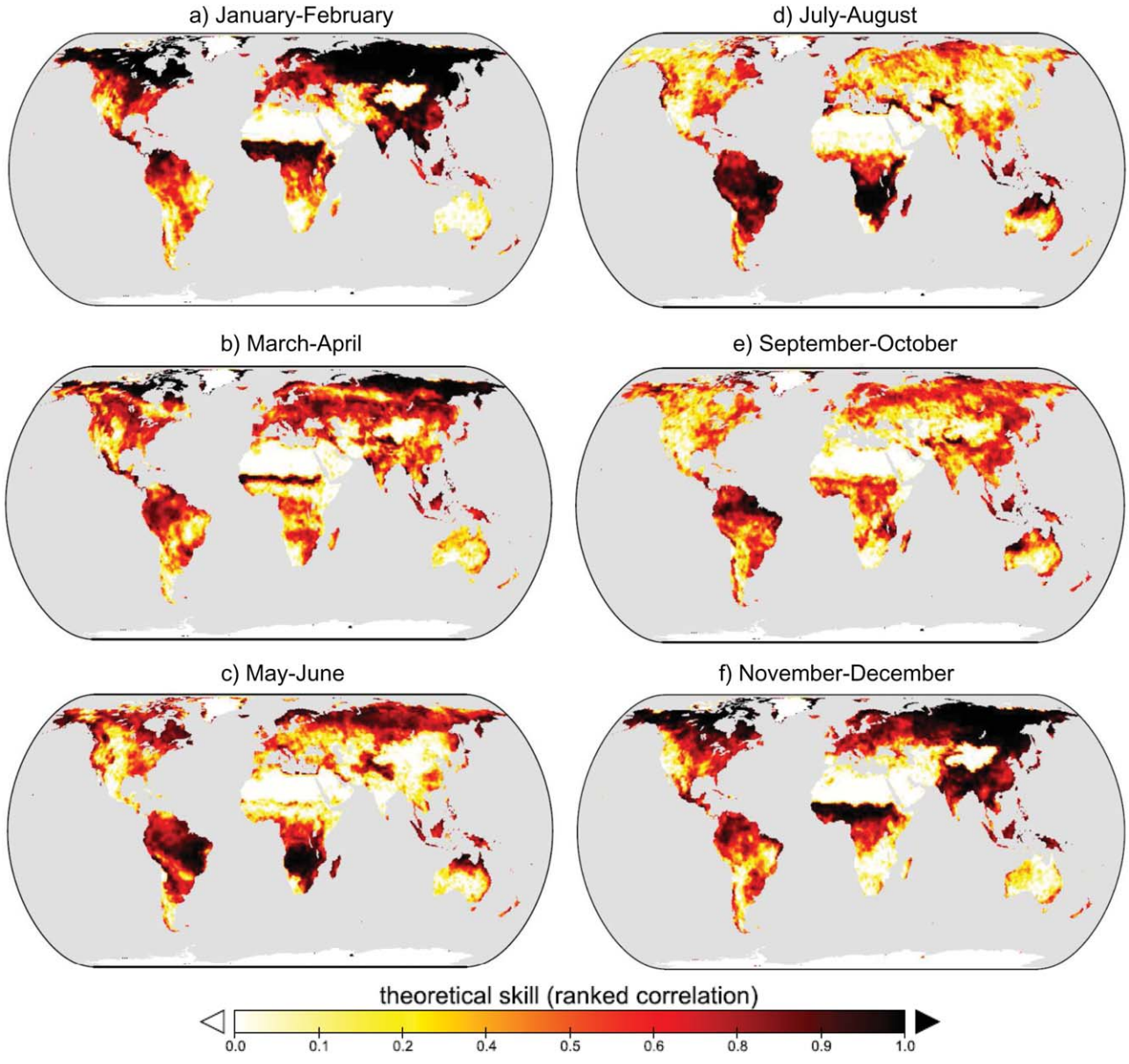

Figure 7. Theoretical prediction skill for the (a–f) six forecast periods, calculated as the ranked correlation (ρ) between forecast ensemble median and model-estimated reference streamflow.

alternative statistical approaches to seasonal climate forecasting. Methods developed to extract predictive value from sea surface temperature or other predictor fields include those based on canonical correlation analysis [*Barnston*, 1994] and principal or independent component analysis [*Westra and Sharma*, 2010; *Westra et al*., 2009] or varimax rotation [*Richman*, 1986]. Any such forecasts should be straightforward to trial in ESP sampling provided they are routinely available with low latency.

[58] Finally, we did not use numerical weather prediction (NWP) forecasts at all. Although NWP skill for precipitation rapidly disappears after around 2 weeks [*Yuan et al*., 2011], that still accounts for about one fifth of the 2 month forecast period and therefore may still increase the skill of cumulative streamflow forecasts. Further analysis is needed to assess to what extent the contribution of climate forecasts from NWP and climate mode forecasts can enhance skill. However, the results of our analysis suggest that for the majority of regions and periods, most of the skill comes from knowing the initial state, at least at two monthly time scales. This confirms results found for the United States [*Shukla and Lettenmaier*, 2011] and Europe [*Bierkens and van Beek*, 2009].

