Earth Syst. Sci. Data, 13, 1531–1545, 2021
https://doi.org/10.5194/essd-13-1531-2021

Open Access Earth System Science Data

HydroGFD3.0 (Hydrological Global Forcing Data): a 25 km global precipitation and temperature data set updated in near-real time

Peter Berg, Fredrik Almén, and Denica Bozhinova

Swedish Meteorological and Hydrological Institute, Folkborgsvägen 17, 60176 Norrköping, Sweden

Correspondence: Peter Berg (peter.berg@smhi.se)

Received: 13 August 2020 – Discussion started: 21 September 2020
Revised: 3 February 2021 – Accepted: 11 February 2021 – Published: 13 April 2021

Abstract. HydroGFD3 (Hydrological Global Forcing Data) is a data set of bias-adjusted reanalysis data for daily precipitation and minimum, mean, and maximum temperature. It is mainly intended for large-scale hydrological modelling but is also suitable for other impact modelling. The data set has an almost global land area coverage, excluding the Antarctic continent and small islands, at a horizontal resolution of 0.25°, i.e. about 25 km. It is available for the complete ERA5 reanalysis time period, currently 1979 until 5 d ago. This period will be extended back to 1950 once the back catalogue of ERA5 is available. The historical period is adjusted using global gridded observational data sets, and to acquire real-time data, a collection of several reference data sets is used. Consistency in time is attempted by relying on a background climatology and only making use of anomalies from the different data sets. Precipitation is adjusted for mean bias as well as the number of wet days in a month. The latter is relying on a calibrated statistical method with input only of the monthly precipitation anomaly such that no additional input data about the number of wet days are necessary. The daily mean temperature is adjusted toward the monthly mean of the observations and applied to 1 h time steps of the ERA5 reanalysis. Daily mean, minimum, and maximum temperature are then calculated. The performance of the HydroGFD3 data set is on par with other similar products, although there are significant differences in different parts of the globe, especially where observations are uncertain. Further, HydroGFD3 tends to have higher precipitation extremes, partly due to its higher spatial resolution. In this paper, we present the methodology, evaluation results, and how to access the data set at https://doi.org/10.5281/zenodo.3871707 (Berg et al., 2020).

1 Introduction

Precipitation (P) and temperature (T) are key driving parameters for many impact models, and there are now many observational data sets available. They differ regarding the spatio-temporal resolution, the historical coverage, and the data sources included in the product. However, when it comes to continuously updated near-real-time data sets, there are very few available data sets. It is therefore challenging to find a product suitable for monitoring and initialization of forecasts for an impact model, i.e. a product that fulfils both a long historical period for calibration and validation as well as real-time updates.

While most data sets now offer a rather long historical period, the real-time availability is a greater challenge. Merged satellite and gauge data sets such as CHIRPS (Funk et al., 2015a), CMORPH (Joyce et al., 2004), and PERSIANN-CDR (Ashouri et al., 2015) offer both high-resolution and near-real-time components but are limited to between the ±50° or ±60° latitude bands. Several data sets have made use of reanalysis data as a basis, adjusted using various gridded observational data sets (Weedon et al., 2011, 2014; Beck et al., 2017; Berg et al., 2018; Cucchi et al., 2020). The advantage is that the reanalysis products are readily available with a large range of variables and output frequencies. Still, the downside with reanalysis products is that especially P is a model product and thereby suffers from model bias. Since the bias can be substantial, several methods have been developed to adjust reanalysis using different methods and reference data sets.

A hydrological-operational-monitoring or forecast product has strong demands on availability and redundancy of the data flows. The data set HydroGFD1 (Hydrological Global Forcing Data; Berg et al., 2018) was constructed and made operational for initializations of the hydrological model HYPE (HYdrological Predictions for the Environment) (Lindström et al., 2010) for different set-ups across the globe. It offered near-real-time updating of daily P and daily T (mean, minimum, and maximum) until the end of the last calendar month. The real-time components of HydroGFD1 were based on ERA-Interim reanalysis, extended by the ECMWF deterministic forecasts and adjusted using monthly mean P from GPCC-Monitoring and GPCC-FirstGuess (Schneider et al., 2018b) products and monthly mean T from GHCN–CAMS (the Global Historical Climate Network combined with the Climate Anomaly Monitoring System) (Fan and Van den Dool, 2008). The follow-up data set HydroGFD2 offered some updates to the methodology and shifted to using primarily the CPC-Unified (Chen et al., 2008) and CPC-Temp (CPCtemp, 2017) products for P and T adjustments, respectively. Both data sets employed a 0.5° resolution and have been operationally produced for a few years now, and we have identified some serious issues regarding the availability of required data sets for successful updates. The largest operational intermission occurred during the government lockdown in the US between 22 December 2018 and 25 January 2019. Neither of the US data sets included in the production were then available, which hampered the production of the HydroGFD data sets and subsequently deteriorated the quality of some operational HYPE models. Both these HydroGFD versions have now become obsolete for real-time production due to the discontinuation of the ERA-Interim production as of August 2019. Data sets using multiple input data sources are less sensitive to such conditions, such as the MSWEP (Multi-Source Weighted-Ensemble Precipitation) data set (Beck et al., 2017).

In this paper, the HydroGFD3.0 system is described, with its range of produced data sets for the period 1979 to near real time at 0.25° resolution and global land coverage. We describe the methodology and the operational production as well as an evaluation of the climatological data set, with comparison to other similar data sources.

2 Data

Table 1 lists the data sources used in the production of the different *tiers* (i.e. production lines with different data sets; see Methods section) of HydroGFD3. From now on, we use the shortened internal abbreviations listed under "Name" in Table 1 when we refer to the data of P or T from each source. ERA5 is the latest global reanalysis product of the ECMWF (Hersbach et al., 2020) and forms the basis for HydroGFD3. This reanalysis product is chosen because our operational forecasts at the SMHI (Swedish Meteorological and Hydrological Institute) are based on the medium-range forecasts of the ECMWF, with the same model as that used for ERA5 and a similar bias, although there are differences in model version. Other reanalysis products would be possible but are not explored here. ERA5 is updated with a 3-month lag, but a new temporary product, ERA5T, is produced with a 5 d lag.

