Evaluating the Suitability of Century-Long Gridded Meteorological Datasets for Hydrological Modeling

CAROLINA MASSMANN

*Institute for Hydrology and Water Management, University of Natural Resources and Life Sciences, Vienna, Austria,
and Department of Civil Engineering, University of Bristol, Bristol, United Kingdom*

(Manuscript received 20 May 2019, in final form 17 January 2020)

ABSTRACT

Recent advances in climate reanalyses have led to the development of meteorological products providing information from the beginning of the last century or even before. As these data sources might be of interest to practitioners in the event of missing data from meteorological stations, it is important to assess their usefulness for different applications. The main objective of this study is to investigate the ability of two long-term reanalysis datasets (CERA-20C and 20CR) and one long-term interpolated dataset (Livneh) for supporting hydrological modeling. The precipitation and temperature data of the three datasets were first compared, downscaled, and then used as inputs to the conceptual hydrological model HBV in 168 basins in the United States. The findings suggest that the quality of all three datasets decreases the further we go back in time. Models calibrated at the beginning of the time series, where the data quality is worse, are only able to capture the general properties of the time series and thus do not show a decrease in performance as the period between calibration and validation becomes larger. The opposite is true for models calibrated at the end of the time series, which show a clear decrease in performance toward the beginning of the century. While the hydrological model driven with the interpolated datasets achieved the best performance, the results obtained with the reanalysis datasets were still informative (i.e., better than the long-term monthly mean), and they matched the performance of the interpolated dataset in a few catchments in the northwestern United States.

1. Introduction

Precipitation and temperature information is especially important in water-related studies as data on these variables are the minimum requirement for setting up and running simple hydrological models. Since many study areas do not have nearby weather stations providing reliable data, it is increasingly common to rely on gridded meteorological datasets. While it is desirable to have data for both variables, precipitation data are more critical as hydrological models tend to be more sensitive to different precipitation than to different temperature

datasets (Essou et al. 2016; Laiti et al. 2018). This discrepancy can be primarily explained by the fact that discharge is driven by precipitation, but it is also the result of the higher homogeneity between temperature datasets in comparison to precipitation datasets (Essou et al. 2016; Behnke et al. 2016).

As the differences between gridded precipitation datasets can be better understood when their data source is considered, the next paragraphs describe the three groups into which meteorological datasets can be classified according to this criterion.

Datasets derived from ground stations interpolate data from weather stations onto a grid, taking into account the impact of orography on the variation and distribution of the considered climate variables. The quality of the product depends on the density and representativeness of the ground stations and on the quality of their readings. In this context, Laiti et al. (2018) found that higher densities of ground stations resulted in better temperature and precipitation estimates. At locations with complex topography, which tend to have a lower density of stations,

---

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

Supplemental information related to this paper is available at the Journals Online website: <https://doi.org/10.1175/JHM-D-19-0113.s1>.

Corresponding author: Carolina Massmann, carolina.massmann@boku.ac.at

DOI: 10.1175/JHM-D-19-0113.1

© 2020 American Meteorological Society. For information regarding reuse of this content and general copyright information, consult the [AMS Copyright Policy](#) (www.ametsoc.org/PUBSReuseLicenses).

Henn et al. (2018) observed more pronounced deviations between precipitation datasets. Regarding the quality of the ground station readings, it is well known that undercatch can have a high influence on measured precipitation. This is especially the case when it falls as snow, in which event errors can reach up to 20%–50% in windy conditions (Rasmussen et al. 2012). Since these errors are carried onto the gridded dataset, we can expect that the datasets will underestimate precipitation in snowy areas (Henn et al. 2018). Other factors influencing the final product are the underlying digital elevation model, the interpolation approach, and the methodology for considering orography.

Datasets derived from modern reanalyses are the result of a combination (through weighted averaging) of different types of observations (e.g., from climate stations, satellites, ships, and balloons) with the a priori predictions of a numerical weather prediction model (Compo et al. 2011). Regarding the seasonal variations in the quality of these estimates, it has been observed that reanalysis datasets tend to better reproduce precipitation in winter, when large-scale systems—which can be well predicted—are the main precipitation source. In summer, when small-scale convective systems dominate, reanalysis data have difficulties correctly reproducing precipitation (Beck et al. 2019). Reanalysis datasets have been used intensively in hydrology, and numerous studies have found that they give acceptable results when used as inputs to hydrological models (Choi et al. 2009; Lauri et al. 2014; Auerbach et al. 2016; Essou et al. 2016). In some cases, hydrological models forced with reanalysis data achieved even better performances than models using data from weather stations. This was the case when the density of weather stations was low and when stations were located in places that are not representative of the average catchment, for instance, when they lie in the valley but an important proportion of the catchment is at higher elevations (Auerbach et al. 2016; Krogh et al. 2015).

Datasets derived from remote sensing sources process data obtained from infrared (IR), passive microwave, and active microwave sensors (Sun et al. 2018). In general, microwave sensors are more suitable for deriving precipitation information than IR sensors (Beck et al. 2017), and almost all satellite products have better performances in summer than in winter (Ciabatta et al. 2017; Beck et al. 2019). This is explained by the sensors' ability to detect

convective storms, which are mostly observed in summer, and by difficulties in identifying snowfall and low-intensity precipitation, which hamper their usefulness in winter. As the large-scale use of sensors started approximately 40 years ago, remote sensing-based meteorological datasets tend to start after 1978 and are therefore not relevant to the present study focusing on datasets starting at the beginning of the last century.

The large number of available datasets makes it difficult to decide a priori on adequate datasets for a given application. Intercomparison studies can offer some relief by analyzing differences between datasets or comparing the products to a benchmark (Beck et al. 2017). Of special hydrological interest are studies using the datasets as inputs to hydrological models and evaluating their quality by assessing the differences between simulated and observed discharges (Lauri et al. 2014; Auerbach et al. 2016; Essou et al. 2016; Beck et al. 2017; Laiti et al. 2018). This approach, however, adds additional uncertainties to the analyses as the results might vary depending on the hydrological model and its parameterization (Mendoza et al. 2015), the potential evapotranspiration (PET) approach (Sperna Weiland et al. 2012), the techniques used for preprocessing the meteorological time series (e.g., bias correction as in Laiti et al. 2018) and the selected ancillary datasets describing, for example, the soil properties (Livneh et al. 2015b).

Most model intercomparison studies have been carried out for the last 35 years as this is the period covered by the largest number of datasets. As the starting date of reanalysis datasets shifts continuously backward, there is an increasing need to carry out evaluations starting at earlier dates. Until recently, all reanalysis products started after 1947, when a substantial increase in the number of upper air observations came about due to the intensified use of radiosondes (Compo et al. 2011). These are devices carrying temperature, humidity, pressure and, optionally, wind sensors providing readings at high vertical resolutions and above 30 km of elevation (Brettle and Galvin 2003). Recent developments in reanalysis products have extended their temporal coverage. This has resulted in historical reanalysis datasets providing data for the complete twentieth century. Extending the coverage of reanalysis products further into the past requires a large effort in data rescue activities, which comprise the identification, scanning, digitalization, and quality checking of potentially useful data sources (Dupigny-Giroux et al. 2007). While this refers primarily to older data sources, it also applies to satellite data as there are huge

amounts of data in older formats and tapes not yet readily available (Poli et al. 2015). The recovery of data from these sources has a high priority, as documentation is often incomplete and there is a decline in the expertise about older instruments, explaining why major efforts are devoted to data rescue activities (Poli et al. 2017). Given that pressure differences have a substantial impact on atmospheric flow, surface pressure data have emerged as a valuable data source (Whitaker et al. 2004). By relying only on surface pressure observations, it is possible to obtain good surface weather maps for the Northern Hemisphere from the late nineteenth century on (Whitaker et al. 2004; Compo et al. 2006). The errors of these maps are of similar magnitude to the errors of current 2–3-day forecasts (Whitaker et al. 2004). This explains the large efforts incurred for obtaining more surface pressure data measurements from land and marine sources, as well as the best tracking pressure data from tropical cyclone observations and reports. These efforts have resulted in a large database, the International Surface Pressure Databank, which stores data between 1768 and 2012 (Cram et al. 2015) and constitutes the backbone of historical reanalysis products.