5.4. Actual Forecast Skill

[59] Actual forecast skill was measured against streamflow records from 6192 catchments worldwide. We found that the realized skill was on average 54% of the theoretical skill. We are aware of one other study where the degradation from theoretical to actual forecast skill was estimated, although a direct comparison is not possible due to the different catchment size ranges, different lead times, and

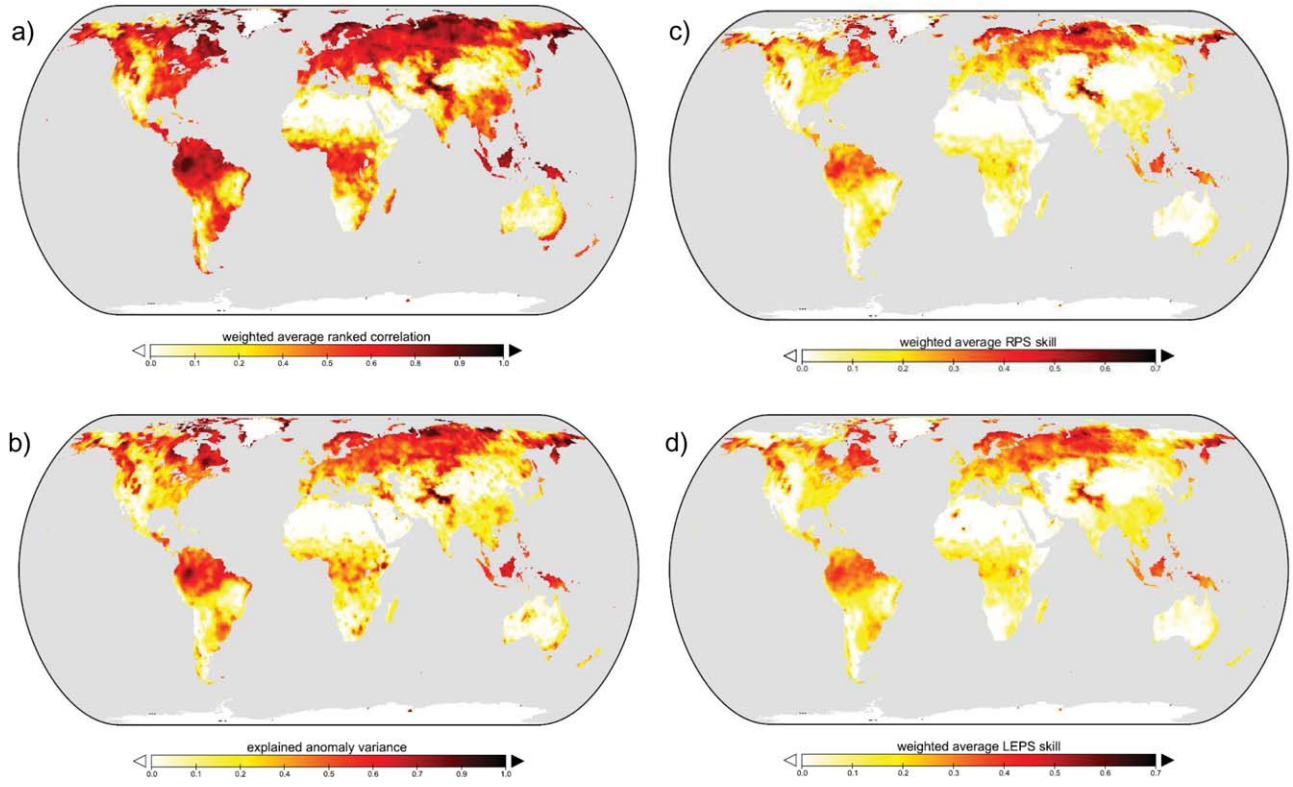

Figure 8. Summary metrics of theoretical skill over the six forecast periods: (a) streamflow variance-weighted ranked correlation (ρ), (b) the total variance in bimonthly streamflow explained, and streamflow variance-weighted (c) RPS and (d) LEPS skill.

different environments involved. *Bierkens and van Beek* [2009] analyzed the theoretical and actual skill of 6 month winter and summer streamflow forecasts for 66 large European catchments (18,500–1,360,000 km^2). They found an average actual skill ρ of 0.31–0.34 for winter and 0.09–0.17 for summer, representing around 50%–70% of theoretical skill. *Koster et al*. [2010] performed a somewhat similar experiment but for 5 month forecasts of spring streamflow for 17 catchments (1863–1,357,667 km^2) in the American west. The forecast showed actual skill in terms of r^2 of up to 0.5 (median 0.21). Koster et al. did not report theoretical skill values, but for the modeled reference they report that r^2 against observed flows was 0.3–0.9 (median 0.73), which suggests a realized skill of around 30%.

[60] We found a strong, near-proportional relationship between retrospective model performance and realized skill. There are several a priori reasons why one would not expect to realize theoretical skill. The forcing data have limitations and errors. There are scaling errors between the 1° resolution and the smaller catchments used in verification. The model structure is an imperfect representation of catchment response. The model parameters may not be optimally specified. The streamflow records have errors themselves. Each of these factors can be expected to affect both retrospective model performance and realized skill in similar ways.

[61] Average model performance and realized skill were greater for catchments that were confirmed to be unimpeded than for catchments for which this was not (yet) confirmed. This may have depressed overall realized skill somewhat, particular for heavily developed catchments, e.g., in Europe (outside France) and Japan. However, the catchments are generally small (<10,000 km^2), and therefore, the influence of large-scale water extraction, impoundments, and regulation is likely to be less than that downstream in large river systems. Other causes for the different statistics for the two data sources are feasible. In particular, we cannot exclude the possibility that the "unconfirmed" catchments outside Europe were mainly located in regions where the model forcing data were of lesser quality (cf. Figure 1b).

[62] The resolution and quality of the 1° forcing data and precipitation in particular are likely factors for all stations. Despite advances in precipitation reanalysis and satellite remote sensing, the quality and resolution of global precipitation data sets are still fundamentally limited by the density, quality, and availability of precipitation gauge measurements. The catchments considered here are small in comparison to the resolution of the forcing (median 899 km^2, ISR 152–4049 km^2). This was necessary, primarily because we wanted to compare skill against catchments in which streamflow is unaffected by downstream processes, but it may also have introduced scaling errors. In a previous analysis [*Van Dijk and Warren*, 2010], we found that the correlation between estimated and recorded streamflow increases as the data are aggregated over more than one catchment: from a median $r^2 = 0.5$ for individual catchments, to ca. $r^2 = 0.7$ for data aggregated over up to 100