As described in Sect. 3, HydroGFD3 is based on a combination of the ERA5 reanalysis with the different data sets as listed in the top section of Table 1. In the following analysis, we compare the different data sets included in the processing and additionally make a state-of-the-art comparison to the data set WFDE5 (Cucchi et al., 2020), which is a new product using the *WATCH forcing data* methodology (Weedon et al., 2011) with ERA5 reanalysis, listed in the bottom section of Table 1.

An issue with global-scale evaluations is that of independence between data sets, and most of the gauge-based data sets listed in Table 1 make use of more or less the same openly available observations, with regional differences. The data sets have, however, been independently generated and use different statistical models for the gridding process. Our aim is to provide a comprehensive overview of HydroGFD3 in comparison to other data sets in order to present its qualities and to point out potential issues. For each of the comparisons in Sect. 5, we chose data sets that are as independent as possible given the limitations just discussed. Our experience from earlier studies is that in-depth evaluation can only be performed at the local scale (e.g. Fallah et al., 2020), and we encourage users of the data set to pursue such evaluations.

3 Method

The main method that HydroGFD3 is building on consists of adding observational monthly anomalies to a background climatology, then adjusting the reanalysis data to that absolute monthly mean. Time steps shorter than the monthly mean are implicitly adjusted following the monthly scaling. A monthly timescale is adopted due to the generally higher availability of observational data sets at this resolution. Further steps assure consistency between different versions of the data set, e.g. regarding spatial coverage. The different steps in producing the HydroGFD3 data sets are presented in detail in the following sections.

3.1 Climatology

The P background climatology is based on chpclim climatology of satellite, gauge, and physiographic indicators (Funk et al., 2015b). We retain the same climatological period (1980–2009) throughout the HydroGFD3 data set. The chpclim data set comes in two versions: one with full coverage for the 50° S–50° N latitude band and one with global

Table 1. Table of model and data sources used in the production of HydroGFD3 as well as the WFDE5 data set used for comparison. Note the lower-case abbreviations used in the main text and in figures, which follow the internal notation used in the data set production. N_{wet} is a measure of the number of wet days in a month. The data set type is marked in parentheses in the leftmost column; r: model reanalysis; g: gauge-based; s: satellite-based. Today's date is marked by "t" in the "Period" column.

Data set	Name	Variables	Resolution	Period	Data reference
ERA5(r)	e5	T, P	1 h; 0.33°	1979–(t – 3 months)	C3S (2020b)
ERA5T(r)	e5t	T, P	1 h; 0.33°	(t – 3 months) – (t – 5 d)	C3S (2020b)
CRUts4.03(g)	cru	T, P, N_{wet}	1 month; 0.5°	1901–(t – 2 months)	Harris and Jones (2019)
GPCCv8(g)	gpcch	P	1 month; 0.25°	1891–2016	Schneider et al. (2018a)
GPCC-monitoringv6(g)	gpccm	P	1 month; 1.0°	1982–(t – 3 months)	Schneider et al. (2018b)
GPCC-First guess(g)	gpccf	P	1 month; 1.0°	2004–(t – 1 month)	Schneider et al. (2018b)
CPC-Unified(g)	cpcp	P	1 d; 0.5°	1979–(t – 2 d)	CPC (2020)
CPC-Temp(g)	cpct	T_{min}, T_{max}	1 d; 0.5°	1979–(t – 2 d)	CPC (2017)
CHPclimv1.0(g,s)	chpclim	P	climatological; 0.05°	(1980–2009)	Funk et al. (2015b)
WFDE5-CRU(r,g)	wfde5-cru	T, P	1 h; 0.5°	1979–2018	C3S (2020a)
WFDE5-GPCC(r,g)	wfde5-gpcc	P	1 h; 0.5°	1979–2016	C3S (2020a)

land coverage. We choose to make the global-coverage version the main choice but add information from the tropical full-coverage version to increase coverage along coastlines and islands. The original 0.05° resolution is remapped to the 0.25° resolution of the HydroGFD3 data set, ensuring conservation of precipitation totals. Some issues with the chpclim data set were identified through visual inspection, with observational artefacts in mid-northern Siberia and underestimation in Scandinavia. Therefore, these two regions were replaced by gpcch climatological data for the 1980–2009 period (see Supplement for details). To avoid introducing sharp borders, a zone of five grid points was used around each area as a linear transition from one data set to another. Since Greenland P is poorly mapped by both satellite and gauge data, we have chosen to let its climatology be defined by e5 rather than any of the data sets.

For T, we use the cpct climatology (1980–2009) with only a remapping to the 0.25° resolution and in-filling of missing data points using e5. The third climatology consists of the wet-day frequency (1980–2009), which is taken from the N_{wet} of the cru data set of gridded station observations of the number of wet days in a month. Both T and P are remapped to the 0.25° resolution using a bilinear-interpolation method.

In a final step, the three climatologies are harmonized by only retaining the grid points that are available consistently in all data sets and all months. This leads also to the final land mask of the HydroGFD3 data set, for which adjusted data are produced.

The elevation is defined by the e5 surface geopotential divided by the gravity of Earth (9.80665 m/s^2).

3.2 Anomaly method

HydroGFD3 makes use of several different data sets, which need to be stitched together in different configurations depending on the use. Without some kind of homogenization between the data sets, sharp changes in the data are unavoidable when switching from one data set to another. The homogenization used here is performed by only making use of anomalies from the different data sets.

In the earlier version HydroGFD1 (Berg et al., 2018), which is closely based on the WFD (WATCH Forcing Data) method (Weedon et al., 2011), each month of the reanalysis data set is adjusted with the absolute monthly mean of the observational data set. This main principle is retained; however, in a new homogenization step we create new absolute observations by first calculating the monthly anomaly compared to the 1980–2009 climatological period calculated for each data set, then adding this anomaly to the HydroGFD3 climatology. Anomalies are additive for T,

$$T_{\mathrm{anom}}(\mathrm{year,\ month}) = T(\mathrm{year,\ month}) - T_{\mathrm{clim}}(\mathrm{month}), \qquad (1)$$

and multiplicative for P,

$$P_{\mathrm{anom}}(\mathrm{year,\ month}) = P(\mathrm{year,\ month})/P_{\mathrm{clim}}(\mathrm{month}). \qquad (2)$$

The reverse operation is applied after replacing the climatology.

