



## Evaluation of gridded meteorological datasets for hydrological modeling

Mélanie Raimonet, Ludovic Oudin, Vincent Thieu, Marie Silvestre, Robert Vautard, Christophe Rabouille, Patrick Le Moigne

### ► To cite this version:

Mélanie Raimonet, Ludovic Oudin, Vincent Thieu, Marie Silvestre, Robert Vautard, et al.. Evaluation of gridded meteorological datasets for hydrological modeling. *Journal of Hydrometeorology*, 2017, 18 (11), pp.3027-3041. 10.1175/JHM-D-17-0018.1 . hal-02881318

**HAL Id: hal-02881318**

**<https://hal.science/hal-02881318>**

Submitted on 26 Jun 2020

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

## Evaluation of Gridded Meteorological Datasets for Hydrological Modeling

MELANIE RAIMONET, LUDOVIC OUDIN, AND VINCENT THIEU

*Sorbonne Universités, UPMC, Univ. Paris 06, CNRS, EPHE, IPSL, UMR 7619 Metis,  
Paris, France*

MARIE SILVESTRE

*Sorbonne Universités, UPMC, Univ. Paris 06, CNRS, FR3020 FIRE, Paris, France*

ROBERT VAUTARD AND CHRISTOPHE RABOUILLE

*Laboratoire des Sciences du Climat et de l'Environnement, UMR CEA-CNRS-UVSQ 8212 et IPSL,  
Gif sur Yvette, France*

PATRICK LE MOIGNE

*CNRM UMR 3589, CNRS/Météo-France, Toulouse, France*

(Manuscript received 20 January 2017, in final form 5 September 2017)


### ABSTRACT

The number and refinement of gridded meteorological datasets are on the rise at the global and regional scales. Although these datasets are now commonly used for hydrological modeling, the representation of precipitation amount and timing is crucial to correctly model streamflow. The Génie Rural à 4 paramètres journalier (GR4J) conceptual hydrological model combined with the CEMANEIGE snow routine was calibrated using four temperature and precipitation datasets [Système d'analyse fournissant des renseignements atmosphériques à la neige (SAFRAN), Mesoscale Analysis (MESAN), E-OBS, and Water and Global Change (WATCH) Forcing Data ERA-Interim (WFDEI)] on 931 French gauged catchments ranging in size from 10 to 10 000 km<sup>2</sup>. The efficiency of the calibrated hydrological model in simulating streamflow was higher for the models calibrated on high-resolution meteorological datasets (SAFRAN, MESAN) compared to coarse-resolution datasets (E-OBS, WFDEI), as well as for reanalysis (SAFRAN, MESAN, WFDEI) compared to datasets based on interpolation only (E-OBS). The systematic decrease in efficiency associated with precipitation bias or temporality highlights that the use of a hydrological model calibrated on meteorological datasets can assess these datasets, most particularly precipitation. It appears essential that datasets account for high-resolution topography to accurately represent elevation gradients and assimilate dense ground-based observation networks. This is particularly emphasized for hydrological applications in mountainous areas and areas subject to finescale events. For hydrological applications on nonmountainous regions, not subject to finescale events, both regional and global datasets give satisfactory results. It is crucial to continue improving precipitation datasets, especially in mountainous areas, and to assess their sensitivity to eventual corrupted observations. These datasets are essential to correct the bias of climate model outputs and to investigate the impact of climate change on hydrological regimes.

### 1. Introduction

*a. The growing number of gridded meteorological datasets: Opportunities and caveats for hydrological studies*

Over the past two decades, gridded meteorological datasets have been increasingly used as inputs to

 Denotes content that is immediately available upon publication as open access.

*Corresponding author:* Melanie Raimonet, melanie.raimonet@upmc.fr

DOI: 10.1175/JHM-D-17-0018.1

© 2017 American Meteorological Society. For information regarding reuse of this content and general copyright information, consult the [AMS Copyright Policy](http://www.ametsoc.org/PUBSReuseLicenses) ([www.ametsoc.org/PUBSReuseLicenses](http://www.ametsoc.org/PUBSReuseLicenses)).

hydrological studies (Habets et al. 2008; Coustau et al. 2015; Soci et al. 2016). Such datasets result from the combination of short-term weather model forecasts with a quite diverse number of observation sources (weather stations, radar, buoys, and satellite products), using different data assimilation techniques. These datasets are continuously evolving, refining spatial coverage, incorporating new data sources, using improved atmospheric models, etc.

Easy access and the ability to mimic point-scale measurements and their spatial and temporal homogeneity (i.e., spatial grid and constant time step) has favored the widespread use of these products in hydrology, most particularly in hydrological modeling applications. This has considerably modified hydrological practices. Classical observation data processing (checking and completing missing precipitation and temperature values, estimates of spatial averages of these variables, etc.) is now less frequently performed by hydrologists for regional and global applications, given that these processes are included in the methodology to derive gridded meteorological datasets.

Gridded meteorological datasets are, however, integrated systems, and it is difficult to fully understand the impact and sensitivity to internal processing of observations, such as treatment of missing data and homogenization. Users, for example, hydrologists, often find it difficult to select one dataset or another, and the choice is often based on pragmatic reasons, for example, the spatial and temporal extent. Therefore, it is important to compare hydrological outputs obtained using different meteorological datasets.

#### *b. Requirements of meteorological datasets for hydrological modeling*

The spatial resolution of datasets must be compatible with the scale of each hydrological study. While most datasets are available at the global scale (Bosilovich et al. 2008), regional datasets at a higher spatial resolution are preferred for regional hydrological studies since they generally assimilate more observations than global ones; resolve physical processes at a finer scale; and consequently account for terrain characteristics, topography, land use, and local weather characteristics. In the particular case of hydrological modeling, Essou et al. (2016b) showed that using regional meteorological datasets instead of global ones for U.S. catchments provided better streamflow simulations, particularly for catchments located in humid continental and subtropical regions. This was explained by a better representation of precipitation seasonality for the regional meteorological datasets than for the global ones. Nevertheless, they also highlighted that, except for

humid continental and subtropical regions, global reanalysis precipitation data were used successfully in the United States, since biases were small enough to be compensated for by hydrological model calibration.

Historical depth is another desirable trait of meteorological datasets. Long-term time series are required for hydrological applications, especially for studies focusing on the impacts of climate variability or climate change on hydrology. For instance, in France, the *Système d'analyse fournissant des renseignements atmosphériques à la neige* (SAFRAN) reanalysis (Quintana-Seguí et al. 2008; Vidal et al. 2010) has increasingly been used since it was developed. This reanalysis covers metropolitan France at a relatively fine spatial scale ( $8\text{ km} \times 8\text{ km}$ ) and is available for a period beginning in 1958. If historical depth is important, the homogeneity of observation series assimilated in reanalyses is crucial and determines the useful length of datasets. As an example, ERA-Interim starts in 1979 because it assimilates satellite data that started in 1979 (Dee et al. 2011), and the ECMWF twentieth-century reanalysis (ERA-20C), which only assimilates observations of surface and mean sea level pressures and surface marine winds, starts in 1900 (Poli et al. 2016).

Evaluation of meteorological datasets is crucial (You et al. 2015) since such products bear limitations that may originate from several sources: low spatial or temporal resolution, a sparse observation station network, misrepresentation of the impact of topography, and atmospheric model biases. Most studies evaluate meteorological datasets by comparing them to ground-based observations (Jones et al. 2016; Quintana-Seguí et al. 2008; Dahlgren et al. 2014) or observation-only-based gridded datasets (You et al. 2015; Essou et al. 2016b; Dayon et al. 2015). These comparisons are intended to detect potential biases, and consequently bias corrections might be proposed and performed (Bastola and Misra 2014; Piani et al. 2010). However, few studies have investigated the ability of meteorological datasets to mimic daily variability.

This paper argues that outputs of conceptual hydrological models dynamically calibrated on a specific meteorological dataset might be additional relevant indicators to appraise the consistency of atmospheric inputs because an integrative variable (streamflow at the outlet of a river basin) is used. The use of conceptual hydrological models to assess the relevance of meteorological datasets is questionable. On one hand, since the development of hydrological models, several authors have pointed out the high sensitivity of hydrological models' streamflow simulations to precipitation data (see, e.g., Dawdy and Bergmann 1969). On the other hand, catchments and hydrological models may damp

out differences in precipitation data because of their storage functional behavior, and model parameter calibration might increase these compensations. However, if the calibration of a conceptual model tends to modify the model parameters to compensate bias in input data to a certain extent, it will compensate the misrepresentation of precipitation temporality very little (Oudin et al. 2006). Therefore, we hypothesize in this paper that the comparison of relative differences between the efficiency of each dataset/calibrated model output might be an indicator of the representation of temperature and precipitation amount and temporality: greater efficiency in streamflow simulations indicates a better representation of temperature and precipitation. Such integrated validations of datasets are particularly relevant when the hydrological model is used for hydrological projections under changing climate conditions (Bourqui et al. 2011; Essou et al. 2016b; Dayon et al. 2015) since climate projections are often downscaled on existing reanalysis grids (Vautard et al. 2013) and since the validation of downscaling methods on present-day data strongly depends on the dataset used (Dayon et al. 2015).

### c. Scope of the paper

The aim of this study was to assess the sensitivity of a hydrological model's streamflow simulation to several gridded meteorological datasets. We used a daily lumped four-parameter hydrological model [Génie Rural à 4 paramètres journalier (GR4J); Perrin et al. 2003] forced by precipitation and potential evaporation calculated from the air temperature. We compared the streamflow output performance of the conceptual hydrological model calibrated using four datasets of precipitation and temperature of various resolutions [SAFRAN, Mesoscale Analysis (MESAN), E-OBS, and Water and Global Change (WATCH) Forcing Data ERA-Interim (WFDEI)]. We examined the applicability of these meteorological datasets to hydrological modeling over metropolitan France on 931 catchments with diverse drainage areas, altitudes, and climatic settings, and we investigated the main reasons for the differences in model performance obtained with the four datasets. We finally discussed the required characteristics of meteorological datasets to force hydrological models, as well as the relevance of using a hydrological model to provide a complementary evaluation tool for these datasets.