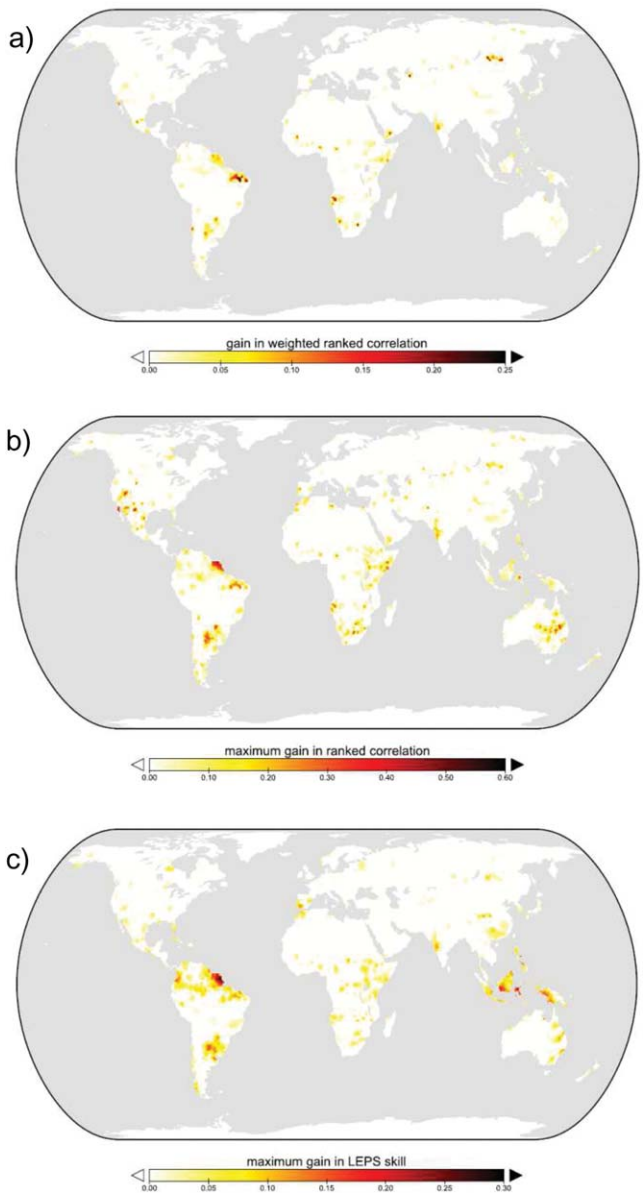

Figure 9. Contribution of climate indices to theoretical skill, calculated as the difference between (a) the variance-weighted the mean and (b) ranked correlation coefficient and (c) LEPS skill for the case with (cf. Figure 8a) and without ensemble sampling based on climate index, respectively.

catchments. This was attributed not only to scale but also the averaging out of streamflow data errors. Moreover, ρ only improved very slightly and therefore seems more robust to scaling. We did not find a statistical relationship between actual skill (ρ) and catchment size for the catchments used in this study.

[63] The model version used here was also implemented in the AWRA system [*Viney et al*., 2012] but using climate forcing at a much higher spatial resolution (0.05°) and based on a larger number of gauges available for Australia [*Jones et al*., 2009]. This provides an opportunity to estimate the influence of the global forcing data used. Measured in the exactly same way (i.e., using ρ for same bimonthly periods and unimpeded catchments, $N = 3515$), the continental model showed a median performance of 0.82 (ISR 0.57–0.93) and the global model used here a median performance of 0.63 (ISR 0.44–0.78). Therefore, the resolution and perhaps quality of the global forcing data appeared to contribute to at least half of the difference between actual and perfect model performance, and hence the unrealized forecast skill. The accuracy of the AWRA estimates has been shown equivalent to other commonly used catchment streamflow models with parameters estimated a priori (in those cases from nearby catchments), but about half of the remaining disagreement in monthly streamflow can be removed through local calibration against the actual streamflow record itself, at least in terms of r^2 [*Viney et al*., 2012; *Zhang et al*., 2011]. Therefore, the lack of model calibration may be responsible for another quarter or so of the unrealized forecast skill. The remaining unrealized skill would be due to a combination of errors in the high-resolution forcing data (the density of the Australian gauging network is highly variable), the model structure, and undetected but inevitable errors in the streamflow records used.

[64] We speculate that the generally poor model performance and realized skill in humid tropical regions are also primarily due to the lack of precipitation gauges and therefore poor quality of precipitation estimates in these regions. Higher-quality and -resolution gridded precipitation data are available for some, typically more densely populated countries and therefore would seem likely to allow enhanced forecast skill. To some extent, precipitation estimates in poorly gauged regions can be improved using remote sensing and NWP outputs [*Ebert et al*., 2007]. The other meteorological variables are likely to be of second-order importance and can arguably be derived with sufficient accuracy from the available products, at least for most locations.

[65] The main sensitivity of forecasts to climate data is probably due to their influence on estimated initial hydrological conditions, rather than due to the forcing data ensemble for the forecast period. Better estimates of initial states can also be achieved by assimilating more direct observations of these. This includes ground-based observations (e.g., of streamflow, snow depth, and groundwater level), but the accessibility and latency of these data sources are likely to be prohibitive at global scale, at least in an operational sense. Remote sensing offers an alternative source of observations. Total water storage estimates derived from missions like the Gravity Recovery and Climate Experiment [*Tapley et al*., 2004] provide a unique opportunity to observe initial catchment state over large areas in a way that can be combined with models [*Van Dijk et al*., 2011a]. Optical and passive and active microwave remote sensing can also help to improve model estimates of initial snow cover and shallow soil moisture state [*Andreadis and Lettenmaier*, 2006; *de Jeu et al*., 2008; *Liu et al*., 2012].