As these newly developed long-term (historical) reanalysis datasets have been mostly evaluated regarding their agreement with large-scale atmospheric patterns (Laloyaux et al. 2018; Compo et al. 2011), it is unclear if they are adequate for hydrological purposes. The main objective of this study is to assess the potential of two long-term reanalysis datasets for supporting hydrological modeling. The analysis is complemented by one long-term interpolated dataset so that the study considers both data sources that are available for long-term hydrological studies, namely, interpolated and reanalysis datasets. The analysis was carried out in two phases. The first phase compared the precipitation and temperature of these three datasets in the last 30 years with a benchmark dataset in 550 catchments in the United States. In the second step, the three long-term datasets were used to force a hydrological model in 168 catchments with more than 70 years of discharge data.

2. Meteorological datasets

This study compares three daily long-term datasets starting at the beginning of the twentieth century and uses one additional dataset as a benchmark. Long-term interpolated datasets at a monthly scale were not considered in the present study [e.g., the CRU datasets, GPCC, Global Historical Climatology Network–Monthly (GHCN-M), and University of Delaware datasets described in Harris et al. (2014), Becker et al. (2013),

Menne et al. (2018), and Willmott and Matsuura 2001, respectively].

a. Interpolated datasets

1) LIVNEH DATASET

Maurer et al. (2002) developed a gridded dataset at a daily time scale for the conterminous United States. Livneh et al. (2013) increased the temporal coverage of the dataset to the period spanning from 1915 to 2011 and increased its spatial resolution to $1/16^\circ$, while following the original methodology as far as possible. A newer version of the dataset (Livneh et al. 2015a) includes Mexico and parts of Canada and thus reduces transboundary discontinuities. However, as this dataset only begins in 1950, it is less suitable for investigating the usefulness for long-term modeling than the previous dataset starting in 1915.

The dataset was derived from precipitation and temperature (minimum and maximum) readings from all COOP (Cooperative Observer Program) stations with more than 20 years of valid data. The data were gridded using the SYMAP (synergraphic mapping algorithm) described by Shepard (1984). The gridded precipitation was then rescaled to match the monthly Parameter-Elevation Regressions on Independent Slopes Model (PRISM) data (Daly et al. 1994). As PRISM data do not correct for snow undercatch, it was expected that the Livneh dataset would also underestimate solid precipitation (Maurer et al. 2002).

2) DAYMET 3 DATASET

The Daymet 3 dataset (Thornton and Running 1999; Thornton et al. 1997, 2000, 2018) was used as a benchmark for evaluating the performance of the other three datasets between 1980 and 2010. The reason for selecting this dataset as benchmark is that it was successfully used in a previous study modeling numerous catchments in the United States (Newman et al. 2015). The dataset provides daily information for seven weather variables in North America from 1980 until today. It was developed by the Oak Ridge National Laboratory Distributed Active Archive Center (ORNL DAAC) and is based on station data of the GHCN-Daily dataset. The data are interpolated with an iterative method that depends on station density [i.e., a smaller (larger) search radius in areas with a high (low) density of weather stations] and that uses the spatial convolution of a truncated Gaussian filter. No additional scaling or correction is performed on the precipitation data.

b. Reanalysis datasets

Two groups have developed reanalysis datasets for the complete twentieth century: the European Centre

for Medium-Range Weather Forecasts (ECMWF) and a cooperation between the National Oceanic and Atmospheric Administration (NOAA) and the University of Colorado.

1) ERA-20 AND CERA-20 DATASETS

The ECMWF released the ERA-20C dataset covering the period between 1900 and 2010 and the CERA-20C dataset with data from 1901 to 2010. The main difference between these datasets is that ERA is based on uncoupled climate models, while CERA is produced by a coupled atmosphere–ocean model (Laloyaux et al. 2018). Uncoupled models do not take the feedback between the ocean and the atmosphere into account, while in CERA different modules interact, making it possible for changes in atmospheric states to directly impact the ocean states and vice versa (Laloyaux et al. 2017). It is thereby expected that coupled models will provide more consistent results and reduce the initialization shock, which is observed when the states of the ocean and atmosphere are not consistent during the initialization of the system (Laloyaux and Dee 2015). More details about the coupling of the components can be found in Laloyaux et al. (2016).

The CERA-20C dataset assimilates surface pressure data, marine wind observations, ocean salinity and temperature profiles using a 4D-Var assimilation approach. In contrast, the ERA-Interim dataset, a modern reanalysis dataset also produced by the ECMWF, primarily assimilates data from satellites (clear-sky radiance, atmospheric motion vectors, scatterometer wind data, ozone retrievals, total precipitable vapor estimates), GPS radio occultation devices (atmospheric refraction) and conventional sources comprising data from pilot balloons, aircraft, wind profilers, ships, drifting buoys and land stations (upper-air temperatures, wind, air-specific humidity, surface pressure, 2-m temperature, 2-m relative humidity, and near-surface winds) (Dee et al. 2011).

Besides considering fewer input variables, historical reanalyses need to manage with fewer available observations in the earlier part of the century, as illustrated by Laloyaux et al. (2018). For example, wind and surface pressure buoy measurements are only available starting from the 1970s, while ship-based measurements of ocean temperature and salinity increased from less than 500 in 1901 to more than 200 000 in 1981. Historical reanalysis datasets, therefore, cannot be expected to compete in quality with modern reanalysis datasets. Their value lies in the long period they cover.

Both datasets (ERA-20C and CERA-20C) are composed of a set of 10 ensemble members, which differ in the initial conditions of the system. This study analyses

the average value of all CERA ensemble members downloaded at $0.125^\circ \times 0.125^\circ$ resolution from the ECMWF web page.

2) TWENTIETH CENTURY REANALYSIS PROJECT DATASETS

The Physical Sciences Laboratory (PSL) of NOAA and the Cooperative Institute for Research in Environmental Sciences (CIRES) of the University of Colorado Boulder have released three versions of the Twentieth Century Reanalysis Project (20CR) dataset (<https://psl.noaa.gov/data/>). The first version covers the period between 1908 and 1958, the second version runs between 1871 and 2012, and the 2c version provides data from between 1851 and 2014. This study used the data from the second version, at a spatial resolution of $2^\circ \times 2^\circ$. This dataset relies on a new version of the NCEP Global Forecast System (GFS) coupled atmosphere–land model for generating a first guess. Observations of surface pressure and sea level pressure observations and reports were then assimilated using an ensemble Kalman filter approach (Compo et al. 2011). The interpolated monthly sea surface temperature and sea ice concentration fields from the Hadley Centre Sea Ice were used as boundary conditions. While the model is run for 56 ensemble members, it is also possible to download the ensemble average, which was used in this study.

3. Methodology

a. Catchments

The first part of this study analyzed the differences in the precipitation and temperature data of 550 GAGES II (Geospatial Attributes of Gauges for Evaluating Streamflow) reference catchments spread across the United States. As they are reference basins, they are among those with the most natural flow. More specifically, we expect them to have limited anthropogenic influence. In the second part, a hydrological model was set up for the 168 basins with available long-term discharge measurements [see section 3d(2)]. The daily discharge time series were downloaded from the U.S. Geological Survey (USGS) web page.

b. Hydrological model

The HBV conceptual hydrological model was used in this study (Bergström 1992; Lindström et al. 1997) in the Python implementation developed by Ayzel (2016a,b). This model was selected because it is fast, is used often in hydrological studies and has been successfully applied in dataset intercomparison studies (Beck et al. 2017).