## 2. Material

### a. Catchment set

The catchment set consisted of 931 gauged stations located throughout France (Fig. 1), which were selected

among more than 3500 stations available in the HYDRO French database (<http://www.hydro.eaufrance.fr/>). The catchment sizes ranged from 10 to 10 000 km<sup>2</sup>, with a great diversity of characteristics in this catchment set (e.g., topography, elevation, and climate; see Table 1).

The selection was made according to the following criteria: 1) no significant direct human influence on flow, 2) high measurement quality, 3) less than 20% missing values over both the 1989–2000 and 2000–10 periods, and 4) catchment size ranging from 10 to 10 000 km<sup>2</sup>. For the first two criteria, qualitative metadata provided by the monitoring authorities were available in the HYDRO database. The third criterion was chosen to ensure robust calibration of the hydrological model, and the focus on the 1989–2010 period was guided by the availability of the different gridded meteorological datasets tested. The fourth criterion is justified by the lumped daily hydrological model used in this study.

### b. Gridded meteorological datasets

Four different gridded datasets were used in this study, namely, SAFRAN, MESAN, E-OBS, and WFDEI. They cover a wide spectrum of available gridded meteorological products with different spatial resolutions and different methods based on interpolation techniques or more sophisticated assimilation systems used to derive climate variables (Table 2).

SAFRAN is an analysis system using an optimal interpolation (OI) method to compute each value analyzed by modifying a first-guess field with the weighted mean of the difference between observed and first-guess values at station locations within a search distance (Durand et al. 1993; Quintana-Seguí et al. 2008; Vidal et al. 2010). The first guess comes from the large-scale ARPEGE operational weather prediction model at a 0.25° resolution, and the observations come from Météo-France monitoring stations ( $n \approx 4100$  stations for precipitation and  $n \approx 1100$  stations for temperature over metropolitan France). The analysis is performed on climatological homogeneous areas, and the variables analyzed are projected to an 8-km regular grid accounting for elevation. The orography is set from a high-resolution 8 km  $\times$  8 km digital elevation model, and temperature and precipitation are determined with a vertical step of 300 m (Quintana-Seguí et al. 2008). SAFRAN is constructed on a dense observation network and comparisons with observations, including for precipitation and temperature. Comparative studies between SAFRAN reanalysis and observations show a good match (Quintana-Seguí et al. 2008; Vidal et al. 2010). SAFRAN is preferentially used by hydrologists in France (Habets et al. 2008; Martin et al. 2016) because it has given satisfactory results in many areas, providing

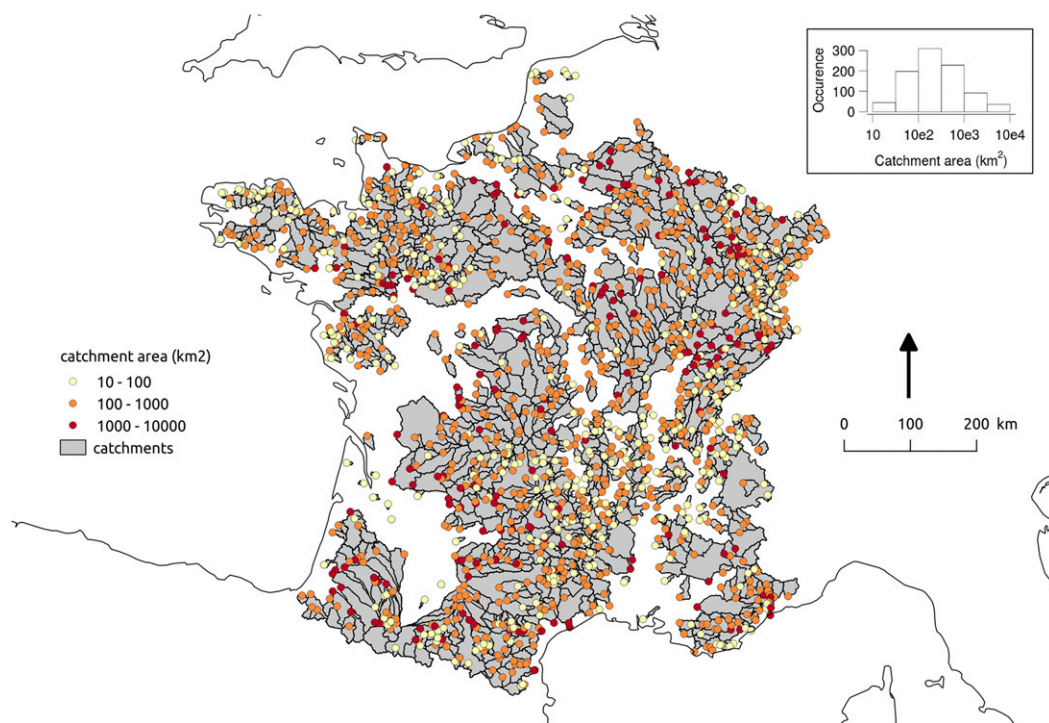


FIG. 1. Map of the 931 catchments studied (gray polygons). Points show the location of each catchment outlet, with a color range indicating the upstream area class. The top-right insert shows the distribution of catchment areas.

a set of consistent meteorological data, such as 2-m temperature and humidity, liquid and solid precipitation, incoming longwave and shortwave radiation, and wind speed. It is also the reference product used for national climate projections. For these reasons, we used SAFRAN as the reference in our study.

MESAN is a European high-resolution surface reanalysis also based on an OI technique (Häggmark et al. 2000; Landelius et al. 2016). The OI technique is based on 1) the HIRLAM operational forecasting system forced by ERA-Interim at the lateral boundaries ( $0.22^\circ$  field) downscaled to the MESAN  $0.05^\circ$  grid as the first guess and 2) observations from automatic stations, including for precipitation and temperature. Observations come from the same dense Météo-France monitoring network as for SAFRAN, with fewer stations for temperature. Häggmark et al. (2000) included a quality-check procedure to remove all observations with large errors using a standard procedure (Lorenc 1981). A high-resolution orography ( $5\text{ km} \times 5\text{ km}$ ) is used for this reanalysis, but orographic effects are not taken into account for daily precipitation (Landelius et al. 2016). We used the datasets interpolated to  $0.11^\circ$  grids (<http://exporter.nsc.liu.se/620eed0cb2c74c859f7d6db81742e114/>).

E-OBS is an observation-only-based gridded dataset of daily precipitation and temperature covering the

European continent. It is interpolated using a three-step methodology of interpolating the daily data for the 1950–present period (Haylock et al. 2008): 1) interpolating the monthly mean to define underlying spatial trends, 2) kriging anomalies with regard to the monthly mean, and 3) applying the interpolated anomaly to the interpolated monthly mean to obtain the final result. The number of ground-based stations is much lower than for SAFRAN and MESAN ( $n = 189$  stations for precipitation and  $n = 171$  stations for temperature over metropolitan France). In this study, we used version 14.0 aggregated on a  $0.22^\circ$  rotated grid, which is the finest resolution available for end-users (<http://www.ecad.eu/download/ensembles/download.php>).

WFDEI is a global meteorological forcing dataset at  $0.5^\circ$  resolution obtained by bias-correcting ERA-Interim (Weedon et al. 2014). ERA-Interim assimilates

TABLE 1. Range of catchment area, elevation, and daily observed water flow for the 931 catchments over the 1990–2010 period.

Characteristics	Median	Min	Max
Catchment area ( $\text{km}^2$ )	616	10.1	9119
Catchment elevation (m)	283	1	2154
Areal averaged observed water flow ( $\text{mm day}^{-1}$ )	0.57	0	381



TABLE 2. Description of daily gridded meteorological datasets used in this study. Asterisks indicate that the number of stations is for the France domain only.

Acronym	Temporal extent	Spatial extent	Spatial resolution	Observation stations	Analysis method for surface $P$ and $T$	References
SAFRAN	1958–present	France	8 km	$P \sim 4100$ $T \sim 1100$	OI	Durand et al. (1993); Quintana-Seguí et al. (2008); Vidal et al. (2010)
MESAN	1989–2010	Europe	$0.11^\circ$ $\sim 12$ km	$P = 4053^*$ $T = 252^*$	OI	Häggmark et al. (2000); Landelius et al. (2016)
E-OBS	1950–present	Europe	$0.22^\circ$ $\sim 25$ km	$P = 189^*$ $T = 171^*$	Kriging	Haylock et al. (2008)
WFDEI	1979–2012	Europe	$0.5^\circ$ $\sim 50$ km		Variational 4D	Weedon et al. (2014)

surface temperature and humidity observations, as well as many other atmospheric variables, but not precipitation, which is a diagnostic variable. WFDEI then corrects ERA-Interim for precipitation biases using data from the Climatic Research Unit (CRU) or from the Global Precipitation Climatology Center (GPCC). In this study, GPCC products were preferred to CRU because of their higher resolution linked to the higher station density (Weedon et al. 2014). Orographic effects on precipitation are not corrected in WFDEI products, possibly leading to some inappropriate precipitation phases. However, the most extreme cases of inappropriate precipitation phase were corrected (Weedon et al. 2014). Given the relatively low spatial resolution, this meteorological forcing is expected to be valuable for hydrological modeling on large catchments ranging from 100 to 10 000 km<sup>2</sup> (Weedon et al. 2010, 2011). This dataset has recently been shown to lead to better streamflow estimations than no-bias-corrected reanalysis using a global conceptual hydrological model in humid continental and subtropical climatic regions (Essou et al. 2016b). Conversely, these authors also showed that using bias-corrected reanalysis deteriorates streamflow simulations in several catchments located in the United States.

Each meteorological dataset was extracted at the catchment scale. Gridded temperature and precipitation values were weighted according to the shared surface of each grid cell within the topographic limits of each catchment to obtain catchment temperature and precipitation. These extractions were made at the daily time step and were used as inputs for the hydrological model.