5.5. Prospects for a Global Seasonal Streamflow Forecasting System

[66] We demonstrated that it is currently feasible to construct a global system to forecast seasonal streamflow. The resulting forecasts are likely to have lesser skill than forecasting systems that use purpose-calibrated models and

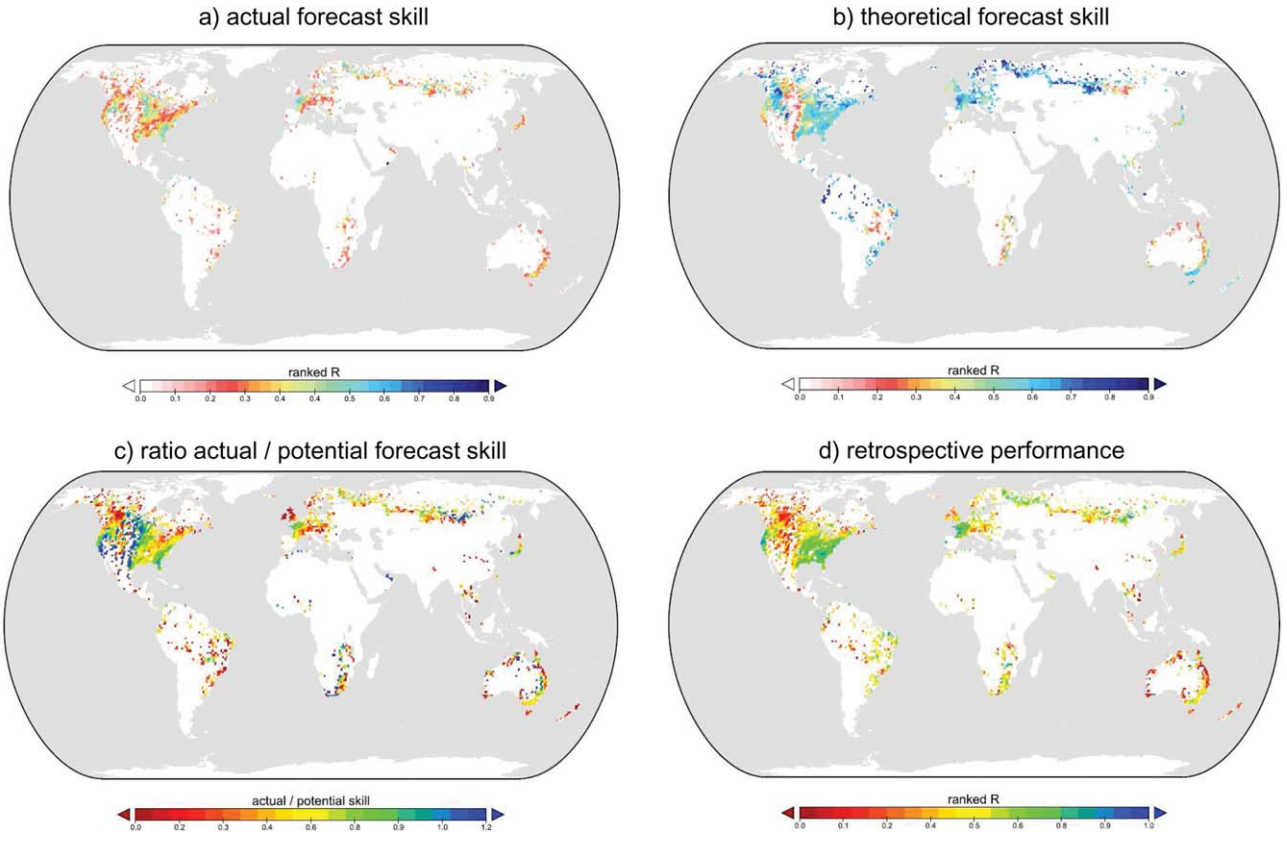

Figure 10. (a) Actual forecast skill against observed streamflow, (b) theoretical forecast skill, (c) realized fraction of theoretical skill (i.e., the ratio of actual over theoretical skill), and (d) retrospective model performance against observed streamflow, given the estimated historic forcing. The metric used in all cases is streamflow variance-weighted ρ (for grid cells with several stations the median ρ for all stations is shown, cf. Figure 1).

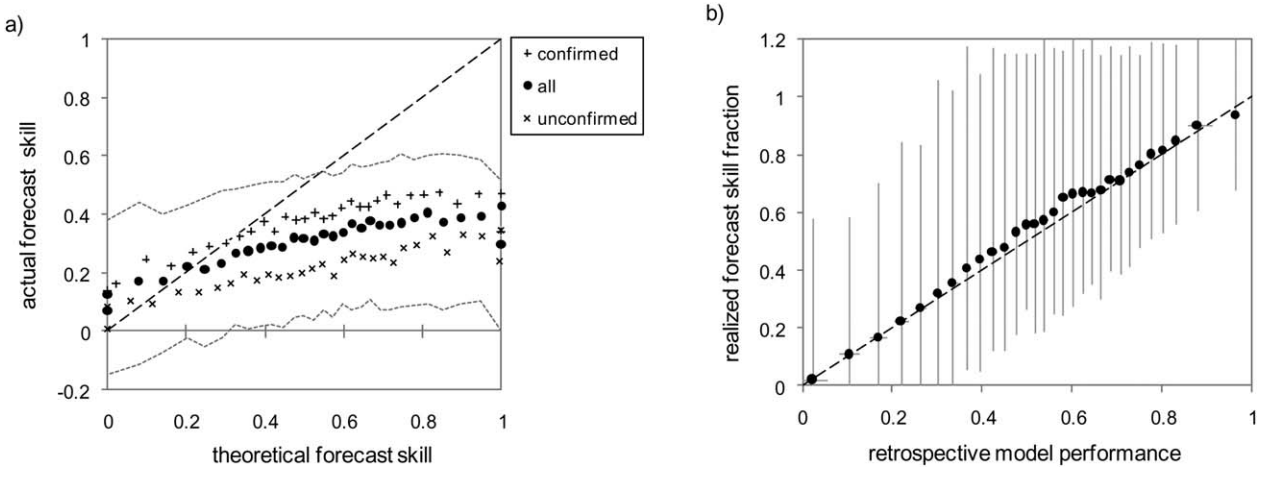

Figure 11. (a) Relationship between theoretical and actual forecast skill based on ranked correlation, showing slight differences in realized skill between stations in catchments confirmed (pluses) and unconfirmed (crosses) to be unimpeded (each symbol shows the median of 567–1221 station-forecast combinations, ordered by theoretical forecast skill; closed dots and dashed line show median and intersextile range for all confirmed and unconfirmed stations combined). (b) Relationship between retrospective performance and realized skill, expressed as the ratio of actual over theoretical forecast skill; in both cases based on flow variance-weighted average ranked correlation (each symbol shows the median of 1169 station-forecast period combinations ordered by model performance; lines indicate the intersextile ranges).

better observations of precipitation and initial catchment conditions. However, where such systems are not available, a large-scale system such as that presented here may still provide useful streamflow forecasts for certain regions and times of year.