3.3 Wet-day frequency

A common issue with coarse-resolution models, such as e5, is a tendency to produce excessive drizzle that reduces the number of dry days in a month. To alleviate potential excessive drizzle, the number of wet days is adjusted before correcting the P amount. This is performed by first determining the target number of wet days in the month, then setting the days with weakest precipitation intensity to 0 until the target is reached. No adjustments are made for too few wet days. The wet-day frequencies in a month are not well covered by observational monitoring records, and the uncertainties are large when available. We have chosen to estimate the number of wet days based on the method of Stillman and Zeng

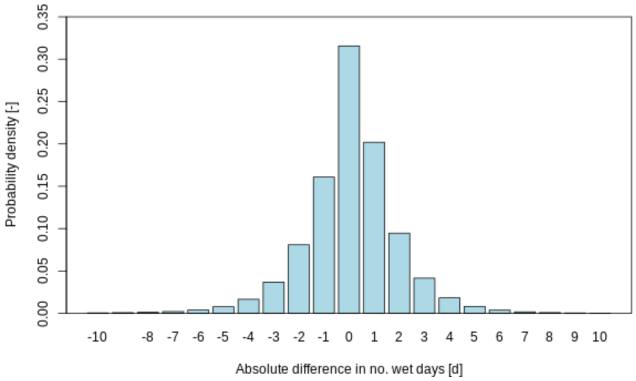

Figure 1. Distribution of the absolute difference in the number of wet days N_{wet} estimated through the Stillman and Zeng (2016) method and gpcch P. The probability density function ranges globally over all land grid points.

(2016). Note that we do not need to define a wet day using a specific threshold value. Instead, the number of wet days in a month is directly defined by the method. The method essentially relates the number of wet days, N_{wet}, to the monthly P anomaly, P_{anom}, using also the climatological wet-day frequency (calculated from cru wet-day data set), $N_{\mathrm{wet}}^{\mathrm{clim}}$ as a predictor, and a tunable constant k.

$$N_{\mathrm{wet}} = P_{\mathrm{anom}}^{k} \cdot N_{\mathrm{wet}}^{\mathrm{clim}} \tag{3}$$

A value of $k = 0.28$ was derived for HydroGFD2.0 by calibration to the cru observations of the number of wet days in a month together with the cpcp P observations. This value is almost half of that found by Stillman and Zeng (2016), which can probably be related to the data sets used but was found to be highly applicable across the world. A verification of this constant was performed with the cru wet days and the gpcch monthly P anomalies; see Fig. 1. This reveals an overall high accuracy of the method, with deviations from observations of mostly only a few days in a month but which can in rare cases be as much as 10 d. On average over the 1980–2009 period and for each single grid point, the deviations are close to 0. Thus, the method works well across all areas and with sufficient precision for our purposes.

3.4 Applied corrections

The production of the corrected data consists of the following steps.

1. Calculate observed anomalies.
2. Construct absolute reference data by adding the anomalies to the HydroGFD3 climatology.
3. Calculate the number of wet days (P only).
4. Remove the weakest excessive wet days in e5 (P only).
5. Calculate the ratio (for P) or difference (for T) between the monthly means of the reference and e5.
6. Apply the ratio or difference to all time steps of e5 within the month.
7. Calculate mean, minimum, and maximum T from the hourly time steps (T only).

For P, the scaling can cause very large values in some cases, e.g. when e5 severely underestimates the number of wet days. Therefore, P is limited to a maximum of 1500 mm/day, which is close to the highest observed record at that timescale.

3.5 Consistency in time and space

To have consistent output in all versions of HydroGFD3, there are internal checks to verify that each of the defined grid points of HydroGFD3 is receiving data after each monthly adjustment. It happens that the land–sea masks of the observational data sets change over time, and they often differ between different data sources. If the anomaly data are not defined for a particular grid point, a search algorithm will identify if there are defined anomalies in grid points within a five-grid-box radius. If the search is successful in finding at least one value, the mean of all values in the search radius is used to fill the grid point value. However, if no defined data are found, the anomaly will be set to 0 for T and 1 for P; in other words, the adjustment will be toward the HydroGFD3 climatology.

3.6 Evaluation

Evaluation of the HydroGFD3 historical data set is presented for the mean climatology of P and T as well as for regional probability distribution functions (PDFs) of daily data and as monthly mean time series. The two latter evaluations are performed for each of the regions defined by Giorgi and Bi (2005) (although we use the correct longitude and latitude coordinates provided by Huebener and Körper, 2013), commonly referred to as Giorgi regions; see Fig. 2. One exception is that we have left out the EQF region in the plots of PDFs since it is contained in other included regions. The reason is that it overlaps other regions, and having only 25 regions simplifies the presentation layout of the plots substantially. For both the PDFs and the time series, only data points in the defined grid points of HydroGFD3 are used. The PDFs are pooling all data in each domain, whereas the time series plots are based on regional averages for each monthly time step.