The model consists of a snow module and a soil module that is made up of two reservoirs connected

through percolation flow. The upper reservoir releases water through evapotranspiration, fast flow, and inter-flow; the release of the lower reservoir can be regarded as groundwater flow. The model was set up as a lumped model with the catchments described as single units accepting as inputs the areal average (over the catchment) of the daily precipitation and the minimum and maximum temperatures. The PET was calculated from the [Hargreaves and Samani \(1982\)](#) formula using the approach outlined by [Allen et al. \(1998\)](#). The model was calibrated with the SCE algorithm ([Duan et al. 1994](#)), the parameter limits suggested by [Ayzel \(2016b\)](#) and using the Nash–Sutcliffe efficiency (NSE) as the objective function.

c. Bias correction and downscaling of the meteorological datasets

The precipitation and temperature time series of the three long-term datasets were downscaled to the Daymet 3 grid, a process which also corrects the bias of the datasets. The approach followed here is based on the empirical transformation of [Panofsky and Brier \(1968\)](#), which is often used in climate studies for correcting the distribution of simulated variables. The transformation matches the simulated distribution with the observed one, a process that is done quantilewise, meaning that the quantiles are independently corrected. The main assumption behind this approach is that climate models can correctly simulate the ranks of the values but not the values themselves ([Feigenwinter et al. 2018](#)).

The study used nonparametric quantile mapping based on empirical quantiles, as [Gudmundsson et al. \(2012\)](#) found that nonparametric approaches were preferable to parametric approaches. A wet day correction, accounting for the excess of wet days simulated by climate models (drizzle effect), was also implemented for the downscaling of precipitation ([Maraun 2016](#)). The correction consists of a simple computation in which precipitation values are sorted in increasing order and the days with the least precipitation are set to zero until the number of days with precipitation in the simulated dataset equals the number of days with precipitation in the observed time series.

The observational dataset was resampled 20 times (bootstrap with replacement), and the quantiles were calculated for each sample. These 20 quantile values were then averaged and used as the observational dataset. The modeled quantiles were directly estimated from the long-term datasets. The empirical cumulative distribution function for the observed and modeled quantiles was then estimated with the R-package implementation “qmap” ([Gudmundsson 2016](#)). This estimate was done using the data for the period between 1995 and 2010 so that it agrees with the second

calibration period [see [section 3d\(2\)](#)] and then used for downscaling the complete precipitation and temperature time series.

d. Study setup

The main objective of this study was to assess if different long-term datasets (CERA-20C, 20CR, and Livneh) are adequate for long-term hydrological modeling. The study consisted of two main parts: in the first, the meteorological variables were compared, while in the second, the model performance that can be achieved with these datasets was determined.

1) COMPARISON OF THE METEOROLOGICAL VARIABLES

As the true value of the meteorological variables is unknown, the Daymet 3 dataset was used as a benchmark with respect to which the meteorological data of the long-term datasets were assessed. This analysis was carried out for 550 catchments in the United States and considered the period between 1980 and 2010. This study relied on four metrics describing different aspects of the meteorological variables that are relevant for hydrological purposes:

Bias can be regarded as a long-term indicator of model performance. Its estimation differs for precipitation and temperature as the former considers the relative bias, while the latter considers the nominal bias:

$$\text{bias}_{\text{pp}} = \frac{\overline{\text{pp}}_s - \overline{\text{pp}}_o}{\overline{\text{pp}}_o}, \quad (1)$$

$$\text{bias}_{\text{te}} = \overline{\text{te}}_s - \overline{\text{te}}_o, \quad (2)$$

where $\overline{\text{pp}}_s$ ($\overline{\text{te}}_s$) and $\overline{\text{pp}}_o$ ($\overline{\text{te}}_o$) are the mean simulated and observed precipitation (temperature), respectively.

Spearman’s rank correlation coefficient for the daily data describes the ability of the dataset to reproduce the daily dynamics.

The mean monthly error (MME) in the winter and summer seasons provides information about the temporal distribution of errors. The MME for the precipitation and temperature in the summer months was calculated using Eqs. (3) and (4):

$$\text{MME}_{\text{summer,pp}} = \frac{\sum_{i=1}^n (\text{pp}_{s,i} - \text{pp}_{o,i})}{\sum_{i=1}^d \text{pp}_{o,i}}, \quad (3)$$

$$\text{MME}_{\text{summer,temp}} = \frac{\sum_{i=1}^n (\text{pp}_{s,i} - \text{pp}_{o,i})}{n}, \quad (4)$$

where n is the number of days in June–August and d is the total number of days in the time series. Analogous formulas were used for the winter season.

2) COMPARISON OF THE PERFORMANCE ACHIEVED BY THE HYDROLOGICAL MODEL FORCED WITH THREE DIFFERENT LONG-TERM METEOROLOGICAL DATASETS

The potential of different datasets for supporting hydrological modeling can be best evaluated by directly analyzing the performance achieved with hydrological models driven with data derived from each dataset. As the main focus of this study lies in long-term hydrological modeling, this part of the study considered all 168 catchments with more than 25 567 days (~ 70 years) of discharge before 2010. The days with available discharge were not necessarily consecutive as many catchments contain gaps in their discharge time series.

The hydrological model was calibrated twice following a split sample approach (Klemeš 1986). The first calibration period consists of the first 16 years with discharge data availability—which vary depending on the catchment—and the second period comprises the years between 1995 and 2010. The models were then run with each calibrated parameter set for the period between 1916 and 2010. The model was calibrated twice to investigate the quality of the datasets at the beginning of the century. By calibrating the model in the last 16 years and then using the optimal parameter sets for modeling the discharge 70 years earlier, we can expect that the performance will be worse than closer to the calibration period. This might be caused by changes in the quality of the datasets or by changes in the climate, catchments or measurements that result in earlier periods that are increasingly out of sample.

Because it is much easier to achieve a high NSE in catchments with pronounced seasonal variability than in catchments with discharge fluctuating around a more or less constant annual value (Schaeffli and Gupta 2007), model performance was evaluated with the mean monthly based NSE_b :

$$NSE_b = 1 - \frac{\sum_{t=1}^N [Q_o(t) - Q_s(t)]^2}{\sum_{t=1}^N [Q_o(t) - Q_{om}(t)]^2}, \quad (5)$$

where $Q_o(t)$ and $Q_s(t)$ are the observed and simulated daily discharges at time t and Q_{om} equals the long-term mean monthly flow of the corresponding month. The results for the standard NSE are presented in the online supplemental material to facilitate comparisons with previous studies that used the NSE for evaluating model results.

Since previous studies found that downscaled meteorological inputs resulted in more realistic simulations of hydrological models (Wilby et al. 2000), the analysis was carried out once for the raw data (i.e., as retrieved from the long-term datasets) and then repeated for the downscaled data.

4. Results and discussion

a. Comparing the average climate variables between 1980 and 2010

Figure 1 compares the raw precipitation data (i.e., not bias corrected) obtained with the three long-term datasets with the Daymet 3 data. The Livneh dataset agreed best with the benchmark as it had the lowest bias and seasonal errors and the highest correlation.