[67] This study is only an initial assessment using one possible system and configuration. For operational implementation of systems such as that presented here, the latency of all data sources would be a critical issue. Moreover, many alternative approaches and configurations are possible. This can include alternative models and forcing data or ensembles of these; assimilation of observed initial conditions; and the use of weather forecasts and seasonal climate predictions. Further improvements in actual seasonal forecast skill may therefore be reasonably expected. The streamflow observations and forecasts produced in this study may serve as a benchmark for future (re)forecast experiments and are available for the Global Energy and Water Cycle Experiment Hydrological Applications Project (GEWEX HAP) Seasonal Forecasting working group activities.

6. Conclusions

[68] We estimated the theoretically achievable skill in bimonthly average streamflow forecasting using an ESP system with post-ESP sampling based on years with similar climate state. Theoretical skill was compared to actual forecast skill calculated for each of the forecast times at 6192 streamflow stations. The following conclusions are drawn:

[69] Significant theoretical skill in bimonthly forecasts is largely due to initial conditions where and when streamflow variations are dominated by snow melt and delayed release of prior precipitation (e.g., in monsoonal regions).

[70] Modest skill is added by considering ENSO conditions for equatorial regions in South America and Southeast Asia, with somewhat more benefit for probabilistic (+11%–24%) than for deterministic median forecasts (+4%–6%).

[71] The actual skill was approximately 54% of the theoretical skill. We could attribute more than half of the unrealized skill to the quality and resolution of the global precipitation data, another quarter to the lack of model calibration, and the remainder to a combination of errors in streamflow measurement and model structure.

[72] A global forecasting system would appear feasible. Skill for any region and period can be estimated as the product of theoretical skill and retrospective model performance.

[73] Further increases in seasonal forecast skill are likely to occur primarily from improvements in observing and assimilating initial conditions (snow, soil, and groundwater storage), but currently available weather forecasts and seasonal climate predictions may also enhance skill in some cases.

[74] **Acknowledgments.** This research was part of the GEWEX HAP Seasonal Forecasting working group activities. The streamflow data were made available by the GRDC and the Ministry of Environment of France. Ulrich Loose and Ludovic Oudin are thanked for assisting us in accessing the GRDC and French streamflow data. Tara Troy, Neil Viney, and Yongqiang Zhang are thanked for collating and cleaning the streamflow data sets for northern Eurasia and Australia, respectively. The GLDAS data used in this study were acquired as part of the mission of NASA's Earth Science Division and archived and distributed by the Goddard Earth Sciences (GES) Data and Information Services Center (DISC). A.I.J.M.v.D. and J.L.P.A. were supported through the WIRADA between the Bureau of Meteorology Water Division and CSIRO's Water for a Healthy Country Flagship Program.

References

Anderson, E. (1996), *II.2 Snow-17 Snow Model, National Weather Service River Forecasting System (NWSRFS) User's Manual,* NWS Hydrol. Lab., Silver Springs, Maryland, U. S.

Andreadis, K. M., and D. P. Lettenmaier (2006), Assimilating remotely sensed snow observations into a macroscale hydrology model, *Adv. Water Res.*, *29*(6), 872–886.

Band, L. E. (2012), Commentary on the progress of the Australian Water Resources Assessment development, in *Proceedings Water Information Research and Development Alliance Science Symposium*, Melbourne, Australia, August. [Available at http://www.csiro.au/WIRADA-Science-Symposium-Proceedings.]

Barnston, A. G. (1994), Linear statistical short-term climate predictive skill in the Northern Hemisphere, *J. Clim.*, *7*, 1513–1564.

Bastiaanssen, W. G. M., M. Menenti, R. A. Feddes, and A. A. M. Holtslag (1998), A remote sensing surface energy balance algorithm for land (SEBAL): 1. Formulation, *J. Hydrol.*, *212*, 198–212.

Bergström, S. (1995), The HBV Model, in *Computer Models of Watershed Hydrology*, edited by V. P. Singh, pp. 443–476, Water Resour. Publ., Highlands Ranch, Colo.

Bierkens, M. F. P., and L. P. H. van Beek (2009), Seasonal predictability of European discharge: NAO and hydrological response time, *J. Hydrometeorol.*, *10*(4), 953–968.

Bohn, T. J., M. Y. Sonessa, and D. P. Lettenmaier (2010), Seasonal hydrologic forecasting: Do multi-model ensemble averages always yield improvements in forecast skill?, *J. Hydrometeorol.*, *11*(6), 1358–1372.

Brutsaert, W. (1975), On a derivable formula for long-wave radiation from clear skies, *Water Resour. Res.*, *11*(5), 742–744.

Cherkauer, K. A., and D. P. Lettenmaier (1999), Hydrologic effects of frozen soils in the upper Mississippi, *J. Geophys. Res.*, *104*(D16), 19,599–19,610.

Cherry, J., H. Cullen, M. Visbeck, A. Small, and C. Uvo (2005), Impacts of the North Atlantic Oscillation on Scandinavian hydropower production and energy markets, *Water Resour. Manage.*, *19*(6), 673–691.

Chiew, F. H. S., S. L. Zhou, and T. A. McMahon (2003), Use of seasonal streamflow forecasts in water resources management, *J. Hydrol.*, *270*(1–2), 135–144.

Day, G. (1985), Extended streamflow forecasting using NWSRFS, *J. Water Resour. Plann. Manage.*, *111*, 157–170.

de Jeu, R., W. Wagner, T. Holmes, A. Dolman, N. van de Giesen, and J. Friesen (2008), Global soil moisture patterns observed by space borne microwave radiometers and scatterometers, *Surv. Geophys.*, *29*(4), 399–420.