### c. Hydrological model

In this study we used the GR4J hydrological model associated with the CEMANEIGE snow accounting routine (Fig. 2).

The GR4J hydrological model is a daily lumped four-parameter model (Perrin et al. 2003). GR4J inputs are

daily mean values of areal precipitation and potential evaporation. Potential evaporation was derived from the mean daily air temperature and latitude of the catchment using an empirical temperature-based formula

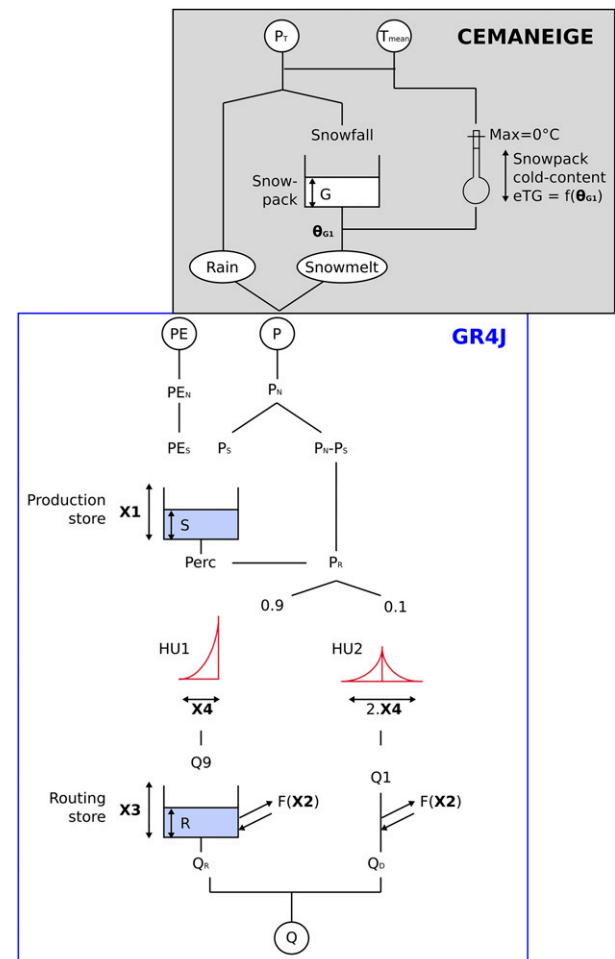


FIG. 2. Overview of the GR4J hydrological model (Perrin et al. 2003) combined with the CEMANEIGE snow module (Valéry et al. 2014).

(Oudin et al. 2005). A snow-accounting routine called CEMANEIGE (Valéry et al. 2014) was combined with GR4J in order to determine the solid fraction of precipitation and the temporal changes in the snowpack on mountainous catchments, based on daily temperature and catchment elevation. Note that this routine has two parameters that can potentially be calibrated along with the GR4J parameters, but in this study we used their default values proposed by Valéry et al. (2014) for France. Using default values was judged necessary to avoid unrealistic parameterization of the snow module that may occur to compensate for errors in gridded meteorological datasets, especially in catchments not affected by snow. The results obtained without the CEMANEIGE routine (not shown) gave the same conclusions as those obtained in this paper with the CEMANEIGE routine, with greater efficiency when using the CEMANEIGE module on snow-affected catchments.

A detailed description of the model can be found in Perrin et al. (2003), and only the main features of the model are described hereafter. Net rainfall (the rainfall amount that reaches the outlet of the catchment) and actual evaporation are calculated as functions of the soil moisture storage level  $S$ , the difference between rainfall and potential evaporation ( $P - PE$ ) and the parameter  $X1$  representing the maximum capacity of the soil moisture storage. Interbasin groundwater flows are parameterized by a second parameter  $X2$ . Positive or negative  $X2$  leads to a catchment water supply or loss, respectively.  $X1$  and  $X2$  are the two possible calibrated parameters that can adjust the catchment's water budget.

Net rainfall is divided into two flow components in the routing function: 1) 90% are routed by a unit hydrograph, with a time-base parameter  $X4$  and a nonlinear routing storage, and 2) the remaining 10% are routed through a unit hydrograph. The  $X3$  parameter represents the maximum level of the routing storage. The two flow components calculated in the routing function are summed to obtain the simulated streamflow at the catchment outlet.

#### *d. Metrics to assess the differences/consistency of meteorological datasets*

All statistical analyses were performed using the R software (<https://cran.r-project.org/>). Metrics were calculated to evaluate 1) hydrological model performance by comparing simulated streamflow using the four meteorological datasets to observed streamflow and 2) meteorological dataset performance using precipitation or temperature for the three datasets tested (MESAN, E-OBS, and WFDEI) compared with the

reference precipitation or temperature (SAFRAN) for the 931 catchments studied.

We chose to calibrate specific sets of hydrological model parameters for each meteorological dataset following a so-called dynamic sensitivity analysis of the hydrological model to forcing data. Oudin et al. (2006) highlighted that dynamic versus static calibration must be chosen depending on the hydrological model used. Dynamic calibration of the conceptual hydrological model is needed on each atmospheric forcing dataset, since a “true” parameter set cannot exist independently of the calibration dataset, while static calibration is preferred for a physical model since a true parameter set could be derived without calibration. Note, however, that even for physical models, Sperna Weiland et al. (2015) highlighted the difficulty of finding optimal parameter sets that can be applied for all atmospheric forcing datasets.

The four parameters of the GR4J model were calibrated using the Kling–Gupta efficiency (KGE) coefficient (Gupta et al. 2009) as the objective function. Because of the parsimony of the model, the calibration does not face the problem of multiple optima regardless of the calibration method (Edijatno et al. 1999; Perrin et al. 2003). In our study, we used the steepest descent method used by Edijatno et al. (1999).

The KGE is a decomposition enabling the estimation of the relative importance of its components (variability  $\alpha$ , bias  $\beta$ , and correlation  $r$ ) in the context of hydrological modeling (Gupta et al. 2009). We used the R package “hydroGOF”, which contains goodness-of-fit (GOF) functions for numerical and graphical comparison of simulated and observed time series, mainly focused on hydrological modeling (Zambrano-Bigiarini 2014):

$$\text{KGE} = 1 - \left[ \sqrt{(r-1)^2 + (\alpha-1)^2 + (\beta-1)^2} \right].$$

The variance ratio  $\alpha$  is the ratio between simulated and reference variances:

$$\alpha = \frac{\sigma_s}{\sigma_{\text{ref}}}.$$

The mean ratio  $\beta$  is the ratio between simulated and reference means:

$$\beta = \frac{\mu_s}{\mu_{\text{ref}}}.$$

The Pearson correlation coefficient  $r$  is calculated between the simulated and reference datasets:

$$r = \frac{\text{COV}_{\text{sref}}}{\sigma_s \times \sigma_{\text{ref}}}.$$

The best performance is obtained for KGE,  $\alpha$ ,  $\beta$ , and  $r$  values close to 1.

While hydrological model outputs are generally validated on (nonsorted) time series, climate model outputs are generally evaluated on (sorted) dataset quantiles. For this reason, we also computed root-mean-square errors (RMSEs) between simulated and observed streamflow quantiles for each catchment.

Significant differences between meteorological datasets for each statistical criterion were assessed using the Kruskal–Wallis nonparametric test for multiple comparisons of groups from the R package “agricolae”. The Kruskal–Wallis test distinguishes groups that have different distributions of their ranked values. The difference between groups is represented by different letters in each subplot presented hereafter. Nonsignificant differences are represented by the same letter. The  $p$  value threshold is 0.05.

To assess the temporal robustness of the model parameters calibrated using each meteorological dataset, we followed a split-sample test procedure (Klemeš 1986), also called “cross validation.” For each catchment and each dataset, the model parameter values are calibrated on the 1989–2000 period, and the streamflow simulations are compared to observations on the 2000–10 period. Then the 2000–10 period is used as the calibration period and the 1989–2000 period as the validation period. In doing such dynamic sensitivity analysis, we allow the calibrated model parameters to eventually adjust their value to the rainfall and temperature inputs.

### 3. Results

This section is structured as follows: the efficiency of the streamflow outputs of the hydrological model calibrated on four gridded meteorological datasets is assessed, and then the reasons for the different levels of efficiency questioning the impact of geographic locations, catchment areas, elevation, and their climatic specificities are investigated.

#### a. Evaluation of streamflow simulation using the different gridded meteorological datasets as inputs

The distributions of the KGE streamflow  $Q$  values and its three components were investigated over the 931 catchments studied and for the four meteorological datasets tested (Fig. 3). The hydrological model forced with SAFRAN and MESAN high-resolution datasets showed the highest statistical efficiencies compared to E-OBS and WFDEI low-resolution datasets regardless of the criteria (Fig. 3). The KGE  $Q$  median

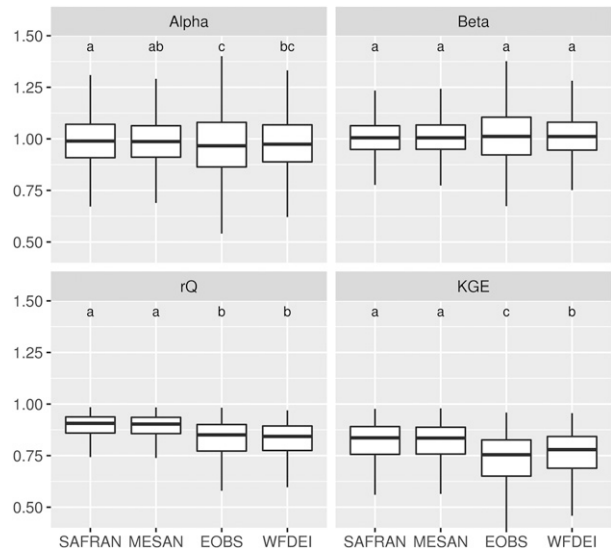


FIG. 3. Statistical criteria (the KGE and its components  $\alpha$ ,  $\beta$ , and  $r$ ) on simulated vs observed streamflow  $Q$  for each gridded meteorological dataset (E-OBS, MESAN, SAFRAN, WFDEI) at a daily resolution for the 931 studied catchments. For each criterion, the letters above the box plot indicate dataset groups of significantly different criteria distributions (Kruskal–Wallis test).

values in validation were higher than 0.75 no matter the meteorological dataset used, with significantly better values for SAFRAN and MESAN (0.84) compared to WFDEI (0.78) and E-OBS (0.75). Then we compared the median values of the three components of this criterion, that is, the variance ratio  $\alpha$ , the mean ratio  $\beta$ , and the Pearson correlation coefficient on streamflow  $rQ$ . The median of the variance ratio  $\alpha$  showed that the variance was well represented using SAFRAN and MESAN (0.99) and was slightly lower with E-OBS and WFDEI (0.97). The median values of the mean ratio  $\beta$  were close to 1 for all datasets, which indicated no general bias of simulated streamflow over the validation periods. Even if not significant, a larger bias was observed for E-OBS compared with other datasets, and much lower first quartile values were observed. The median values of the Pearson correlation coefficient  $rQ$  were significantly higher for SAFRAN (0.91) and MESAN (0.90) than for E-OBS (0.85) and WFDEI (0.84). For all criteria, lower first quartile values were generally observed for E-OBS, and to a lesser extent, for WFDEI.