Regarding bias, the CERA and 20CR datasets tended to underestimate precipitation. An exception was observed in the central part of the country, where the 20CR dataset clearly overestimated precipitation, while the CERA datasets had smaller absolute biases. An analysis of the seasonal performance (the last two rows in Fig. 1) indicated that in the West, the differences with the benchmark were mostly relevant in winter, while the errors in the Central and East regions were more pronounced in summer. The biases in the Southeast can be explained by the land–atmosphere interactions, which affected the parameterization of the model and the resulting local precipitation patterns. This was enhanced by the strong humidity gradients and large moisture fluxes (Essou et al. 2016). In the West, precipitation is regular, abundant, and strongly linked to the Pacific Ocean. Essou et al. (2016) explain that reanalysis products have difficulties estimating the amount of precipitation between autumn and spring in this area. This result agrees with the results of the present study as the largest bias in the West was observed in winter.

The correlation of the Livneh dataset reached values around 0.6 across the entire country. The 20CR dataset showed similar correlations to the Livneh dataset on the West Coast but displayed lower values (around 0.4) in the Central region. The correlation increased again farther east without, however, reaching the correlations observed for the Livneh dataset. In a similar way, the correlation for the CERA dataset reached the highest values at the West Coast, but the values were overall lower than those observed with the other two datasets. Correlation decreased toward the east and exhibited values close to zero in the Northeast. These results agree with those of Essou et al. (2016), who found that the daily correlation of precipitation tended to be high in the West and lower in the Central and Southeast regions of the United States for all the considered datasets.

FIG. 1. Comparison of the precipitation derived from the three datasets with the benchmark, the Daymet 3 precipitation. The plots show (row 1) the long-term bias, (row 2) the daily rank correlation, and the MME in the (row 3) summer and (row 4) winter months.

Figure 2 shows the results for the raw temperature of the datasets (i.e., not bias corrected). The Livneh dataset, with smallest biases and highest correlations, had the best performance. The catchments with the largest absolute magnitude of the long-term biases were located in the Rocky Mountains for all three datasets. Regarding the performance in different seasons, there were no large differences in the error patterns between summer and winter for the 20CR and Livneh datasets. On the contrary, the CERA dataset overestimated summer temperatures in the West while underestimating temperatures in the winter season. The seasonal variability of the absolute magnitude of the biases tended to be larger in summer than in winter for the reanalysis datasets, while the Livneh dataset showed larger absolute biases in winter. The annual temperature patterns of the Livneh dataset agreed, therefore, better with the finding of Essou et al. (2016), who observed that temperature biases were larger in winter than in summer.

The daily correlation coefficients for the temperature were relatively high for all datasets across the United States compared to the correlations obtained for precipitation. This confirms the observations by Essou et al.

(2016) and Laiti et al. (2018), who found that the differences between the datasets were smaller for temperature than for precipitation.

It is important to note that the benchmark dataset (Daymet 3) does not correspond to the true values, which are unknown, but were derived from station data. The fact that both the Daymet 3 and Livneh datasets are interpolated might explain the closer agreement between these datasets. Furthermore, neither the Daymet 3 nor Livneh datasets corrected for snow undercatch. The biases observed in mountainous regions for the Livneh dataset indicate, therefore, only that both datasets agree well, not that there is no actual bias.

b. Impact of bias correction on model performance

This section compares the performance of the hydrological model driven first with the raw and then with the bias-corrected climate inputs. The performance achieved with the raw Livneh dataset agreed well with the performance obtained with the Daymet 3 dataset (Fig. 3). These results match the patterns described by Newman et al. (2015), who observed that the model had the best results in the Northwest and Southeast regions, while the

FIG. 2. Comparison of the temperature derived from the three datasets with the benchmark, the Daymet 3 temperature. The plots show (row 1) the long-term bias, (row 2) the daily rank correlation, and the MME in the (row 3) summer and (row 4) winter months.

lowest performance was obtained in the Great Plains. Using the two raw reanalysis datasets as model inputs resulted in much lower model performances in most parts of the United States. The performances of both datasets agree well, although the correlations (for precipitation and temperature) are higher for the 20CR than for the CERA datasets in the Central and East regions of the country. These results indicate that the correlation values might not be necessarily a good predictor of model performance (see next section for a detailed discussion).

The impact of using bias-corrected input data instead of raw data is shown in the third row of Fig. 3. The impact was small for the Livneh dataset, where the bias-corrected dataset led to slight increases in model performance in the southern half of the United States, while there was a decrease in model performance in some catchments located in the northern half. The situation was different for the reanalysis datasets, which evidenced a large increase in the NSE values in the West and a slight, but consistent decrease in the Gulf Coast. In the West, where performance increased after correcting the bias, the precipitation of the reanalysis datasets had high correlation values (the

second row in Fig. 1) and also high biases (the first row in Fig. 1). This finding agrees with the results of Raimonet et al. (2017), who found that the main factor explaining the low performance of hydrological models is errors in the short-term dynamics of the precipitation input (i.e., the correlation). This indicates that bias correction might increase model performance if the underlying data have an adequate quality and the main problem is the existence of a bias (Laiti et al. 2018; Essou et al. 2016). The results obtained in the Central and East regions, however, indicate that information about the bias and correlation of the meteorological variables does not suffice for predicting the need for bias correction as it is observed that the impact of bias correction between CERA and 20CR datasets does not differ much here, regardless of the higher correlations obtained by the 20CR datasets.

Figure 4 quantifies the changes in performance due to the use of bias-corrected inputs instead of raw inputs. There were decreases in the observed performance in a small number of catchments, no changes for most catchments and considerable increases in performance for 50–60 catchments.

FIG. 3. Nash–Sutcliffe efficiency (NSE_b) achieved during calibration (1995–2010) using (top) the Daymet 3 meteorological dataset and (middle) the raw (i.e., not bias corrected) three long-term datasets. (bottom) The difference in the NSE_b between the model calibrated with the bias-corrected meteorological data and the raw data. Positive differences indicate that the model driven with the downscaled data achieved a better performance.

c. Long-term performance of the hydrological model driven with different datasets

The long-term model performance achieved with the bias-corrected time series was analyzed for the complete length of the available discharge time series (Fig. 5). The first row of plots shows the results when the model is calibrated in the first 16 years with discharge data; the second row of plots shows the results when the model is calibrated in the second calibration period. Hydrographs for a few watersheds are presented in the supplemental material.

The model driven with the Livneh dataset performed significantly better than the models driven with the reanalysis datasets. A comparison between the CERA and the 20CR results indicates that the CERA datasets resulted in better performances. This is somewhat surprising, as Figs. 1 and 2 show that the 20CR dataset had higher correlations for precipitation and temperature than the CERA dataset. One factor explaining this might be that Figs. 1 and 2 show the correlation between the reanalysis datasets and the benchmark (Daymet 3 dataset) but not with the true, unknown, meteorological conditions. Although the high correlation between the Livneh and the Daymet 3 datasets coupled with the high performances achieved with the Livneh dataset might

suggest that the Daymet 3 data is a good representation of the true meteorological variables, this might not necessarily be the case as it is possible that model parameters compensate for errors in the model inputs. Furthermore, correlation relationships are not transitive, meaning that if we assume that the Daymet 3 dataset is correlated with the true value of the meteorological variables, it does not follow from this that a dataset with a high correlation with the Daymet 3 dataset will also have a high correlation with the true values of the variables.

FIG. 4. CDF plot showing the improvement in the calibrated NSE_b value when using bias-corrected meteorological inputs instead of the raw inputs. Negative values indicate that the NSE achieved with raw inputs is higher than the NSE_b achieved with the bias-corrected inputs.

FIG. 5. NSE_b value achieved for each catchment in 12 periods from 1915 to 2010 with the bias-corrected climate datasets. Periods without discharge data have a white color. (a) The results obtained with the three datasets with the model calibrated in the first 16 years with discharge data; (b) the results of the model calibrated in the last 16 years with discharge. The y axis depicts the values for the catchments sorted by increasing runoff coefficient.