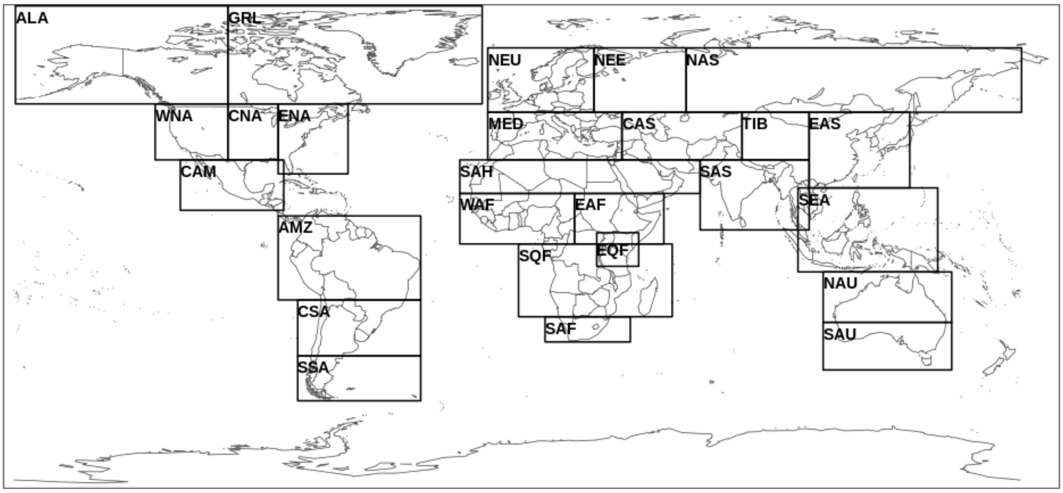

Figure 2. Evaluation regions as defined by Giorgi and Bi (2005) and employed in the PDF and time series analysis.

4 Data sets

HydroGFD3 is built up by different data sets depending on the time period and the tier; see schematic in Fig. 3.

The historical period (1979–2016) is built on e5, corrected with the gpcch and cru data sets, respectively, for P and T. There is only one tier produced for this period; e5 will later be released back to 1950, and the HydroGFD3 historical data will then cover that period as well.

After 2016, in the "extended" and "near-real-time" periods, there are three tiers built on different data sets. Tier 1 is the primary choice and follows the gpccm (for the e5 period) and gpccf (for the e5t period) products for P adjustments, and the cpct product (for the complete period) is used for T. Tier 2 builds instead on the cpcp and cpct products. Note that the Tier 1 and Tier 2 T products are identical and are only repeated here for simplification of the schematic. In practice, there is no Tier 2 for T, and the tiers are anyway not necessarily used consistently for T and P together since the data sets are completely independent. Tier 3 is the final resort if none of the data sets for a variable are available. It is performing only a climatological correction of e5 or e5t by calculating anomalies of the reanalysis and adding this to or multiplying it by the HydroGFD3 climatology. Since it does not make use of any observational data sets, it has received the internal file naming convention "none". For P, also the number of wet days is adjusted according to the description in Sect. 3.3, using the reanalysis anomalies as a predictor.

A closer-to-real-time product is possible, with the daily time step cpcp and cpct products being available with a 2 d latency and e5t available at 5 d latency. The adjustment of the e5t data is then based on the latest available 30 d, synchronized between the data sets, and is therefore called "trailing".

Operational aspects

The HydroGFD3 data sets are updated at regular intervals. The "extended" period is updated each month as new e5 and other data sets become available. Each tier works independently and can therefore become available at different times.

The "near-real-time" period is updated at earliest 5 d into the new month, when e5t is available. By then, the cpcp and cpct products are generally available, but gpccf normally needs a few days more. Tier 3 needs no additional data sets and is available together with e5t but is produced at the calendar month time step like the other products. The priority order is independent for each variable and goes from Tier 1–3.

Finally, the "trailing" updates are performed along with e5t and cpcp and cpct updates and is normally available at a 5 d time lag.

5 Results

5.1 Climatology

The climatological period of HydroGFD3 is set to 1980–2009 and is consistently used in this section. Figure 4 presents the annual mean climatology of HydroGFD3 for both P and T as well as the bias of the e5 reanalysis to this climatology; e5 has in general a wet and cold bias in mountainous regions in most of the world. The Arctic is generally wetter and warmer in e5; note that Greenland P is bias-free per definition since the HydroGFD3 climatology uses e5 there. The tropics are generally drier and colder in e5.

Figures S1–S4 show the seasonal HydroGFD3 climatology and biases of e5. The bias patterns are rather stable across the seasons, although the magnitude changes somewhat. Most striking are the relative changes in western Africa in the December–February period, but this is the dry period

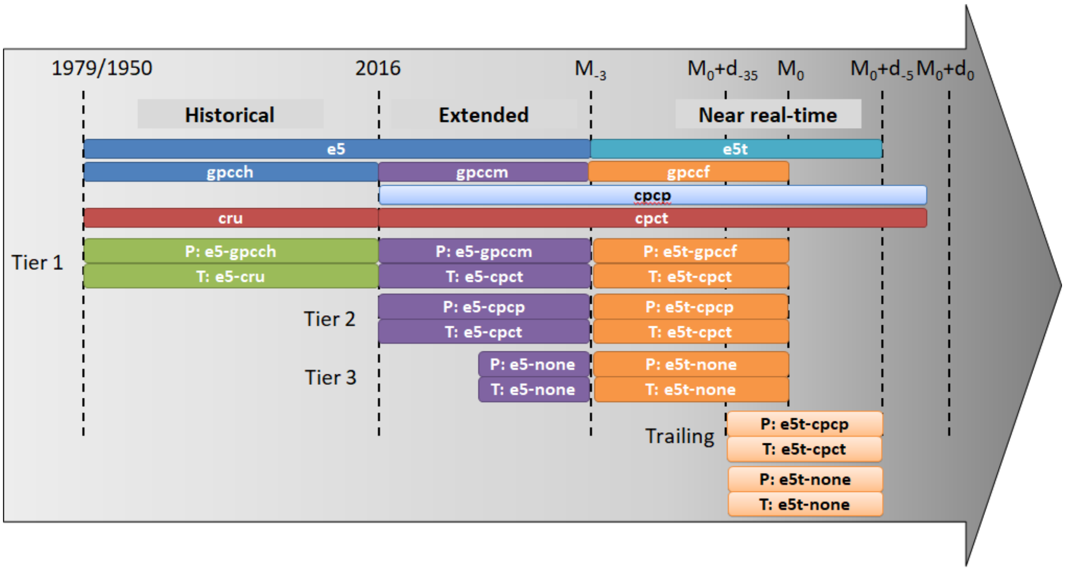

Figure 3. Schematic of the different HydroGFD3 products on a non-linear time axis. The top bars show the original data sources, and the Tier 1–3 and trailing products are shown below. Abbreviations follow Table 1. The time axis denotes years with significant changes in data sources, and the later time marks are relative to the 1st of the current month, M_0, and the current day, d_0. The sub-script for the month is in months and for the days in days.