KGE  $Q$  distributions were not significantly different for SAFRAN and MESAN, but different for E-OBS and WFDEI (Fig. 4, left). Comparing the KGE  $Q$  values obtained with each meteorological dataset (E-OBS, MESAN, WFDEI) to the SAFRAN reference dataset (Fig. 4, right), the absence of difference between



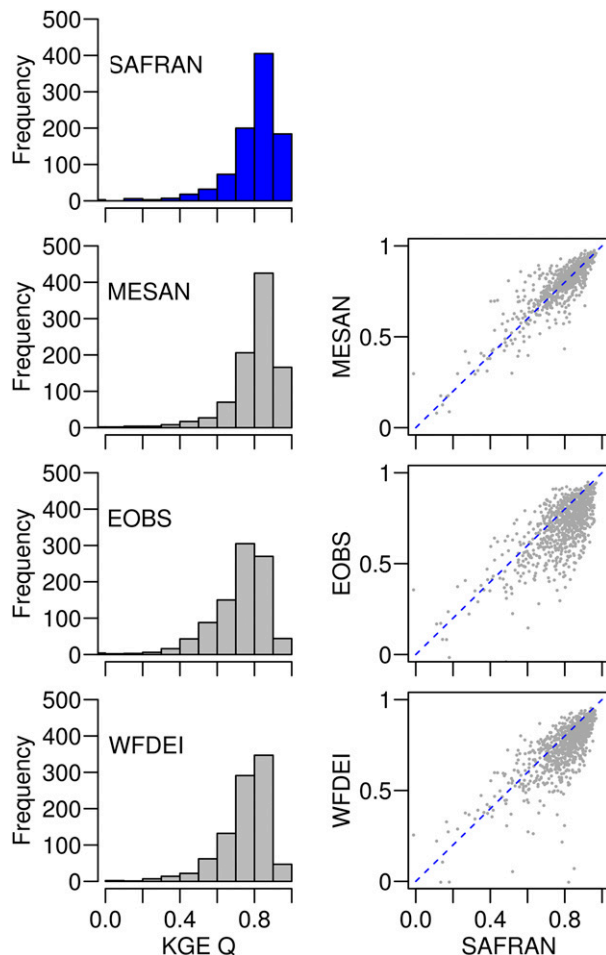


FIG. 4. (left) Distribution of calibrated KGE on simulated vs observed  $Q$  for each gridded meteorological dataset. The SAFRAN reference is in blue. (right) Comparison of each dataset KGE  $Q$  with the SAFRAN KGE  $Q$  reference. The 1:1 line is dashed blue;  $n = 931$  catchments.

SAFRAN and MESAN was confirmed by the points following the 1:1 line, while many catchments showed lower performance for WFDEI and much lower for E-OBS.

We then calculated RMSE values on simulated versus observed  $Q$  quantiles (Fig. 5). The RMSE median values calculated on  $Q$  quantiles were not significantly different for SAFRAN and MESAN ( $0.087 \text{ mm day}^{-1}$ ), E-OBS ( $0.095 \text{ mm day}^{-1}$ ), and WFDEI ( $0.092 \text{ mm day}^{-1}$ ).

#### *b. Potential factors controlling hydrological model efficiency forced by meteorological datasets*

To investigate possible areas with poor streamflow simulation efficiency for some meteorological datasets, we analyzed the spatial distribution of KGE  $Q$  for SAFRAN (the reference) and the KGE  $Q$  anomalies

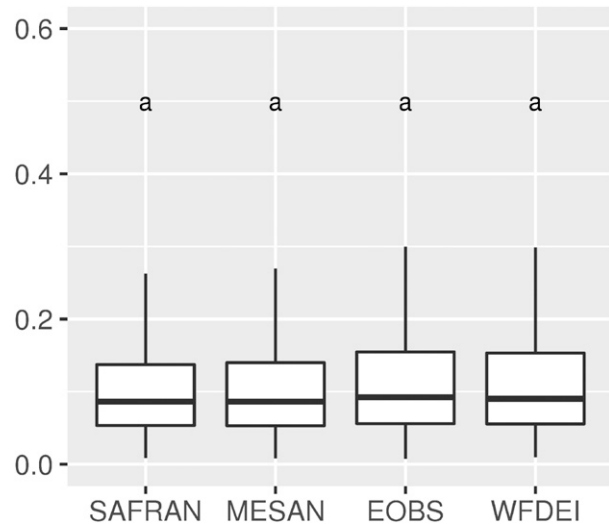


FIG. 5. RMSEs ( $\text{mm day}^{-1}$ ) on streamflow quantiles;  $n = 931$  catchments.

of MESAN, E-OBS, and WFDEI datasets compared to the reference (Fig. 6). Most KGE  $Q$  values were  $>0.6$  (green) when using SAFRAN precipitation and temperature as atmospheric forcing (Fig. 6a), which confirms the quality of the GR4J/CEMANEIGE hydrological model calibrated on SAFRAN. Most of the lowest hydrological model efficiency values were observed in mountainous areas. As expected from the results presented in Fig. 4, MESAN showed the lowest (negative and positive) anomalies to SAFRAN, while WFDEI and even more E-OBS showed the highest negative anomalies, especially in mountainous regions (Figs. 6b–d).

We investigated the influence of the mean catchment altitude on the KGE  $Q$  values for the SAFRAN reference and on the KGE  $Q$  anomalies of each meteorological dataset compared to SAFRAN (Fig. 7). KGE  $Q$  values for the SAFRAN reference significantly decreased for high-elevation catchments (Fig. 7a). Increasing mean altitude significantly decreased KGE  $Q$  for MESAN (Fig. 7b), but even more for E-OBS and WFDEI (Figs. 7c,d).

We then investigated the influence of catchment size on the KGE  $Q$  values for the SAFRAN reference and on the KGE  $Q$  anomalies of each meteorological dataset compared to SAFRAN (Fig. 8). KGE  $Q$  values for the SAFRAN reference significantly decreased when catchment size decreased (Fig. 8a). While the altitude had a stronger influence for MESAN, E-OBS, and WFDEI compared to the SAFRAN reference, the influence of catchment size was similar for MESAN, E-OBS, and WFDEI compared to the reference (Figs. 8b–d).

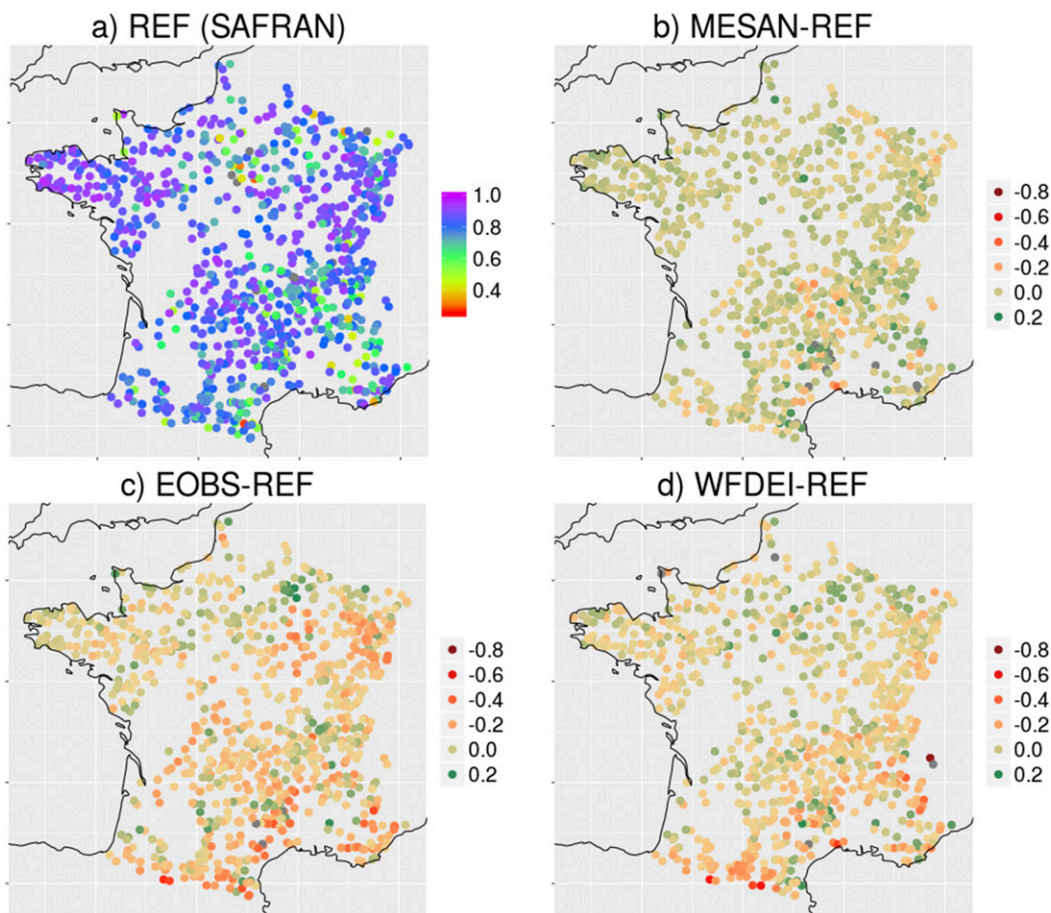


FIG. 6. (a) KGE values obtained on simulated vs observed  $Q$  for the SAFRAN reference and (b)–(d) KGE  $Q$  anomalies of each gridded meteorological dataset (MESAN, E-OBS, WFDEI) compared to the SAFRAN reference. Negative anomalies compared to SAFRAN account for 51%, 80%, and 79% of the 931 studied catchments for MESAN, E-OBS, and WFDEI, respectively.