Doubkova, M., A. I. J. M. van Dijk, D. Sabel, W. Wagner, and G. Bloeschl (2012), Evaluation of the predicted error of the soil moisture retrieval from C-band SAR by comparison against modelled soil moisture estimates over Australia, *Remote Sens. Environ.*, *120*, 188–196.

Drosdowsky, W. (2005), The latitude of the subtropical ridge over eastern Australia: The L index revisited, *Int. J. Clim.*, *25*(10), 1291–1299.

Ebert, E. E., J. E. Janowiak, and C. Kidd (2007), Comparison of near-real-time precipitation estimates from satellite observations and numerical models, *Bull. Am. Meteorol. Soc.*, *88*(1), 47–64.

Feng, X., T. DelSole, and P. Houser (2011), Bootstrap estimated seasonal potential predictability of global temperature and precipitation, *Geophys. Res. Lett.*, *38*, L07702, doi:10.100/2010GL046511.

Ferguson, R. I. (1999), Snowmelt runoff models, *Prog. Phys. Geogr.*, *23*(2), 205–227.

Forootan, E., J. Awange, J. Kusche, B. Heck, and A. Eicker (2012), Independent patterns of water mass anomalies over Australia from satellite data and models, *Remote Sens. Environ.*, *124*, 427–443.

Franz, K. J., H. C. Hartmann, S. Sorooshian, and R. Bales (2003), Verification of National Weather Service ensemble streamflow predictions for water supply forecasting in the Colorado River basin, *J. Hydrometeorol.*, *4*(6), 1105–1118.

Frost, A., et al. (2012), Australian water balance assessment: operational challenges, in *Proceedings Water Information Research*

and Development Alliance Science Symposium, Melbourne, Australia, August. [Available at: http://www.csiro.au/WIRADA-Science-Symposium-Proceedings.]

Gobena, A. K., and T. Y. Gan (2010), Incorporation of seasonal climate forecasts in the ensemble streamflow prediction system, *J. Hydrol.*, *385*(1–4), 336–352.

Hamlet, A. F., and D. P. Lettenmaier (1999), Columbia River streamflow forecasting based on ENSO and PDO climate signals, *J. Water Resour. Plann. Manage.*, *125*(6), 333–341.

Hamlet, A. F., D. Huppert, and D. P. Lettenmaier (2002), Economic value of long-lead streamflow forecasts for Columbia River hydropower, *J. Water Resour. Plann. Manage.*, *128*(2), 91–101.

Hansen, M. C., R. S. DeFries, J. R. G. Townshend, M. Carroll, C. Dimiceli, and R. A. Sohlberg (2003), Global percent tree cover at a spatial resolution of 500 meters: First results of the MODIS vegetation continuous fields algorithm, *Earth Interact.*, *7*(10), 1–15.

Hurrell, J. W., and C. Deser (2009), North Atlantic climate variability: The role of the North Atlantic Oscillation, *J. Mar. Syst.*, *78*(1), 28–41.

Jones, D., W. Wang, and R. Fawcett (2009), High-quality spatial climate data-sets for Australia, *Aust. Meteorol. Oceanogr. J.*, *58*, 233–248.

Kaplan, A., M. A. Cane, Y. Kushnir, A. C. Clement, M. B. Blumenthal, and B. Rajagopalan (1998), Analyses of global sea surface temperature 1856–1991, *J. Geophys. Res.*, *103*(C9), 18,567–18,589.

King, E. A., et al. (2012), An operational actual evapotranspiration product for Australia, in *Proceedings Water Information Research and Development Alliance Science Symposium, Melbourne*, Australia, August. [Available at http://www.csiro.au/WIRADA-Science-Symposium-Proceedings.]

Koster, R. D., S. P. P. Mahanama, B. Livneh, D. P. Lettenmaier, and R. H. Reichle (2010), Skill in streamflow forecasts derived from large-scale estimates of soil moisture and snow, *Nat. Geosci.*, *3*(9), 613–616.

Lavers, D., L. Luo, and E. F. Wood (2009), A multiple model assessment of seasonal climate forecast skill for applications, *Geophys. Res. Lett.*, *36*, L23711, doi:10.1029/2009GL041365.

Lindström, G., B. Johansson, M. Persson, M. Gardelin, and S. Bergström (1997), Development and test of the distributed HBV-96 hydrological model, *J. Hydrol.*, *201*(1–4), 272–288.

Liu, Y. Y., W. A. Dorigo, R. M. Parinussa, R. A. M. De Jeu, W. Wagner, M. F. McCabe, J. P. Evans, A. I. J. M. van Dijk (2012), Trend-preserving blending of passive and active microwave soil moisture retrievals, *Rem. Sens. Env.*, *123*, 280–297.

Luo, L., and E. F. Wood (2008), Use of Bayesian merging techniques in a multimodel seasonal hydrologic ensemble prediction system for the eastern United States, *J. Hydrometeorol.*, *9*(5), 866–884.

Mason, S. J. (2008), Understanding forecast verification statistics, *Meteorol. Appl.*, *15*(1), 31–40.

McVicar, T. R., and D. L. B. Jupp (1999), Estimating one-time-of-day meteorological data from standard daily data as inputs to thermal remote sensing based energy balance models, *Agric. For. Meteorol.*, *96*(4), 219–238.

Monteith, J. L. (1965), Evaporation and environment, *Symp. Soc. Exp. Biol.*, *19*, 205–224.

Moody, E. G., M. D. King, S. Platnick, C. B. Schaaf, and G. Feng (2005), Spatially complete global spectral surface albedos: Value-added datasets derived from Terra MODIS land products, *IEEE Trans. Geosci. Remote. Sens.*, *43*(1), 144–158.

Mutziger, A. J., C. M. Burt, D. J. Howes, and R. G. Allen (2005), Comparison of measured and FAO-56 modeled evaporation from bare soil, *J. Irrig. Drain. Eng.*, *131*, 59–72.

Nan, S., and J. Li (2003), The relationship between the summer precipitation in the Yangtze River valley and the boreal spring Southern Hemisphere annular mode, *Geophys. Res. Lett.*, *30*(24), 2266, doi:10.1029/2003GL018381.