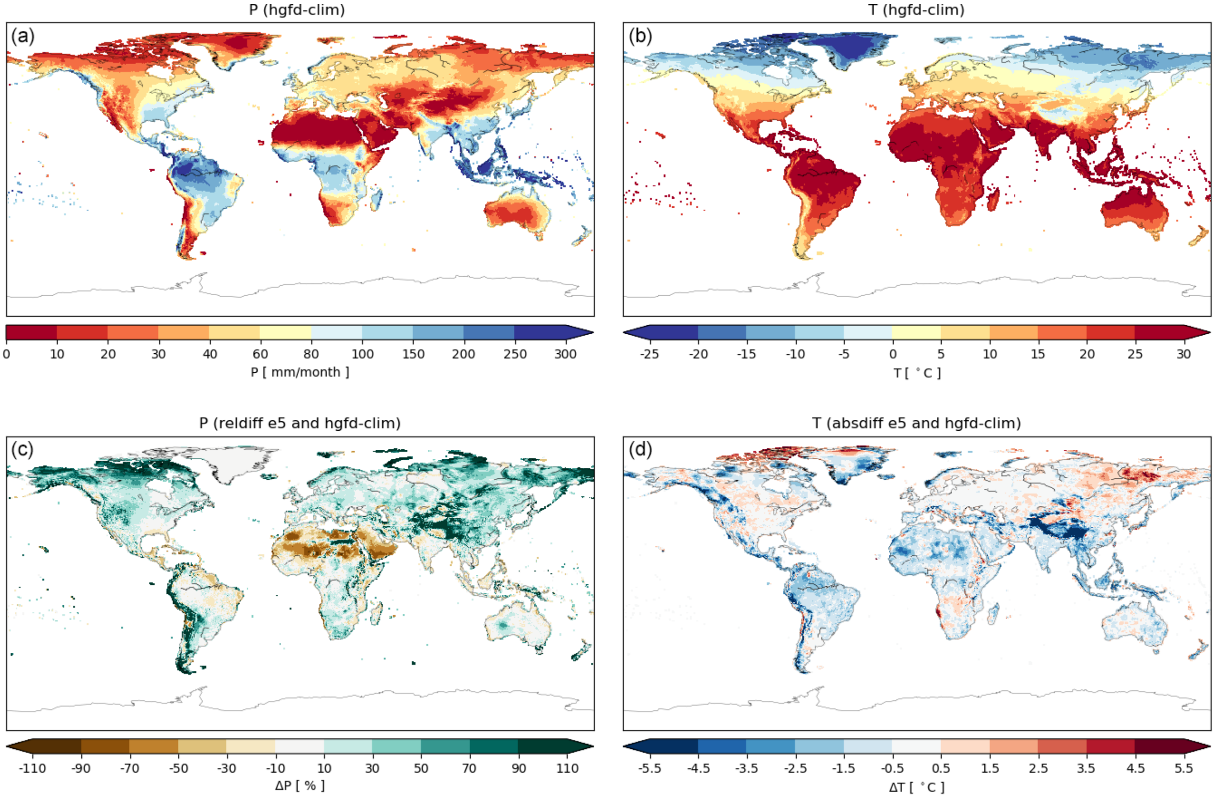

Figure 4. The baseline HydroGFD3 annual mean climatology for P **(a)** and T **(b)**. The bottom row shows the bias of the e5 reanalysis for each variable to the climatology. Note that the lack of P bias for e5 in Greenland is due to the definition of using e5 climatology for that region.

there, and the relative changes are therefore comparing low numbers, which tend to exaggerate the absolute term differences.

We also compare the HydroGFD3 climatologies to other data sets, mainly with a focus on data with daily time steps that could be used equally for the historical period but also to gpcch, which is the main background data set for anomalies in the historical period. Figure 5 shows the annual mean difference in P of gpcch, cpcp, wfde5-gpcc, and wfde5-cru to the HydroGFD3 climatology. Differences to gpcch are generally within ±10 %, except for parts of the Andes mountain range, the Canadian Arctic, the dry north of Africa, the Himalayan plateau, and Greenland. These are all dry and/or snowy regions, with an inherent observational uncertainty, adding the lower gauge network density in the areas. The presented differences between the data sets are considered well inside this expected uncertainty range. We also remark that uncertainties in Greenland are especially large due to few observations and difficult conditions, and data for this region should be used carefully with HydroGFD3 and other data sets alike. The cpcp data set is generally drier, especially in the Arabian Peninsula; wfde5-gpcc and wfde5-cru are both generally wetter than the HydroGFD3 climatology, especially in the cold seasons (see Fig. S5–S8). This is due to the gauge corrections applied in the wfde5 data, which is also the reason for wfde5-gpcc not being identical to gpcch, which it is based on. There are also discrepancies in large dry desert areas such as the Sahara desert, which arise due to differences in the way the number of wet days is calculated in the different data sets. The WFDE5 implementation would produce NaN (not a number) in division by 0 if the number of wet days was 0, which has not happened so far (reviewer comment by Graham Weedon). In HydroGFD3, division by 0 does occur and is solved by setting the ratio to 0 when the calculated number of dry days equals 0. An incompatibility between P and no observed wet days can act to remove P completely for some months, therefore making a drier data set. Seasonal differences (Fig. S5–S8) show similar patterns as the annual mean for most of the regions but can also differ substantially in some regions. One region that stands out is southern Africa in June–August, where both gpcch and cpcp show much wetter conditions (Fig. S7).