We finally investigated if the loss or gain in hydrological efficiency ( $KGE\ Q$ ) was related to the properties of the atmospheric forcing inputs (precipitation  $P$  and temperature  $T$ ) tested: MESAN, E-OBS, and WFDEI compared to the SAFRAN reference (Fig. 9). The characteristics of atmospheric forcing were investigated by comparing each component of the  $KGE\ P$  and  $KGE\ T$  calculated between each dataset and the SAFRAN reference over the entire simulation period (1989–2010) for precipitation and air temperature, respectively.

For precipitation, variance and mean ratios ( $\alpha$  and  $\beta$ ) differed significantly between the three datasets (Fig. 9, upper panel). They were close to 1 for MESAN (median = 1.01 and 1.00, respectively), indicating not significantly different variance and mean compared to SAFRAN. Variance and mean were much lower for E-OBS than for SAFRAN, with lower  $\alpha$  and  $\beta$  values (0.88 and 0.87). For WFDEI,  $\alpha$  was lower (0.97) and  $\beta$  slightly higher (1.03) than for SAFRAN.

The correlation coefficient  $rP$  showed a different pattern, with a significantly decreasing correlation coefficient from higher values for MESAN (0.98) to lower values for E-OBS (0.92) and even lower for WFDEI (0.71). Therefore, the low performance of the hydrological model is likely due primarily to bias in precipitation for E-OBS and to poor correlation in precipitation for WFDEI.

For temperature, fewer interdataset differences were observed for each component of the  $KGE\ T$  (Fig. 9, lower panel). The variance ratio  $\alpha$  was high and not significantly different for MESAN and E-OBS (median = 0.996), and for E-OBS and WFDEI (0.994), indicating similar variance compared to SAFRAN. The mean ratio  $\beta$  was equal to 1 for MESAN and significantly higher for E-OBS (1.04) and WFDEI (1.07). The  $rT$  Pearson correlation coefficient was quite high for MESAN (median = 0.996) and slightly but significantly lower for E-OBS (0.991) and WFDEI (0.987).

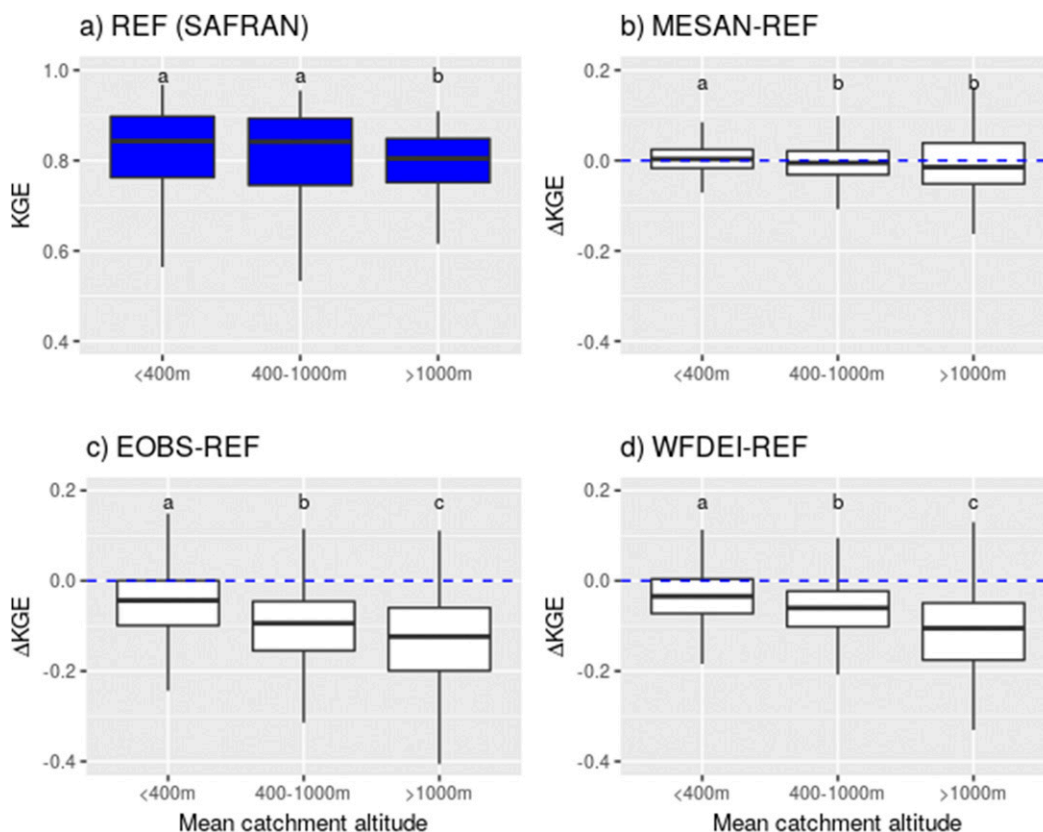


FIG. 7. (a) KGE values on simulated vs observed  $Q$  for the SAFRAN reference and (b)–(d) KGE  $Q$  anomalies for the other gridded meteorological datasets (MESAN, E-OBS, WFDEI) compared to the SAFRAN reference. Values are given for three altimetric classes (<400 m, 400–1000 m, >1000 m). The dashed blue line represents zero differences of KGE  $Q$  values for each gridded meteorological dataset compared to the KGE  $Q$  values computed with SAFRAN, and  $n = 931$  catchments.

#### 4. Discussion

##### a. Main requirements of gridded meteorological datasets for hydrological modeling

The hydrological model used in this study appears sensitive to the choice of the gridded meteorological datasets since the datasets lead to different levels of efficiency in terms of streamflow simulations. Disentangling the factors explaining these differences is not easy as more than one characteristic is different between each dataset (e.g., the resolution, the assimilation method, the number of monitoring stations, the type of observations, the temporal coverage), but it is possible to identify which factors have a major impact (i.e., high-resolution accounting for topography and high observation density).

Accounting for topography is essential to improve the representation of precipitation and temperature gradients with altitude and is one of the most constraining factors in actual meteorological datasets. This implies high spatial resolution datasets, either based on dense

observation networks or on models representing small-scale processes and accounting for elevation gradients. This is highlighted by comparing the hydrological model efficiency obtained with the 8-km SAFRAN and the 50-km WFDEI. The high resolution of SAFRAN accounts for catchment topography down to  $64 \text{ km}^2$  and calculates temperature and precipitation with a 300-m vertical step, leading to high model efficiency. On the contrary, the low-resolution WFDEI does not integrate correction of precipitation for orographic effects and shows lower model efficiency. Accounting for topography is particularly essential in mountainous catchments to have adequate estimations of snow. SAFRAN gives the best hydrological model efficiencies in mountainous catchments, most probably because it takes into account a high-resolution topography, as well as the elevation gradient to compute temperature and precipitation variables. Limitations can, however, arise for small catchments. Indeed, regardless of the meteorological datasets used, the efficiency to model streamflow is lower for catchments with surface area lower than  $100 \text{ km}^2$  (Fig. 7),

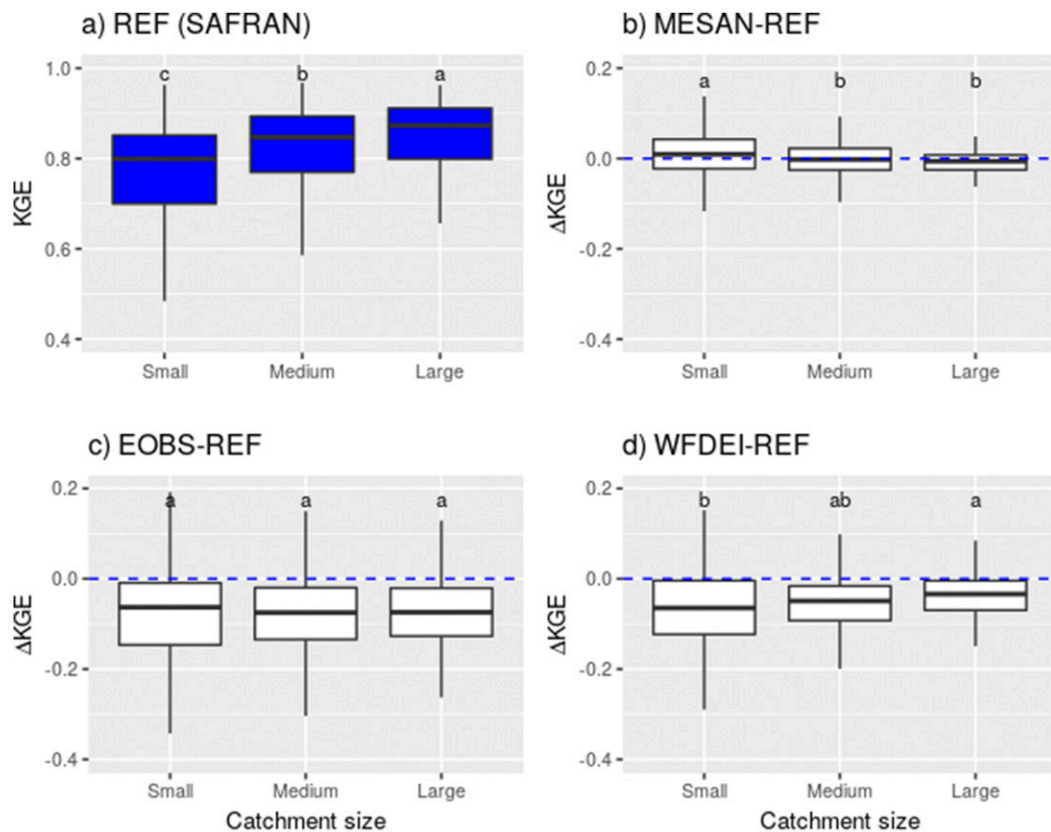


FIG. 8. As in Fig. 7, but values are given for three catchment area classes (small,  $<100 \text{ km}^2$ ; medium,  $100\text{--}1000 \text{ km}^2$ ; large,  $>1000 \text{ km}^2$ ).