Another factor that could have influenced the results was the use of the NSE_b as objective function for calibration and for evaluating the model. It has been observed that evaluation criteria squaring the errors might be sensitive to the results of a small fraction of the considered time period (Berthet et al. 2010). The 20CR dataset could thus have problems in the representation of extreme precipitation events that strongly affect

overall performance when the model is calibrated using an objective function that gives a high weight to these extreme errors.

Performance tended to be higher during calibration than during validation when the model was calibrated using the last 16 years of data (the second row of plots). While this agrees with usual observations in hydrological modeling (e.g., Klemes 1986), it seems not to hold for

FIG. 6. Boxplots representing the distribution of the NSE_b values in 12 periods. (a) The results obtained with the three datasets with the model calibrated in the first 16 years with discharge data; (b) the results of the model calibrated in the last 16 years with discharge.

the models calibrated during the first 16 years, as confirmed by the boxplots presented in Fig. 6. This could be explained by the low quality of the data at the beginning of the twentieth century, which hinders an adequate fit of the models. The models only captured in this case the general properties of the climate time series, resulting in models that are less sensitive to changes in climate between the calibration and validation periods and, in this way, have a more constant performance in time. On the other hand, the models calibrated in the last 16 years show a trend of decreasing performance as the length of the period between calibration and validation increases. In other words, the NSE value is lower at the beginning of the century than toward the end. This is caused by model calibration, which adjusts the parameters to the conditions observed in the calibration period and results in a decrease in performance in the noncalibrated periods. As the probability of being out of sample due to changes in the climate, catchment properties or the approaches used for measuring climate variables or discharge (Merz et al. 2011; Kling et al. 2012) increases with the length of time between calibration and validation, there is a related increase in the probability of having a decrease in model performance. A factor that might

explain the decrease in the performance of the model driven by the Livneh dataset is the use of PRISM monthly normals for rescaling the precipitation. As these normals are computed for 30 years, they are less representative as we move away from the midpoint of the averaging period.

An analysis of the temporal performance patterns of both calibration periods helps to visualize the impact of data quality on model performance. Figure 6 shows that the performance patterns achieved with the CERA dataset with the parameters of the first and second calibration periods agree well. For example, in the left panels for the CERA datasets, the median of the bars increases from 1919 to 1935, decreases in 1943 and increases again until 1959. This is quantified in Fig. 7, which shows a histogram of the rank correlation of model performance achieved with both parameter sets (i.e., one for each calibration period) in each catchment. The patterns agree best for the CERA and 20CR datasets as they have a high fraction of catchments with correlations above 0.75. A possible explanation for the lower correlation observed for the 20CR than the CERA dataset in Fig. 7 might be related to the high variance in the NSE values for the 1935 and 1943 periods (Fig. 6b),

FIG. 7. Histogram for the rank correlation between the models calibrated at the beginning of the time series and the end of the time series.

which could indicate the presence of larger errors in the climate time series that are compensated when the calibration period includes these periods (Fig. 6a) but lead to large decreases in performance when these periods are used to validate the model.

The correlations tend to be lower for the Livneh dataset. This might be explained again by a better fit of the hydrological model forced with the Livneh dataset to the characteristics of the calibration period. The meteorological properties in the calibration period do have, therefore, a larger impact on the modeled hydrograph and on model performance in the validation period. The reanalysis datasets, on the other hand, do not allow for such a close fit of the model parameters. This means that these vary more consistently in time as they are less constrained by the properties in the calibration period and that they primarily reflect the impact of climate variability. This suggests that models driven with the reanalysis datasets might be better suited for investigating how specific characteristics of the time series (climate and discharge) influence models results (e.g., Merz et al. 2011; Li et al. 2015).

From a spatial perspective, the Livneh dataset had a high correlation for temperature and precipitation across the entire United States, while there were clear differences in model performance (Fig. 3) that do not match the correlation patterns but which are explained by model deficiencies, for example in modeling arid (Atkinson et al. 2002) or groundwater dominated watersheds (Massmann 2020).

d. Contributions of long-term climate datasets to hydrological knowledge

Long-term datasets are a valuable source of information for understanding slow processes and extreme

events (e.g., floods) and identifying subtle and complex phenomena, of which climate change is the best example (Burt 1994). Long-term studies and datasets further constitute a vivid example of the impact that past conditions and interventions have on the present state of the environment. The dissemination of such examples might facilitate the engagement of the general population in activities aimed at mitigating and adapting to climate change.

Besides climate and discharge time series, it is desirable to have access to the related metadata of these measurements. Information about changes in land cover and land use, water use and agricultural practices is further needed to understand the relationships between society and hydrological systems. Finally, information about specific events (e.g., floods and droughts), as well as measurement data for other variables (e.g., groundwater levels and snow extent and depth), can also be critical for improving our understanding and representation of hydrologic systems (Kling et al. 2012).

It is often said that the past is the key to the future. This applies to hydrology as our understanding of how previous changes in climate have affected hydrological processes can be applied to extrapolate what we can expect in the future. However, there are large differences in the uncertainties involved in modeling past and future climates, and the lessons we learn about modeling the past might be less relevant to helping us model the future than expected. The largest uncertainties in the CERA dataset are caused by a lack of information about the initial state of the ocean. This is why 10 probable ocean states for 1900 were used to initialize the CERA model, which resulted in a 10-member ensemble (de Boissésón and Balmaseda 2016). The spread of this ensemble decreases with time as more and more observations are assimilated into the model. There is also uncertainty regarding the model variability and a small uncertainty regarding the emissions of climate change gases. In contrast, for climate change projections, we have little uncertainty regarding the initial state, some uncertainty related to internal model variability (which is especially important for middle-term simulations) and large uncertainties regarding future emissions, which are especially relevant for long-term projections (Northrop and Chandler 2014).

e. Limitations of the study

The main limitation of this study is that it focuses only on the uncertainties of the meteorological dataset, disregarding the impact of other sources of uncertainty, such as the model structure, model parameterization or approach for estimating the PET.

As the Penman–Monteith formula is physically based, it is the recommended approach for estimating the PET. Its disadvantage is that it requires data for a large number of climate variables, hindering its use in data-scarce situations (Hargreaves and Allen 2003). In these cases, it is recommended to use the Hargreaves formula, which has much lower data requirements and gives results that agree well with the results obtained with the Penman–Monteith formula in a large range of climates (Hargreaves and Allen 2003). While it would have been possible to obtain data for additional climate variables (e.g., radiation and humidity), thus enabling the use of the Penman–Monteith approach, it was decided to rely on the Hargreaves formula instead. In this way, the results between the datasets depended only on temperature and precipitation, which are the climate variables that hydrological models are most sensitive to and which, therefore, were of the most interest to this study.

While it would be interesting to repeat the analysis with different PET formulas, it is not expected that this would lead to large changes in the results and conclusions. Previous studies have found that different PET formulas do indeed exhibit large differences in the estimated PET, but these differences become less important when models are calibrated to match the observed discharge (Sperna Weiland et al. 2012; Seiller and Anctil 2016).

Other factors influencing the outcome of this study were the selection of the hydrological model and its parameterization. Regarding model complexity, Raimonet et al. (2017) hypothesize that simpler hydrological models (i.e., models having few parameters) are more adequate for evaluating input datasets as their fewer degrees of freedom make it less likely that they will compensate for errors in the dataset. The calibration approach (Mendoza et al. 2015) and approach used for parameterizing parameters that are not calibrated (Livneh et al. 2015b) also influence the results. However, as the main objective of this study was to evaluate the meteorological datasets rather than find the best model and calibration approach for each catchment, it was decided to focus on one model and calibration approach.