For T, we compare to cru only since cpct is used to build the climatology on top of which cru anomalies are added, and wfde5-cru is adjusted to cru and is per definition identical regarding climatology. Figure 6 shows the absolute difference in cru and the HydroGFD3 climatology for each season of the climatological year. The largest differences are in the Arctic region, where gauge availability is low. In other regions, such as south-central Africa, the Himalayan plateau, and other orographic regions, the differences are very consistent over all seasons, with deviations up to a few degrees Celsius. This makes us suspect that they are due to differences in the elevation used for the different data sets. The cpct data set does not come with any information on the elevations used. The use of anomalies from the cru and cpct in constructing the final data set removes such effects, but the climatological difference remains.

5.2 Distributions

Figure 7 shows the PDFs for the complete time period 1980–2009 for P and for each of the data sets e5, hgfd3, cpcp, wfde5-cru, and wfde5-gpcc. In these plots, the spread between the coloured lines representing direct observations or e5 adjusted to observations can be interpreted as indicators of the uncertainty in the observed state. Many regions show fairly high agreement between the data sets, including the original e5 data. In some regions, there is a large spread in the observations, and e5 is somewhere in between, e.g. in ALA, GRL, TIB, and SAH. Again, these regions have large observational uncertainty, making it difficult to determine a ground truth. However, in other regions e5 is deviating significantly in part of the distribution, such as in SSA and WAF moderate intensities and AMZ and EAF extreme intensities.

HydroGFD3 tends to have higher extremes than other data sets. This is partly a resolution effect due to the 0.25° resolution of HydroGFD3 and 0.5° of the other data sets used here. A coarser resolution will move all higher intensities toward the lower intensities (to the left in the PDF plots). That the effect differs between regions is because the extremes are also modulated by the magnitude of the applied correction, i.e. the applied scaling. A scaling factor above 1 will increase the extremes and below 1 will decrease them. The baseline climatology therefore has an impact on the extremes. Also the wet-day calculation of HydroGFD3 can affect the results, and we find that the dry regions, e.g. SAH and MED, have more dry days in HydroGFD3 than in the other data sets. When e5 only gives few P days, while the observational anomaly is high, the scaling factor can become very large, and the only process to limit this is the upper limit of 1500 mm/d, which is seldom reached. The wfde5-gpcc, which has a similar methodology as HydroGFD3, still has lower extremes. Besides the above-mentioned undercatch corrections, the lower extremes may be due to the upper threshold applied to each hour, as can be seen in the original wfde5 code in the CDS (Climate Data Store) catalogue (Copernicus Climate Change Service , 2020a).

For T, the general shapes of the PDFs agree across all data sets and regions (Fig. 8). However, there are sometimes substantial differences between e5 and the observational data sets. Typically, e5 displays issues around 0 °C, which is common in global models and related to melting conditions. There are also seasonal offsets outside the range of the observations. HydroGFD3 remains fairly close to cpct and wfde5-cru in most cases. Orographic effects on T were not accounted for in this comparison, which can explain some of the differences in regions with varying orography such as TIB.

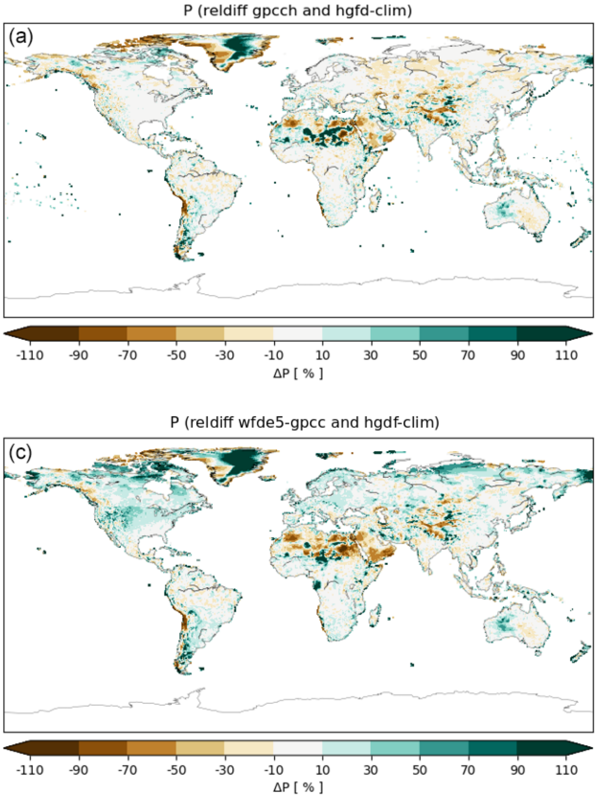
P (reldiff gpcch and hgfd-clim)
(a)
P (reldiff cpc and hgfd-clim)
(b)
P (reldiff wfde5-gpcc and hgdf-clim)
(c)
P (reldiff wfde5-cru and hgdf-clim)
(d)
-110 -90 -70 -50 -30 -10 10 30 50 70 90 110
ΔP [%]

Figure 5. Relative difference in data sets to the HydroGFD3 annual mean P climatology for the period 1980–2009; gpcch **(a)**, cpcp **(b)**, wfde5-gpcc **(c)**, and wfde5-cru **(d)**.

5.3 Temporal trends

To get an impression of the temporal trends and to identify potential issues in the time series, we also investigate the time series as an average over the Giorgi regions. To emphasize differences between the data sets, we discuss mainly differences relative to a common reference, here chosen to be e5. In other words, we present the inverse bias of e5 compared to each observational source.

Figure 9 shows the results for P for the period 1980–2019, and the absolute values are shown in Fig. S9. Note that wfde5-gpcc ends in 2016, wfde5-cru ends in 2018, and gpccm and gpccfg are only available for the last years. The most striking feature is the strong deviations in cpcp for many of the regions. It also varies significantly with time by changing variance, e.g. in SEA, changing mean value, e.g. in CAS, SAS, and AMZ. In some years, there are significant offsets compared to surrounding years, e.g. in 2014 in NEU, NEE, CAS, and MED. Likely, these issues are due to variations in the underlying station network, but we have not verified this. All data sets show signs of an annual cycle in their anomalies to e5 in colder regions, which is indicative of differences between warm- and cold-season precipitation. The wfde5-gpcc and wfde5-cru data sets display stronger anomalies over the annual cycle in the colder regions compared to other data sets. This is likely due to the undercatch corrections, which are larger for snowy conditions. As expected, HydroGFD3 follows the general trends of wfde5-gpcc, and the other data sets have similar trends besides the cpcp deviations just discussed. The gpccm and gpccf have similar mean and variance as gpcch in the overlapping period and show generally consistent behaviour for the later years. However, some larger anomalies occur in, for example, CAN, CAM, SQF, and SAH.