as the resolution of the meteorological dataset becomes coarse compared to the catchment size. Conversely, the systematic increase of efficiency of the calibrated hydrological model with increasing catchment size may be due to the increased damping effect of large catchments (explaining their lower sensitivity to forcing data).

If the resolution is high but the observation network is sparse and/or the physical model used to create reanalyses does not represent small-scale processes, high spatial heterogeneity will not be accounted for. Using SAFRAN 8 km or MESAN 12 km, both built on a large number of observations, gives similarly high efficiencies of the hydrological model outputs. However, when comparing SAFRAN (8 km) or MESAN (12 km) to E-OBS (25 km), we observe a significant decrease in the efficiency of the hydrological model to estimate streamflow with E-OBS. This might be related to the limited number of available monitoring stations in France for E-OBS ( $n = 171$  for temperature;  $n = 189$  for precipitation) compared to the large number of observations used for SAFRAN and MESAN ( $n \sim 1100$  and  $250$  for temperature;  $n \sim 4100$  for precipitation). Based on the available monitoring stations, E-OBS cannot

achieve a spatial representation as good as SAFRAN and MESAN. The hydrological model efficiency using E-OBS was especially low in mountainous regions, thus reinforcing the findings of [Isotta et al. \(2015\)](#) that showed higher daily precipitation RMSE for E-OBS than for MESAN in the French Alps, mostly due to lower station density. In addition, the lowest model performance using SAFRAN is often observed in mountainous regions where weather stations are often sparse, and precipitation events are spatially heterogeneous and not often captured by the gauge network ([Durand et al. 1993](#); [Prein and Gobiet 2017](#)). This raises the question of the representativeness of sparse observations in mountainous regions. Even if accounting for quality-controlled observations is better than not doing it, analyses of precipitations based on sparse observations in mountainous regions could lead to larger errors in precipitation estimations than with dense observation networks. In addition, stations are not all located at the same altitude. Therefore, reconstruction of the precipitation field at a given elevation needs to account for stations at different altitudes, and a correction of the precipitation amount with altitude should usually be used



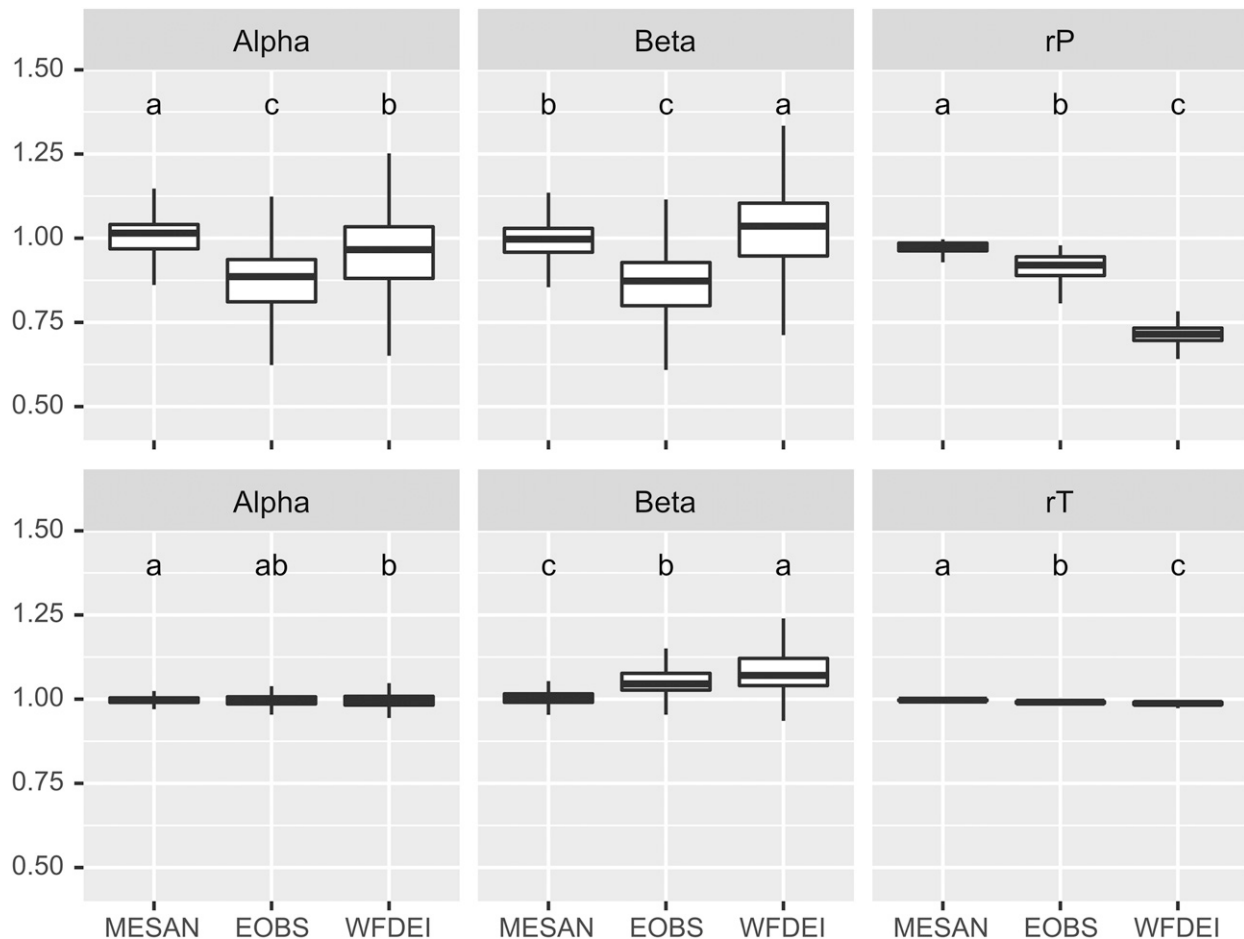


FIG. 9. Statistical criteria ( $\alpha$ ,  $\beta$ , and  $r$ ) computed between each gridded meteorological dataset (E-OBS, MESAN, WFDEI) and for the SAFRAN reference for the 931 catchments for daily (top) precipitation and (bottom) temperature.

in the analysis of precipitation. Thus, underlying dense networks of ground-based observation stations in meteorological datasets with correction for station altitude, particularly in mountainous areas, appear essential to correctly account for spatial heterogeneities and give a fine description of the meteorological situation (Garcia et al. 2008). This is particularly crucial in order to correctly capture the phases of precipitation, that is, to distinguish between rain and snow, which will impact streamflow simulations.

The use of assimilation schemes and/or meteorological model outputs to create meteorological datasets can possibly have an impact on the efficiency of hydrological outputs. As an example, the better efficiency calculated using WFDEI compared to E-OBS shows that, even if the resolution of WFDEI is coarse, accounting for processes in the forecast model and using bias correction give better results than gridded observations only. This particularly leads to better precipitation mean and

variance. This is in agreement with recent findings in eastern U.S. catchments showing that bias-corrected dynamic downscaling of global reanalysis gave better streamflow estimations than biased ones, global reanalysis, or synthetically generated meteorological forcing from a weather generator (Bastola and Misra 2014). However, we also show that, even if precipitation mean and variance are better for WFDEI than for E-OBS, the correlation of daily amounts to reference SAFRAN precipitations is much lower, indicating that the temporality of precipitation events is less accurately represented in WFDEI. The ineffectiveness of WFDEI to represent daily temporality of precipitation might be explained by the lack of assimilation of ground-based precipitation observations with the physical weather model (Weedon et al. 2014). The weather model is implemented at a coarse resolution, which does not account for fine spatial and temporal scale processes. Indeed, this product



results from a forecast model at the global scale, which is appropriate for the representation of global-scale processes but not for local processes.

The use of gridded meteorological datasets for hydrological purposes is often based on pragmatic reasons, such as the time period covered by the dataset or the spatial and temporal resolution, and intercomparisons to select the most adequate product are often left out because of time constraints. However, our results show the necessity of evaluating meteorological datasets before using them, especially for models like GR4J that are sensitive to meteorological forcing.

*b. On the use of a hydrological model to assess the relevance of meteorological datasets*

Using a conceptual hydrological model dynamically calibrated on meteorological datasets highlights how crucial the quality of precipitation and temperature datasets is to correctly simulate streamflow. Precipitation is particularly different among the tested datasets, most probably because of the difficulty of capturing or modeling the spatial distribution of precipitation compared to more continuous variables like temperature. Given the high sensitivity of hydrological model outputs to precipitation, the efficiency of a hydrological model depends strongly on the accuracy of the precipitation dataset used. This confirms the recent findings of [Essou et al. \(2016b\)](#), who showed the sensitivity of a global conceptual model to precipitation inputs on 370 U.S. catchments of a similar size range (104–10 325 km<sup>2</sup>). However, the accuracy of precipitation datasets may not be optimal for all regions. For SAFRAN, which is the meteorological reanalysis reference in France, the lowest hydrological model efficiencies are observed in mountainous catchments. This is linked to the inability of SAFRAN to represent intrazone variability and strong horizontal gradients in mountainous regions and by high spatial variability of convective precipitation along the Rhone and on the Mediterranean border of the Massif Central ([Quintana-Seguí et al. 2008](#)).