5. Summary and conclusions

This study investigated three long-term meteorological datasets: two reanalysis products starting at the beginning of last century (CERA and 20CR) and one interpolated dataset (Livneh) starting in 1915. The main aim was to discover how well they could support long-term hydrological modeling.

The first part of the study compared the precipitation and temperature time series of the long-term datasets

with the Daymet 3 dataset, which was used as a benchmark. The results agreed with previous studies, indicating that temperature is better estimated than precipitation. Regarding the data sources, the interpolated dataset agreed well with the benchmark for both climate variables across the whole country. The reanalysis products are adequate for modeling the short-term dynamics of precipitation in the Northwest but had difficulties in the Central and East regions of the United States. The temperature dynamics were adequately reproduced across the United States with lower agreements in the Rocky Mountains.

Regarding the data preprocessing steps, the precipitation bias correction had a much larger impact on the performance of the reanalysis datasets than on the interpolated dataset. While the improvements resulting from the bias correction were spatially concentrated in the West, where the correlation and the bias with the benchmark were high, it was seen in the Central and East regions of the country that higher correlations with the Daymet 3 benchmark do not guarantee that the performance will improve after correcting the bias.

The best model performance was achieved with the Livneh dataset. This result was not unexpected as it is an interpolated dataset and thus relied directly on available temperature and precipitation measurements, whereas the reanalysis datasets primarily used pressure data to estimate these variables. The analysis further suggests that the optimal calibration approach of the models might depend on the objective of the study as models focusing on the relationship between climate variability and model performance might benefit from a “light” calibration approach that avoids overfitting. On the other hand, studies designed for detecting trends and investigating the continuous decrease in performance in time due to the increase of the period between calibration and validation could take advantage of parameter sets that are optimized for achieving the best agreement between the observed and simulated datasets in the calibration period.

This study showed that the model driven with the reanalysis datasets achieved, in most cases, lower performances than the model driven with the interpolated dataset. The reanalysis datasets had, however, in most cases a mean monthly based NSE value larger than zero, indicating that the reanalysis datasets provide information that can be helpful for hydrological purposes as the model results represent an improvement over the long-term monthly averages. In small areas in the Northwest, the reanalysis datasets achieved performances comparable to the interpolated dataset. This is an encouraging result, which raises hope that in the future, as more

analog measurements are digitized and made available to developers of climate models, the quality of the estimates for other areas and for earlier time periods could be improved.

Acknowledgments. Support for the 20CR dataset was provided by the United States Department of Energy's Office of Science Innovative and Novel Computational Impact on Theory and Experiment program, the Office of Biological and Environmental Research (BER), and the National Oceanic and Atmospheric Administration Climate Program Office. I thank Andrew Newman and two anonymous reviewers for providing many detailed comments that helped to improve this manuscript considerably. This project was funded by the Austrian Science Fund (FWF) project J-3689-N29.

Data availability statement. The primary data used in the study are publicly available from U.S. organizations. More detailed information about how it can be accessed is found in the text. Processed data are available from the author upon request.

REFERENCES

- Allen, R. G., L. S. Pereira, D. Raes, and M. Smith, 1998: Crop evapotranspiration: Guidelines for computing crop water requirements. FAO Irrigation and Drainage Paper 56, 300 pp., www.fao.org/docrep/X0490E/X0490E00.htm.
- Atkinson, S. E., R. A. Woods, and M. Sivapalan, 2002: Climate and landscape controls on water balance model complexity over changing timescales. *Water Resour. Res.*, **38**, 1314, <https://doi.org/10.1029/2002WR001487>.
- Auerbach, D. A., Z. M. Easton, M. T. Walter, A. S. Flecker, and D. R. Fuka, 2016: Evaluating weather observations and the climate forecast system reanalysis as inputs for hydrologic modelling in the tropics. *Hydrol. Processes*, **30**, 3466–3477, <https://doi.org/10.1002/hyp.10860>.
- Ayzel, G., 2016a: LHMP: First major release. Zenodo, <https://doi.org/10.5281/zenodo.59680>.
- , 2016b: LHMP: Lumped hydrological modelling playground. Zenodo, <https://doi.org/10.5281/zenodo.59501>.
- Beck, H. E., and Coauthors, 2017: Global-scale evaluation of 22 precipitation datasets using gauge observations and hydrological modelling. *Hydrol. Earth Syst. Sci.*, **21**, 6201–6217, <https://doi.org/10.5194/hess-21-6201-2017>.
- , and Coauthors, 2019: Daily evaluation of 26 precipitation datasets using Stage-IV gauge-radar data for the CONUS. *Hydrol. Earth Syst. Sci.*, **23**, 207–224, <https://doi.org/10.5194/hess-23-207-2019>.
- Becker, A., P. Finger, A. Meyer-Christoffer, B. Rudolf, K. Schamm, U. Schneider, and M. Ziese, 2013: A description of the global land-surface precipitation data products of the Global Precipitation Climatology Centre with sample applications including centennial (trend) analysis from 1901–present. *Earth Syst. Sci. Data*, **5**, 71–99, <https://doi.org/10.5194/essd-5-71-2013>.
- Behnke, R., S. Vavrus, A. Allstadt, T. Albright, W. E. Thogmartin, and V. C. Radeloff, 2016: Evaluation of downscaled, gridded climate data for the conterminous United States. *Ecol. Appl.*, **26**, 1338–1351, <https://doi.org/10.1002/15-1061>.
- Bergström, S., 1992: The HBV model – Its structure and applications. SMHI Rep. RH4, Swedish Meteorological and Hydrological Institute, Norrköping, Sweden, 44 pp., https://www.smhi.se/polopoly_fs/1.83592!/Menu/general/extGroup/attachmentColHold/mainCol1/file/RH_4.pdf.
- Berthet, L., V. Andréassian, C. Perrin, and C. Loumagne, 2010: How significant are quadratic criteria? Part II On the relative contribution of large flood events to the value of a quadratic criterion. *Hydrol. Sci. J.*, **55**, 1063–1073, <https://doi.org/10.1080/02626667.2010.505891>.
- Bettle, M. J., and J. F. P. Galvin, 2003: Back to basics: Radiosondes: Part I – The instrument. *Weather*, **58**, 336–341, <https://doi.org/10.1256/wea.126.02A>.
- Burt, T. P., 1994: Long-term study of the natural environment - Perceptive science or mindless monitoring? *Prog. Phys. Geogr.*, **18**, 475–496, <https://doi.org/10.1177/030913339401800401>.
- Choi, W., S. J. Kim, P. F. Rasmussen, and A. R. Moore, 2009: Use of the North American regional reanalysis for hydrological modelling in Manitoba. *Can. Water Resour. J.*, **34**, 17–36, <https://doi.org/10.4296/cwrj3401017>.
- Ciabatta, L., A. C. Marra, G. Panegrossi, D. Casella, P. Sanò, S. Dietrich, C. Massari, and L. Brocca, 2017: Daily precipitation estimation through different microwave sensors: Verification study over Italy. *J. Hydrol.*, **545**, 436–450, <https://doi.org/10.1016/j.jhydrol.2016.12.057>.
- Compo, G. P., J. S. Whitaker, and P. D. Sardeshmukh, 2006: Feasibility of a 100-year reanalysis using only surface pressure data. *Bull. Amer. Meteor. Soc.*, **87**, 175–190, <https://doi.org/10.1175/BAMS-87-2-175>.
- , and Coauthors, 2011: The Twentieth Century Reanalysis project. *Quart. J. Roy. Meteor. Soc.*, **137**, 1–28, <https://doi.org/10.1002/qj.776>.
- Cram, T. A., and Coauthors, 2015: The International Surface Pressure Databank version 2. *Geosci. Data J.*, **2**, 31–46, <https://doi.org/10.1002/gdj3.25>.
- Daly, C., R. P. Neilson, and D. L. Phillips, 1994: A statistical-topographic model for mapping climatological precipitation over mountainous terrain. *J. Appl. Meteor.*, **33**, 140–158, [https://doi.org/10.1175/1520-0450\(1994\)033<0140:ASTMFM>2.0.CO;2](https://doi.org/10.1175/1520-0450(1994)033<0140:ASTMFM>2.0.CO;2).
- de Boissésou E., and M. Balmaseda, 2016: An ensemble of 20th century ocean reanalyses for providing ocean initial conditions for CERA-20C coupled streams. ERA Rep. Series 24, 35 pp., <https://www.ecmwf.int/node/16456>.
- 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, <https://doi.org/10.1002/qj.828>.
- Duan, Q., S. Sorooshian, and V. K. Gupta, 1994: Optimal use of the SCE-UA global optimization method for calibrating watershed models. *J. Hydrol.*, **158**, 265–284, [https://doi.org/10.1016/0022-1694\(94\)90057-4](https://doi.org/10.1016/0022-1694(94)90057-4).
- Dupigny-Giroux L.-A., T. F. Ross, J. D. Elms, R. Truesdell, and S. R. Doty, 2007: NOAA's Climate Database Modernization Program: Rescuing, archiving, and digitizing history. *Bull. Amer. Meteor. Soc.*, **88**, 1015–1017, <https://doi.org/10.1175/1520-0477-88.7.1007>.
- Essou, G. R., F. Sabarly, P. Lucas-Picher, F. Brissette, and A. Poulin, 2016: Can precipitation and temperature from meteorological reanalyses be used for hydrological modeling? *J. Hydrometeorol.*, **17**, 1929–1950, <https://doi.org/10.1175/JHM-D-15-0138.1>.