For T, the anomalies to e5 (see Fig. 10 and Fig. S10) retain a clear annual cycle in many regions. Sometimes, the annual cycle is mainly for wfde5-cru (e.g, NEU, TIB, SAS) but often for all data sets. HydroGFD3 and cpct are in general close to each other because of the HydroGFD3 climatology reducing the offset to 0. However, cpct has some clear “break points” in its time series in some regions. For example, in NEU, there is a marked change in the magnitude of the anomalies from about 0 to 0.5 °C to −0.5 to 0.5 °C in about 2006. A similar change about that time is visible also for EAS, GRL, MED, SAS, and NAU. Because the climatologies are calculated for the period 1980–2009, part of these changes are included with the earlier weaker variability. HydroGFD3 is based on cru anomalies pre-2016, but from 2016 on, also its variability is subjected to the changes in cpct.

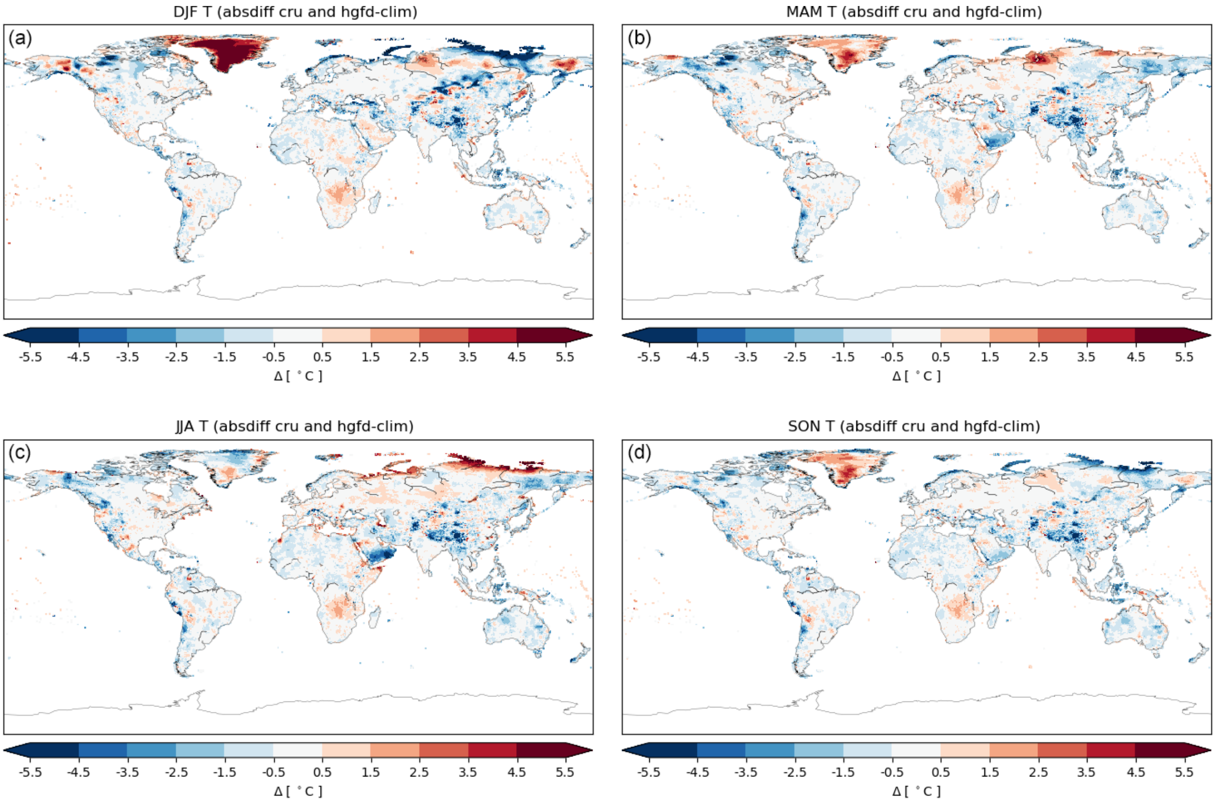
DJF T (absdiff cru and hgfd-clim)
(a)
MAM T (absdiff cru and hgfd-clim)
(b)
JJA T (absdiff cru and hgfd-clim)
(c)
SON T (absdiff cru and hgfd-clim)
(d)
-5.5 -4.5 -3.5 -2.5 -1.5 -0.5 0.5 1.5 2.5 3.5 4.5 5.5
Δ[°C]

Figure 6. Absolute-difference T climatology for the period 1980–2009 between cru and HydroGFD3 for each season: **(a)** December–February, **(b)** March–May, **(c)** June–August, and **(d)** September–November.

Some regions display a significant offset between the data sets, such as SEA, CSA, MED, TIB, and SAS, with cru having generally lower T values. Interestingly, changes in cpct after 2006 often act to reduce the offset to e5.

5.4 Extending to near real time

The near-real-time products, in Fig. 3 called “trailing”, use the daily updates of the cpcp and cpct observations. They are therefore subject to the quality of the cpcp and cpct products and the changes in time as discussed in the previous section. This product follows HydroGFD3 fairly closely to that shown in Figs. 9 and 10 as the main-version Tier 2 is also based on cpcp and cpct but with corrections at calendar months.

In addition, also the “none” products are created with the trailing time window. These only replace the e5 climatology with that of HydroGFD3 and are the simplest form of corrections of the mean. They act as the last failsafe option in the production chain before defaulting to uncorrected e5 data. We do not present this product in the time series plots since it would only constitute a constant annual cycle offset in comparison to e5.