Pagano, T., D. Garen, and S. Sorooshian (2004), Evaluation of official western US seasonal water supply outlooks, 1922–2002, *J. Hydrometeorol.*, *5*(5), 896–909.

Palmer, T. N., and D. L. T. Anderson (1994), The prospects for seasonal forecasting—A review paper, *Q. J. R. Meteorol. Soc.*, *120*(518), 755–793.

Pappenberger, F., J. Thielen, and M. Del Medico (2011), The impact of weather forecast improvements on large scale hydrology: Analysing a decade of forecasts of the European Flood Alert System, *Hydrol. Processes*, *25*(7), 1091–1113.

Peeters, L., R. Crosbie, R. Doble, and A. Van Dijk (2013), Conceptual evaluation of continental land-surface model behaviour, *Environ. Model. Softw.*, *43*, 49–59.

Peña-Arancibia, J., A. van Dijk, M. Mulligan, and L. Bruijnzeel (2010), The role of climatic and terrain attributes in estimating baseflow recession in tropical catchments, *Hydrol. Earth Sys. Sci.*, *14*(11), 2193–2205.

Peña Arancibia, J. L., A. I. J. M. Van Dijk, M. P. Stenson, and N. R. Viney (2011), Opportunities to evaluate a landscape hydrological model (AWRA-L) using global data sets, in *Proceedings of the 19th International Congress on Modelling and Simulation MODSIM2011,* edited by F. Chan, D. Marinova, and R. S. Anderssen, Modell. and Simul. Soc. of Aust. and N. Z.December [Available at http://www.mssanz.org.au/modsim2011/I11/pena_arancibia.pdf.]

Potts, J. M., C. K. Folland, I. T. Jolliffe, and D. Sexton (1996), Revised ''LEPS'' scores for assessing climate model simulations and long-range forecasts, *J. Clim.*, *9*, 34–53.

Richman, M. B. (1986), Rotation of principal components, *J. Climatol.*, *6*, 293–335.

Ritchie, J. W., C. Zammit, and D. Beal (2004), Can seasonal climate forecasting assist in catchment water management decision-making?: A case study of the Border Rivers catchment in Australia, *Agric. Ecosyst. Environ.*, *104*(3), 553–565.

Rodell, M., P. Houser, U. Jambor, J. Gottschalck, K. Mitchell, C. Meng, K. Arsenault, B. Cosgrove, J. Radakovich, and M. Bosilovich (2004), The Global Land Data Assimilation System, *Bull. Am. Meteorol. Soc.*, *85*(3), 381–394.

Rodwell, M. J., and F. J. Doblas-Reyes (2006), Medium-range, monthly, and seasonal prediction for Europe and the use of forecast information, *J. Clim.*, *19*(23), 6025–6046.

Rodwell, M. J., and C. K. Folland (2002), Atlantic air–sea interaction and seasonal predictability, *Q. J. R. Meteorol. Soc.*, *128*(583), 1413–1443.

Saha, S., et al. (2006), The NCEP climate forecast system, *J. Clim.*, *19*(15), 3483–3517.

Saji, N. H., B. N. Goswami, P. N. Vinayachandran, and T. Yamagata (1999), A dipole mode in the tropical Indian Ocean, *Nature*, *401*(6751), 360–363.

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.

Sheffield, J., G. Goteti, and E. F. Wood (2006), Development of a 50-year high-resolution global dataset of meteorological forcings for land surface modeling, *J. Clim.*, *19*(13), 3088–3111.

Shi, X., A. W. Wood, and D. P. Lettenmaier (2008), How essential is hydrologic model calibration to seasonal streamflow forecasting?, *J. Hydrometeorol.*, *9*(6), 1350–1363.

Shukla, S., and D. P. Lettenmaier (2011), Seasonal hydrologic prediction in the United States: Understanding the role of initial hydrologic conditions and seasonal climate forecast skill, *Hydrol. Earth Syst. Sci.*, *15*(11), 3529–3538.

Shuttleworth, W. J. (1992), Evaporation, in *Handbook of Hydrology*, edited by D. R. Maidment, pp. 4.1–4.53, McGraw-Hill.

Stenson, M. P., et al. (2012), Operationalising the Australian Water Resources Assessment (AWRA) system, in *Proceedings Water Information Research and Development Alliance Science Symposium*, Melbourne, Australia, August. [Available at http://www.csiro.au/WIRADA-Science-Symposium-Proceedings.]

Tapley, B. D., S. Bettadpur, J. C. Ries, P. F. Thompson, and M. M. Watkins (2004), GRACE measurements of mass variability in the earth system, *Science*, *305*(5683), 503–505.

Thom, A. S. (1975), 3. Momentum, mass and heat exchange of plant communities, in: JL Monteith (Ed.), *Vegetation and the Atmosphere: Principles*, p. 57, Academic Press, London.

Tregoning, P., S. McClusky, A. I. J. M. Van Dijk, R. Crosbie, and J. L. Pena-Arancibia (2012), Assessment of GRACE satellites for groundwater estimation in Australia, Waterlines Rep. Ser. 71. Natl. Water Comm., Canberra. [Available at http://nwc.gov.au/__data/assets/pdf_file/0003/21468/71-Assessment-of-GRACE-satellites-for-groundwater-estimation-in-Australia.pdf.]

Trenberth, K. E., and J. W. Hurrell (1994), Decadal atmosphere-ocean variations in the Pacific, *Clim. Dyn.*, *9*, 303–319.

Troup, A. J. (1965), The 'southern oscillation,' Q. J. R. *Meteorol. Soc.*, *91*(390), 490–506.

Van Dijk, A. I. J. M. (2009), Climate and terrain factors explaining streamflow response and recession in Australian catchments, *Hydrol. Earth Syst. Sci.*, *14*, 159–169.

Van Dijk, A. I. J. M. (2010a), Selection of an appropriately simple storm runoff model, *Hydrol. Earth Syst. Sci.*, *14*(3), 447–458.