In this paper, we use a dynamic approach consisting of calibrating the model parameters for each of the meteorological datasets. This is warranted both by the conceptual nature of the hydrological model and by the fact that no dataset can be considered a priori as a reference for the calibration. Consequently, the calibration of the model parameters might compensate for certain errors in the meteorological forcing. However, we show that even if conceptual hydrological models are dynamically tuned on each atmospheric forcing, the calibration procedure does not offset biased or out-of-phase precipitation datasets. Indeed, even if WFDEI shows a precipitation mean and variance close

to SAFRAN, the misrepresentation of temporality (i.e., low correlation on daily precipitation values) is the main factor leading to worse streamflow simulations. For E-OBS, the underestimation of the precipitation mean is not compensated for by model calibration, and the model efficiency strongly decreases. Our results based on 931 catchments in France strengthen the findings of [Oudin et al. \(2006\)](#), who showed that the dynamically calibrated GR4J model efficiency had a very high sensitivity to corrupted precipitation inputs compared with temperature for 12 U.S. catchments.

Interestingly, [Essou et al. \(2016a\)](#) showed that their conceptual hydrological model was able to compensate for errors of four gridded datasets of various resolution levels in the United States. The results might be partly explained by the larger number of calibrated parameters in their hydrological model (23 parameters) compared with our study (four parameters). This suggests that conceptual hydrological models with a limited number of calibrated parameters would possibly be more likely to evaluate and intercompare meteorological datasets because of their lower degree of freedom. However, comparisons between the two calibrated models forced by the same datasets are needed to confirm this statement. Further comparisons with a mechanistic approach, that is, a distributed hydrological model, would be interesting to evaluate the impact of the spatialization of the meteorological inputs on the spatial patterns of streamflow.

Quantifying hydrological modeling efficiency driven by meteorological forcing appears to be an interesting tool to evaluate meteorological datasets for two main reasons. First, the sensitivity of hydrological model outputs to meteorological inputs allows identifying the most appropriate meteorological datasets. Second, as hydrological processes are related to the daily covariation of precipitation and temperature and are characterized by an important memory effect, using a hydrological model forced by meteorological datasets is an integrative tool to evaluate precipitation and temperature datasets and their covariation, as well as the memory effect. For these reasons, we argue that, in addition to the necessary evaluation of meteorological inputs for hydrological purposes, this evaluation may be also useful for climate impact studies. Indeed, climate model outputs, particularly precipitation, are not assessed on time series but rather on sorted, averaged, or event values, for example, event frequency, intensity, total, and duration ([Loikith et al. 2017](#)). This type of validation for precipitation is mostly explained by the lower confidence in model estimates for precipitation than for temperature ([Randall et al. 2007](#)). In our study, we show that the RMSE on (sorted) streamflow quantiles does not allow discrimination between gridded

meteorological datasets, contrary to statistical criteria calculated on (unsorted) time series. We thus stress that climate impact studies need evaluations on unsorted variables, as the timing of events is important to quantify impacts, for example, hydrological and biogeochemical impacts.

## 5. Conclusions

In this paper, we show that for hydrological applications, it appears essential that gridded meteorological datasets account for high-resolution topography to accurately represent elevation gradients and assimilate dense ground-based observations. This is emphasized for mountainous areas and/or areas subject to finescale events while acceptable streamflow simulations were obtained for low-altitude catchments, regardless of the meteorological input used.

Even if the calibration of hydrological model parameters is dynamic, using a dataset that misrepresents precipitation amount or timing results in a decrease in streamflow output efficiency. Therefore, our methodology based on dynamic calibration of a conceptual model on meteorological datasets appears to be an interesting tool to assess the consistency of meteorological datasets for hydrological applications, complementary to classical validation procedures of these datasets. More precisely, using hydrological modeling allows the evaluation of the covariability of precipitation and temperature, as well as their temporality, because of the relative inertia of the catchments.

In this paper, the similar efficiencies of the model calibrated using SAFRAN and MESAN highlight the possible use of both reanalyses to model hydrology at the scale of France. Even if the hydrological model efficiency is not tested outside of France in our study, reanalyses like MESAN are expected for hydrological studies covering Europe, as MESAN is the most complete homogeneous temperature and precipitation dataset available for Europe (Landelius et al. 2016). Landelius et al. (2016), however, stressed the need for open national databases, the lack of which leads to limitations for high-resolution datasets in large areas of Europe.

Both SAFRAN and MESAN seem to be useful as a reference to validate atmospheric outputs of climate models over their respective coverage period, at least for France. This strengthens the use of SAFRAN and MESAN reanalyses to correct biases of atmospheric outputs of regional climate models at the scale of France, for example, DRIAS (<http://www.drias-climat.fr>) or at the European scale, for example, CORDEX ([www.cordex.org](http://www.cordex.org)) for climate change impact studies.

Quantifying the sensitivity of this approach to known gridded meteorological dataset errors would allow reaching more quantitative conclusions. ERA5, at a

30-km resolution, has just been launched and will provide a set of approximately 10 runs with perturbations of initial conditions and observations in order to provide uncertainty estimation. It would be advantageous to test the response of simulated streamflow to these perturbations in order to evaluate their sensitivity to known errors in atmospheric forcing. It would be even more useful to also perform the same test on higher-resolution meteorological datasets.

**Acknowledgments.** This work was supported by Labex L-IPSL, which is funded by ANR (Grant ANR-10-LABX-0018), and EC2CO MARICCA, which is funded by INSU/CNRS. The authors are grateful to Josette Garnier for her advice on these projects, Andrew Wood for the editorial work, and two anonymous reviewers for their insightful comments that improved the manuscript.

## REFERENCES

- Bastola, S., and V. Misra, 2014: Evaluation of dynamically downscaled reanalysis precipitation data for hydrological application. *Hydrol. Processes*, **28**, 1989–2002, doi:[10.1002/hyp.9734](https://doi.org/10.1002/hyp.9734).
- Bosilovich, M. G., J. Chen, F. R. Robertson, and R. F. Adler, 2008: Evaluation of global precipitation in reanalyses. *J. Appl. Meteor. Climatol.*, **47**, 2279–2299, doi:[10.1175/2008JAMC1921.1](https://doi.org/10.1175/2008JAMC1921.1).
- Bourqui, M., T. Mathevet, J. Gailhard, and F. Hendrickx, 2011: Hydrological validation of statistical downscaling methods applied to climate model projections. *IAHS Publ.*, **344**, 32–38, <http://iahs.info/uploads/dms/16758.09-32-38-344-33-JH02-abstract144-Bourqui-et-A-COR.pdf>.
- Coustau, M., F. Rousset-Regimbeau, G. Thirel, F. Habets, B. Janet, E. Martin, C. de Saint-Aubin, and J.-M. Soubeyroux, 2015: Impact of improved meteorological forcing, profile of soil hydraulic conductivity and data assimilation on an operational hydrological ensemble forecast system over France. *J. Hydrol.*, **525**, 781–792, doi:[10.1016/j.jhydrol.2015.04.022](https://doi.org/10.1016/j.jhydrol.2015.04.022).
- Dahlgren, P., P. Källberg, T. Landelius, and P. Undén, 2014: Comparison of the regional reanalyses products with newly developed and existing state-of-the art systems. EURO4M Project Rep. 2.9, 19 pp., <http://www.euro4m.eu/Deliverables.html>.
- Dawdy, D. R., and J. M. Bergmann, 1969: Effect of rainfall variability on streamflow simulation. *Water Resour. Res.*, **5**, 958–966, doi:[10.1029/WR005i005p00958](https://doi.org/10.1029/WR005i005p00958).
- Dayon, G., J. Boé, and E. Martin, 2015: Transferability in the future climate of a statistical downscaling method for precipitation in France. *J. Geophys. Res. Atmos.*, **120**, 1023–1043, doi:[10.1002/2014JD022236](https://doi.org/10.1002/2014JD022236).
- Dee, D. P., and Coauthors, 2011: The ERA-Interim reanalysis: Configuration and performance of the data assimilation system. *Quart. J. Roy. Meteor. Soc.*, **137**, 553–597, doi:[10.1002/qj.828](https://doi.org/10.1002/qj.828).
- Durand, Y., E. Brun, L. Merindol, G. Guyomarch, B. Lesaffre, and E. Martin, 1993: A meteorological estimation of relevant parameters for snow models. *Ann. Glaciol.*, **18**, 65–71, doi:[10.1017/S0260305500011277](https://doi.org/10.1017/S0260305500011277).
- Edijatno, N., X. Yang, Z. Makhlof, and C. Michel, 1999: GR3J: A daily watershed model with three free parameters. *Hydrol. Sci. J.*, **44**, 263–277, doi:[10.1080/02626669909492221](https://doi.org/10.1080/02626669909492221).
- Essou, G. R. C., R. Arsenault, and F. P. Brissette, 2016a: Comparison of climate datasets for lumped hydrological