- Feigenwinter, I., S. Kotlarski, A. Casanueva, A. M. Fischer, C. Schwierz, and M. A. Liniger, 2018: Exploring quantile mapping as a tool to produce user-tailored climate scenarios for Switzerland. Tech. Rep. MeteoSwiss 270, 44 pp., https://www.meteoschweiz.admin.ch/content/dam/meteoswiss/en/service-und-publikationen/publikationen/doc/MeteoSchweiz_Fachbericht_270_final.pdf.
- Gudmundsson, L., 2016: qmap: Statistical transformations for post-processing climate model output, version 1.0-4. R package, <https://cran.r-project.org/web/packages/qmap/qmap.pdf>.
- , J. B. Bremnes, J. E. Haugen, and T. Engen-Skaugen, 2012: Technical note: Downscaling RCM precipitation to the station scale using statistical transformations - A comparison of methods. *Hydrol. Earth Syst. Sci.*, **16**, 3383–3390, <https://doi.org/10.5194/hess-16-3383-2012>.
- Hargreaves, G. H., and Z. A. Samani, 1982: Estimating potential evapotranspiration. *J. Irrig. Drain. Div.*, **108**, 225–230.
- , and R. G. Allen, 2003: History and evaluation of Hargreaves evapotranspiration equation. *J. Irrig. Drain. Eng.*, **129**, 53–63, [https://doi.org/10.1061/\(ASCE\)0733-9437\(2003\)129:1\(53\)](https://doi.org/10.1061/(ASCE)0733-9437(2003)129:1(53)).
- Harris, I., P. Jones, T. Osborn, and D. Lister, 2014: Updated high-resolution grids of monthly climatic observations – The CRU TS3.10 dataset. *Int. J. Climatol.*, **34**, 623–642, <https://doi.org/10.1002/joc.3711>.
- Henn, B., A. J. Newman, B. Livneh, C. Daly, and J. D. Lundquist, 2018: An assessment of differences in gridded precipitation datasets in complex terrain. *J. Hydrol.*, **556**, 1205–1219, <https://doi.org/10.1016/j.jhydrol.2017.03.008>.
- Klemeš, V., 1986: Operational testing of hydrological simulation models. *Hydrol. Sci. J.*, **31**, 13–24, <https://doi.org/10.1080/02626668609491024>.
- Kling, H., M. Fuchs, and M. Paulin, 2012: Runoff conditions in the upper Danube basin under an ensemble of climate change scenarios. *J. Hydrol.*, **424–425**, 264–277, <https://doi.org/10.1016/j.jhydrol.2012.01.011>.
- Krogh, S. A., J. W. Pomeroy, and J. McPhee, 2015: Physically based mountain hydrological modeling using reanalysis data in Patagonia. *J. Hydrometeor.*, **16**, 172–193, <https://doi.org/10.1175/JHM-D-13-0178.1>.
- Laiti, L., S. Mallucci, S. Piccolroaz, A. Bellin, D. Zardi, A. Fiori, G. Nikulin, and B. Majone, 2018: Testing the hydrological coherence of high-resolution gridded precipitation and temperature data sets. *Water Resour. Res.*, **54**, 1999–2016, <https://doi.org/10.1002/2017WR021633>.
- Laloyaux, P., and D. Dee, 2015: CERA: A coupled data assimilation system for climate reanalysis. *ECMWF Newsletter*, No. 144, ECMWF, Reading, United Kingdom, 15–20, <https://doi.org/10.21957/0dx9o7bg>.
- , M. Balmaseda, D. Dee, K. Mogensen, and P. Janssen, 2016: A coupled data assimilation system for climate reanalysis. *Quart. J. Roy. Meteor. Soc.*, **142**, 65–78, <https://doi.org/10.1002/qj.2629>.
- , F. de Boissésón, and P. Dahlgren, 2017: CERA-20C: An earth system approach to climate reanalysis. *ECMWF Newsletter*, No. 150, ECMWF, Reading, United Kingdom, 25–30, <https://doi.org/10.21957/ffs36birj2>.
- , and Coauthors, 2018: CERA-20C: A coupled reanalysis of the twentieth century. *J. Adv. Model. Earth Syst.*, **10**, 1172–1195, <https://doi.org/10.1029/2018MS001273>.
- Lauri, H., T. A. Räsänen, and M. Kumm, 2014: Using reanalysis and remotely sensed temperature and precipitation data for hydrological modeling in monsoon climate: Mekong River case study. *J. Hydrometeor.*, **15**, 1532–1545, <https://doi.org/10.1175/JHM-D-13-084.1>.
- Li, H., S. Beldring, and C.-Y. Xu, 2015: Stability of model performance and parameter values on two catchments facing changes in climatic conditions. *Hydrol. Sci. J.*, **60**, 1317–1330, <https://doi.org/10.1080/02626667.2014.978333>.
- 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**, 272–288, [https://doi.org/10.1016/S0022-1694\(97\)00041-3](https://doi.org/10.1016/S0022-1694(97)00041-3).
- Livneh, B., E. A. Rosenberg, C. Lin, B. Nijssen, V. Mishra, K. M. Andreadis, E. P. Maurer, and D. P. Lettenmaier, 2013: A long-term hydrologically based dataset of land surface fluxes and states for the conterminous United States: Update and extensions. *J. Climate*, **26**, 9384–9392, <https://doi.org/10.1175/JCLI-D-12-00508.1>.
- , T. J. Bohn, D. S. Pierce, F. Munoz-Ariola, B. Nijssen, R. Vose, D. Cayan, and L. D. Brekke, 2015a: A spatially comprehensive, hydrometeorological data set for Mexico, the U.S., and southern Canada 1950–2013. *Sci. Data*, **2**, 150042, <https://doi.org/10.1038/sdata.2015.42>.
- , R. Kumar, and L. Samaniego, 2015b: Influence of soil textural properties on hydrologic fluxes in the Mississippi river basin. *Hydrol. Processes*, **29**, 4638–4655, <https://doi.org/10.1002/hyp.10601>.
- Maraun, D., 2016: Bias correcting climate change simulations - A critical review. *Curr. Climate Change Rep.*, **2**, 211–220, <https://doi.org/10.1007/s40641-016-0050-x>.
- Massmann, C., 2020: Identification of factors influencing hydrologic model performance using a top-down approach in a large number of U.S. catchments. *Hydrol. Processes*, **34**, 4–20, <https://doi.org/10.1002/hyp.13566>.
- Maurer, E. P., A. W. Wood, J. C. Adam, B. Nijssen, and D. P. Lettenmaier, 2002: A long-term hydrologically based dataset of land surface fluxes and states for the conterminous United States. *J. Climate*, **15**, 3237–3251, [https://doi.org/10.1175/1520-0442\(2002\)015<3237:ALTHBD>2.0.CO;2](https://doi.org/10.1175/1520-0442(2002)015<3237:ALTHBD>2.0.CO;2).
- Mendoza, P. A., and Coauthors, 2015: Effects of hydrologic model choice and calibration on the portrayal of climate change impacts. *J. Hydrometeor.*, **16**, 762–780, <https://doi.org/10.1175/JHM-D-14-0104.1>.
- Menne, M. J., C. N. Williams, B. E. Gleason, J. J. Rennie, and J. H. Lawrimore, 2018: The Global Historical Climatology Network monthly temperature dataset, version 4. *J. Climate*, **31**, 9835–9854, <https://doi.org/10.1175/JCLI-D-18-0094.1>.
- Merz, R., J. Parajka, and G. Blöschl, 2011: Time stability of catchment model parameters: Implications for climate impact analyses. *Water Resour. Res.*, **47**, W02531, <https://doi.org/10.1029/2010WR009505>.
- Newman, A. J., and Coauthors, 2015: Development of a large-sample watershed-scale hydrometeorological data set for the contiguous USA: Data set characteristics and assessment of regional variability in hydrologic model performance. *Hydrol. Earth Syst. Sci.*, **19**, 209–223, <https://doi.org/10.5194/hess-19-209-2015>.
- Northrop, P. J., and R. E. Chandler, 2014: Quantifying sources of uncertainty in projections of future climate. *J. Climate*, **27**, 8793–8808, <https://doi.org/10.1175/JCLI-D-14-00265.1>.
- Panofsky, H. A., and G. W. Brier, 1968: *Some Applications of Statistics to Meteorology*. Earth and Mineral Sciences Continuing Education, College of Earth and Mineral Sciences, 224 pp.
- Poli, P., R. Saunders, and D. Sante, 2015: Rescuing satellite data for climate reanalysis. *ECMWF Newsletter*, No. 144, ECMWF, Reading, United Kingdom, 8–9, <https://www.ecmwf.int/node/14588>.