Van Dijk, A. I. J. M. (2010b), AWRA Technical Report 3, Landscape Model (Version 0.5) Technical Description, WIRADA/CSIRO Water for a Healthy Country Flagship, Canberra. [Available at http://www.clw.csiro.au/publications/waterforahealthycountry/2010/wfhc-aus-water-resources-assessment-system.pdf.]

Van Dijk, A. I. J. M., and L. A. Bruijnzeel (2001), Modelling rainfall interception by vegetation of variable density using an adapted analytical model. Part 1: Model description, *J. Hydrol.*, *247*(3–4), 230–238.

Van Dijk, A. I. J. M., and L. J. Renzullo (2011), Water resource monitoring systems and the role of satellite observations, *Hydrol. Earth Syst. Sci.*, *15*, 39–55.

Van Dijk, A. I. J. M., and G. A. Warren (2010), *AWRA Technical Report 4, Evaluation Against Observations,* WIRADA/CSIRO Water for a Healthy Country Flagship, Canberra. [Available at http://www.clw.csiro.au/publications/waterforahealthycountry/2010/wfhc-awras-evaluation-against-observations.pdf.]

Van Dijk, A. I. J. M., L. J. Renzullo, and M. Rodell (2011), GRACE water storage estimates help evaluate and improve the Australian water resources assessment system, *Water Resour. Res.*, *47*, W11524, doi:10.1029/2011WR010714.

Van Dijk, A. I. J. M., et al. (2012a), Design and development of the Australian Water Resources Assessment system, in *Proceedings Water Information Research and Development Alliance Science Symposium*, Melbourne, Australia, August. [Available at http://www.csiro.au/WIRADA-Science-Symposium-Proceedings.]

Van Dijk, A. I. J. M., J. L. Peña-Arancibia, and L. A. Bruijnzeel (2012b), Land cover and water yield: Inference problems when comparing catchments with mixed land cover, *Hydrol. Earth Syst. Sci.*, *16*(9), 3461–3473.

Van Dijk, A. I. J. M., H. E. Beck, R. S. Crosbie, R. A. M. de Jeu, Y. Y. Liu, G. M. Podger, B. Timbal, and N. R. Viney (2013), The millennium drought in southeast Australia (2001–2009): Natural and human causes and implications for water resources, ecosystems, economy and society, *Water Resour. Res.*, *49*, 2729–2746, doi:10.1002/wrcr.20123.

Van Oldenborgh, G., M. Balmaseda, L. Ferranti, T. Stockdale, and D. Anderson (2005), Did the ECMWF seasonal forecast model outperform statistical ENSO forecast models over the last 15 years?, *J. Clim.*, *18*(16), 3240–3249.

Viney, N. R., A. I. J. M. Van Dijk, and J. Vaze (2012), Comparison of models and methods for estimating spatial patterns of streamflow across a spatial domain, in *Proceedings Water Information Research and Development Alliance Science Symposium*, Melbourne, Australia, August. [Available at http://www.csiro.au/WIRADA-Science-Symposium-Proceedings.]

Wang, Q. J., D. E. Robertson, and F. H. S. Chiew (2009), A Bayesian joint probability modeling approach for seasonal forecasting of streamflows at multiple sites, *Water Resour. Res.*, *45*, W05407, doi:10.1029/2008WR007355.

Weisheimer, A., F. J. Doblas-Reyes, T. N. Palmer, A. Alessandri, A. Arribas, M. Deque, 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*, L21711, doi:10.1029/2009GL040896.

Werner, K., D. Brandon, M. Clark, and S. Gangopadhyay (2004), Climate index weighting schemes for NWS ESP-based seasonal volume forecasts, *J. Hydrometeorol.*, *5*(6), 1076–1090.

Westra, S., and A. Sharma (2010), An upper limit to seasonal rainfall predictability?, *J. Clim.*, *23*, 3332–3351.

Westra, S., C. Brown, U. Lall, I. Koch, and A. Sharma (2009), Interpreting variability in global SST data using independent component analysis and principal component analysis, *Int. J. Climatol.*, *30*, 333–346, doi:10.1002/joc.1888.

Wilks, D. S. (1995), Forecast verification, *Stat. Methods Atmos. Sci.*, *2*, 260–276.

Wilks, D. S. (2008), Improved statistical seasonal forecasts using extended training data, *Int. J. Climatol.*, *28*, 1589–1598, doi:10.1002/joc.1661.

Wood, A., and D. Lettenmaier (2006), A test bed for new seasonal hydrologic forecasting approaches in the western United States, *Bull. Am. Meteorol. Soc.*, *87*, 1699–1712.

Yebra, M., A. Van Dijk, R. Leuning, A. Huete, and J. P. Guerschman (2013), Evaluation of optical remote sensing to estimate actual evapotranspiration and canopy conductance, *Remote Sens. Environ.*, *129*, 250–261.

Yuan, X., E. F. Wood, L. Luo, and M. Pan (2011), A first look at Climate Forecast System version 2 (CFSv2) for hydrological seasonal prediction, *Geophys. Res. Lett.*, *38*(13), L13402, doi:10.1029/2011GL047792.

Zaitchik, B. F., M. Rodell, and F. Olivera (2010), Evaluation of the Global Land Data Assimilation System using global river discharge data and a source to sink routing scheme, *Water Resour. Res.*, *46*, W06507, doi:10.1029/2009WR007811.

Zhang, Y., J. M. Wallace, and D. S. Battisti (1997), ENSO-like interdecadal variability: 1900–93, *J. Clim.*, *10*, 1004–1020.

Zhang, Y. Q., N. R. Viney, F. H. S. Chiew, A. I. J. M. Van Dijk, and Y. Y. Liu (2011), Improving hydrological and vegetation modelling using regional model calibration schemes together with remote sensing data, in *Proceedings of the 19th International Congress on Modelling and Simulation MODSIM2011*, edited by F. Chan, D. Marinova, and R. S. Andersen, Modell. and Simul. Soc. of Aust. and N. Z., December. [Available at http://www.mssanz.org.au/modsim2011/I4/zhang.pdf.]