- modeling over the continental United States. *J. Hydrol.*, **537**, 334–345, doi:[10.1016/j.jhydrol.2016.03.063](https://doi.org/10.1016/j.jhydrol.2016.03.063).
- , F. Sabarly, P. Lucas-Picher, F. Brissette, and A. Poulin, 2016b: Can precipitation and temperature from meteorological reanalyses be used for hydrological modeling? *J. Hydrometeorol.*, **17**, 1929–1950, doi:[10.1175/JHM-D-15-0138.1](https://doi.org/10.1175/JHM-D-15-0138.1).
- Garcia, M., C. D. Peters-Lidard, and D. C. Goodrich, 2008: Spatial interpolation of precipitation in a dense gauge network for monsoon storm events in the southwestern United States. *Water Resour. Res.*, **44**, W05S13, doi:[10.1029/2006WR005788](https://doi.org/10.1029/2006WR005788).
- Gupta, H. V., H. Kling, K. K. Yilmaz, and G. F. Martinez, 2009: Decomposition of the mean squared error and NSE performance criteria: Implications for improving hydrological modelling. *J. Hydrol.*, **377**, 80–91, doi:[10.1016/j.jhydrol.2009.08.003](https://doi.org/10.1016/j.jhydrol.2009.08.003).
- Habets, F., and Coauthors, 2008: The SAFRAN-ISBA-MODCOU hydrometeorological model applied over France. *J. Geophys. Res.*, **113**, D06113, doi:[10.1029/2007JD008548](https://doi.org/10.1029/2007JD008548).
- Häggmark, L., K.-I. Ivarsson, S. Gollvik, and P.-O. Olofsson, 2000: MESAN: An operational mesoscale analysis system. *Tellus*, **52A**, 2–20, doi:[10.3402/tellusa.v52i1.12250](https://doi.org/10.3402/tellusa.v52i1.12250).
- Haylock, M. R., N. Hofstra, A. M. G. Klein Tank, E. J. Klok, P. D. Jones, and M. New, 2008: A European daily high-resolution gridded data set of surface temperature and precipitation for 1950–2006. *J. Geophys. Res.*, **113**, D20119, doi:[10.1029/2008JD010201](https://doi.org/10.1029/2008JD010201).
- Isotta, F. A., R. Vogel, and C. Frei, 2015: Evaluation of European regional reanalyses and downscalings for precipitation in the Alpine region. *Meteor. Z.*, **24**, 15–37, doi:[10.1127/metz/2014/0584](https://doi.org/10.1127/metz/2014/0584).
- Jones, R. W., I. A. Renfrew, A. Orr, B. G. M. Webber, D. M. Holland, and M. A. Lazzara, 2016: Evaluation of four global reanalysis products using in situ observations in the Amundsen Sea Embayment, Antarctica. *J. Geophys. Res. Atmos.*, **121**, 6240–6257, doi:[10.1002/2015JD024680](https://doi.org/10.1002/2015JD024680).
- Klemes, V., 1986: Operational testing of hydrological simulation models. *Hydrol. Sci. J.*, **31**, 13–24, doi:[10.1080/0266668609491024](https://doi.org/10.1080/0266668609491024).
- Landelius, T., P. Dahlgren, S. Gollvik, A. Jansson, and E. Olsson, 2016: A high-resolution regional reanalysis for Europe. Part 2: 2D analysis of surface temperature, precipitation and wind. *Quart. J. Roy. Meteor. Soc.*, **142**, 2132–2142, doi:[10.1002/qj.2813](https://doi.org/10.1002/qj.2813).
- Loikith, P. C., D. E. Waliser, J. Kim, and R. Ferraro, 2017: Evaluation of cool season precipitation event characteristics over the Northeast US in a suite of downscaled climate model hindcasts. *Climate Dyn.*, doi:[10.1007/s00382-017-3837-0](https://doi.org/10.1007/s00382-017-3837-0), in press.
- Lorenc, A. C., 1981: A global three-dimensional multivariate statistical interpolation scheme. *Mon. Wea. Rev.*, **109**, 701–721, doi:[10.1175/1520-0493\(1981\)109<0701:AGTDSM>2.0.CO;2](https://doi.org/10.1175/1520-0493(1981)109<0701:AGTDSM>2.0.CO;2).
- Martin, E., and Coauthors, 2016: On the use of hydrological models and satellite data to study the water budget of river basins affected by human activities: Examples from the Garonne basin of France. *Surv. Geophys.*, **37**, 223–247, doi:[10.1007/s10712-016-9366-2](https://doi.org/10.1007/s10712-016-9366-2).
- Oudin, L., F. Hervieu, C. Michel, C. Perrin, V. Andréassian, F. Anctil, and C. Loumagne, 2005: Which potential evapotranspiration input for a lumped rainfall–runoff model?: Part 2—Towards a simple and efficient potential evapotranspiration model for rainfall–runoff modelling. *J. Hydrol.*, **303**, 290–306, doi:[10.1016/j.jhydrol.2004.08.026](https://doi.org/10.1016/j.jhydrol.2004.08.026).
- , C. Perrin, T. Mathevet, V. Andréassian, and C. Michel, 2006: Impact of biased and randomly corrupted inputs on the efficiency and the parameters of watershed models. *J. Hydrol.*, **320**, 62–83, doi:[10.1016/j.jhydrol.2005.07.016](https://doi.org/10.1016/j.jhydrol.2005.07.016).
- Perrin, C., C. Michel, and V. Andréassian, 2003: Improvement of a parsimonious model for streamflow simulation. *J. Hydrol.*, **279**, 275–289, doi:[10.1016/S0022-1694\(03\)00225-7](https://doi.org/10.1016/S0022-1694(03)00225-7).
- Piani, C., G. P. Weedon, M. Best, S. M. Gomes, P. Viterbo, S. Hagemann, and J. O. Haerter, 2010: Statistical bias correction of global simulated daily precipitation and temperature for the application of hydrological models. *J. Hydrol.*, **395**, 199–215, doi:[10.1016/j.jhydrol.2010.10.024](https://doi.org/10.1016/j.jhydrol.2010.10.024).
- Poli, P., and Coauthors, 2016: ERA-20C: An atmospheric reanalysis of the twentieth century. *J. Climate*, **29**, 4083–4097, doi:[10.1175/JCLI-D-15-0556.1](https://doi.org/10.1175/JCLI-D-15-0556.1).
- Prein, A. F., and A. Gobiet, 2017: Impacts of uncertainties in European gridded precipitation observations on regional climate analysis. *Int. J. Climatol.*, **37**, 305–327, doi:[10.1002/joc.4706](https://doi.org/10.1002/joc.4706).
- Quintana-Seguí, P., and Coauthors, 2008: Analysis of near-surface atmospheric variables: Validation of the SAFRAN analysis over France. *J. Appl. Meteor. Climatol.*, **47**, 92–107, doi:[10.1175/2007JAMC1636.1](https://doi.org/10.1175/2007JAMC1636.1).
- Randall, D. A., and Coauthors, 2007: Climate models and their evaluation. *Climate Change 2007: The Physical Science Basis*, S. Solomon et al., Eds., Cambridge University Press, 589–662.
- Soci, C., E. Bazile, F. Besson, and T. Landelius, 2016: High-resolution precipitation re-analysis system for climatological purposes. *Tellus*, **68A**, 29879, doi:[10.3402/tellusa.v68.29879](https://doi.org/10.3402/tellusa.v68.29879).
- Sperna Weiland, F. C., J. A. Vrugt, R. L. P. H. van Beek, A. H. Weerts, and M. F. P. Bierkens, 2015: Significant uncertainty in global scale hydrological modeling from precipitation data errors. *J. Hydrol.*, **529**, 1095–1115, doi:[10.1016/j.jhydrol.2015.08.061](https://doi.org/10.1016/j.jhydrol.2015.08.061).
- Valéry, A., V. Andréassian, and C. Perrin, 2014: ‘As simple as possible but not simpler’: What is useful in a temperature-based snow-accounting routine? Part 1 – Comparison of six snow accounting routines on 380 catchments. *J. Hydrol.*, **517**, 1166–1175, doi:[10.1016/j.jhydrol.2014.04.059](https://doi.org/10.1016/j.jhydrol.2014.04.059).
- Vautard, R., T. Noël, L. Li, M. Vrac, E. Martin, P. Dandin, J. Cattiaux, and S. Joussaume, 2013: Climate variability and trends in downscaled high-resolution simulations and projections over metropolitan France. *Climate Dyn.*, **41**, 1419–1437, doi:[10.1007/s00382-012-1621-8](https://doi.org/10.1007/s00382-012-1621-8).
- Vidal, J.-P., E. Martin, L. Franchistéguy, M. Baillon, and J.-M. Soubeyrou, 2010: A 50-year high-resolution atmospheric reanalysis over France with the Safran system. *Int. J. Climatol.*, **30**, 1627–1644, doi:[10.1002/joc.2003](https://doi.org/10.1002/joc.2003).
- Weedon, G. P., S. Gomes, P. Viterbo, H. Österle, J. C. Adam, N. Bellouin, O. Boucher, and M. Best, 2010: The WATCH forcing data 1958–2001: A meteorological forcing dataset for land surface and hydrological models. WATCH Tech. Rep. 22, 41 pp., <http://www.eu-watch.org/publications/technical-reports>.
- , and Coauthors, 2011: Creation of the WATCH forcing data and its use to assess global and regional reference crop evaporation over land during the twentieth century. *J. Hydrometeorol.*, **12**, 823–848, doi:[10.1175/2011JHM1369.1](https://doi.org/10.1175/2011JHM1369.1).
- , G. Balsamo, N. Bellouin, S. Gomes, M. J. Best, and P. Viterbo, 2014: The WFDEI meteorological forcing data set: WATCH forcing data methodology applied to ERA-Interim reanalysis data. *Water Resour. Res.*, **50**, 7505–7514, doi:[10.1002/2014WR015638](https://doi.org/10.1002/2014WR015638).
- You, Q., J. Min, W. Zhang, N. Pepin, and S. Kang, 2015: Comparison of multiple datasets with gridded precipitation observations over the Tibetan Plateau. *Climate Dyn.*, **45**, 791–806, doi:[10.1007/s00382-014-2310-6](https://doi.org/10.1007/s00382-014-2310-6).
- Zambrano-Bigiarini, M., 2014: hydroGOF: Goodness-of-fit functions for comparison of simulated and observed hydrological time series. R package, accessed 19 December 2016, <https://cran.r-project.org/web/packages/hydroGOF/index.html>.