- , and Coauthors, 2017: Recent advances in satellite data rescue. *Bull. Amer. Meteor. Soc.*, **98**, 1471–1484, <https://doi.org/10.1175/BAMS-D-15-00194.1>.
- Raimonet, M., L. Oudin, V. Thieu, M. Silvestre, R. Vautard, C. Rabouille, and P. Le Moigne, 2017: Evaluation of gridded meteorological datasets for hydrological modeling. *J. Hydrometeor.*, **18**, 3027–3041, <https://doi.org/10.1175/JHM-D-17-0018.1>.
- Rasmussen, R., and Coauthors, 2012: How well are we measuring snow: The NOAA/FAA/NCAR winter precipitation test bed. *Bull. Amer. Meteor. Soc.*, **93**, 811–829, <https://doi.org/10.1175/BAMS-D-11-00052.1>.
- Schaeffli, B., and H. V. Gupta, 2007: Do Nash values have value? *Hydrol. Processes*, **21**, 2075–2080, <https://doi.org/10.1002/hyp.6825>.
- Seiller, G., and F. Anctil, 2016: How do potential evapotranspiration formulas influence hydrological projections? *Hydrol. Sci. J.*, **61**, 2249–2266, <https://doi.org/10.1080/02626667.2015.1100302>.
- Shepard, D. S., 1984: Computer mapping: The SYMAP interpolation algorithm. *Spatial Statistics and Models*, G. L. Gaile and C. J. Willmott, Eds., Theory and Decision Library, Vol. 40, Springer, 133–145, https://doi.org/10.1007/978-94-017-3048-8_7.
- Sperna Weiland, F. C., C. Tisseuil, H. H. M. Dürr, M. Vrac, and L. P. H. van Beek, 2012: Selecting the optimal method to calculate daily global reference potential evaporation from CFSR reanalysis data for application in a hydrological model study. *Hydrol. Earth Syst. Sci.*, **16**, 983–1000, <https://doi.org/10.5194/hess-16-983-2012>.
- Sun, Q., C. Miao, Q. Duan, H. Ashouri, S. Sorooshian, and K.-L. Hsu, 2018: A review of global precipitation data sets: Data sources, estimation, and intercomparisons. *Rev. Geophys.*, **56**, 79–107, <https://doi.org/10.1002/2017RG000574>.
- Thornton, P. E., and S. W. Running, 1999: An improved algorithm for estimating incident daily solar radiation from measurements of temperature, humidity, and precipitation. *Agric. For. Meteorol.*, **93**, 211–228, [https://doi.org/10.1016/S0168-1923\(98\)00126-9](https://doi.org/10.1016/S0168-1923(98)00126-9).
- , —, and M. A. White, 1997: Generating surfaces of daily meteorological variables over large regions of complex terrain. *J. Hydrol.*, **190**, 214–251, [https://doi.org/10.1016/S0022-1694\(96\)03128-9](https://doi.org/10.1016/S0022-1694(96)03128-9).
- , H. Hasenauer, and M. A. White, 2000: Simultaneous estimation of daily solar radiation and humidity from observed temperature and precipitation: An application over complex terrain in Austria. *Agric. For. Meteorol.*, **104**, 255–271, [https://doi.org/10.1016/S0168-1923\(00\)00170-2](https://doi.org/10.1016/S0168-1923(00)00170-2).
- , M. M. Thornton, B. W. Mayer, Y. Wei, R. Devarakonda, R. S. Vose, and R. B. Cook, 2018: Daymet: Daily surface weather data on a 1-km grid for North America, version 2. Oak Ridge National Laboratory Distributed Active Archive Center, accessed 25 October 2016, <https://doi.org/10.3334/ORNLDAAAC/1219>.
- Whitaker, J. S., G. P. Compo, X. Wei, and T. M. Hamill, 2004: Reanalysis without radiosondes using ensemble data assimilation. *Mon. Wea. Rev.*, **132**, 1190–1200, [https://doi.org/10.1175/1520-0493\(2004\)132<1190:RWRUED>2.0.CO;2](https://doi.org/10.1175/1520-0493(2004)132<1190:RWRUED>2.0.CO;2).
- Wilby, R. L., L. E. Hay, W. J. Gutowski, R. W. Arritt, E. S. Takle, Z. Pan, G. H. Leavesley, and M. P. Clark, 2000: Hydrological responses to dynamically and statistically downscaled climate model output. *Geophys. Res. Lett.*, **27**, 1199–1202, <https://doi.org/10.1029/1999GL006078>.
- Willmott, C. J., and K. Matsuura, 2001: Terrestrial precipitation: 1900–2017 gridded monthly time series. Dept. of Geography, University of Delaware, http://climate.geog.udel.edu/~climate/html_pages/Global2017/README.GlobalTsP2017.html.