6 Discussion

Compared to similar data sets based on reanalysis, such as WFDE5 and MSWEP, HydroGFD3 differs in that it has its own climatological background and performs the corrections based on anomalies of that same climatological time period. The reason for using this method is to be able to switch data sets closer to real time without “jumps” in the time series. This works well as long as the real-time data set retains its climatological state, which seems to be the case for gpccm and gpccf compared to gpcch. However, cpct and cpcp both cause issues due to changes in the time series towards the end of the time period, in about 2006. The bias of e5 is still reduced, which brings validity to the method. A future development could be to instead retain trends from the ERA5 reanalysis and explore the use of shorter periods for calculating anomalies of the observed data. This would reduce discontinuities in the time series but would remove the potential benefits of using trends from the observations.

HydroGFD3 has generally higher extremes than the other analysed data sets. This is especially so in drier regions where an interplay between the estimation of the number of wet days and the scaling causes fewer wet days and larger scaling factors. In effect, this leads to enlarging the tail of the distribution, e.g. in the MED and SAH region in Fig. 7. It is

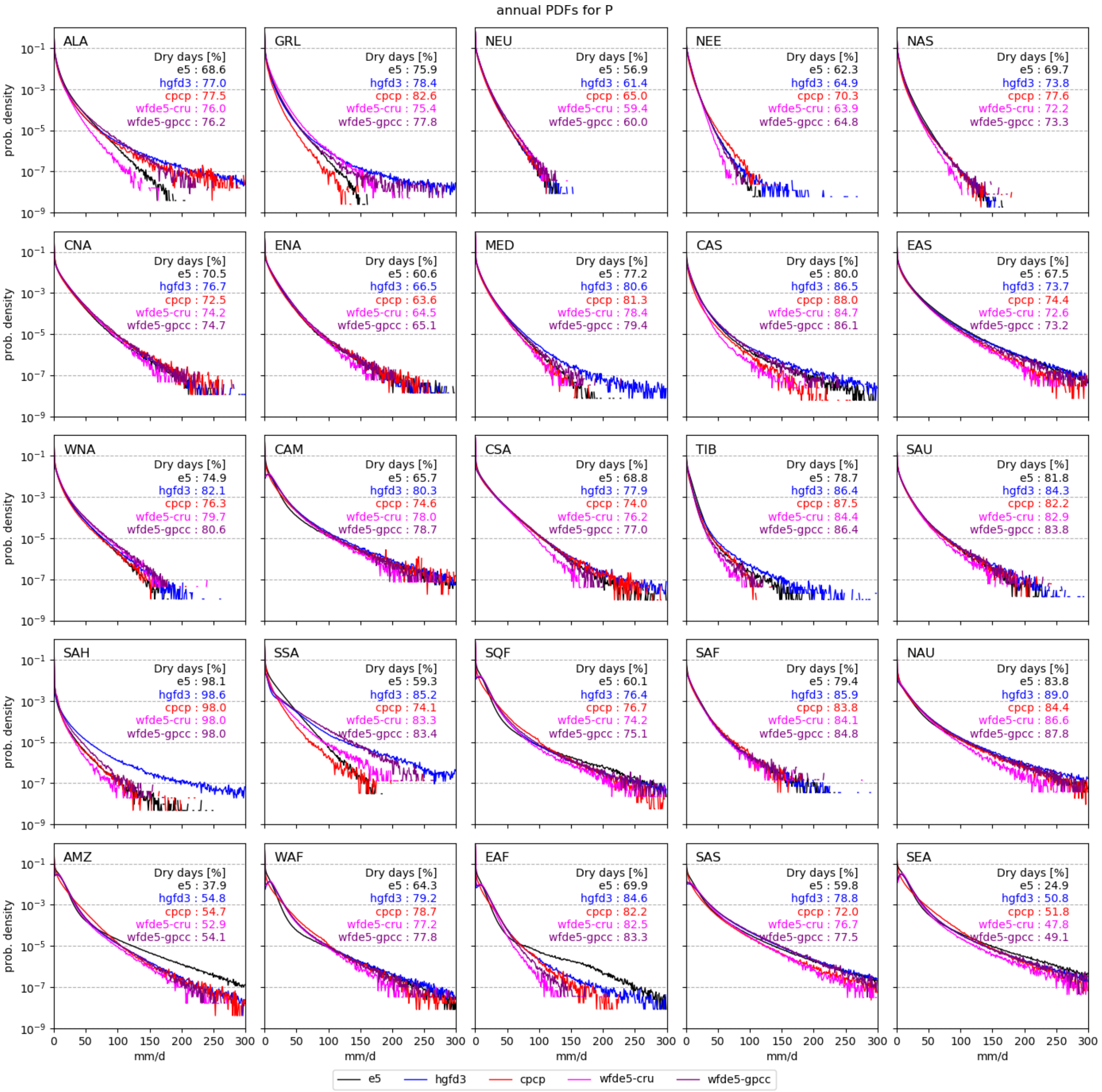

Figure 7. P PDFs of each Giorgi region for the data sets with daily output data in the period 1980–2009. The table in each plot states the percentage of dry days for each data set, i.e. the percentage of data in the first bin of 0–1 mm/d.

possible to restrict the scaling by only allowing the scaling factor to be a few times the original value, but such restrictions would in turn impact the monthly mean. A potential method would be to "borrow" P from adjacent grid points on e5's excessive dry days, thereby reducing the scaling factors. This topic is being investigated for future updates of the methodology.

The regional analysis shows clearly that the observational data sets give substantially different results in some regions. Diverse results are more common in data-sparse regions or in regions where data are not generally available to all data sets. It is therefore difficult to determine which is closer to the truth in a global assessment like this, and more detailed regional studies, such as Fallah et al. (2020), are needed.

The current main usage of the data set is to initialize different HYPE forecasting models around the world, e.g. in Europe (Hundecha et al., 2016), the Niger River (Andersson et al., 2017), and worldwide (Arheimer et al., 2020). This has influenced some of the choices for the set-up, such as the use of only the ERA5 reanalysis model, among other reanalysis

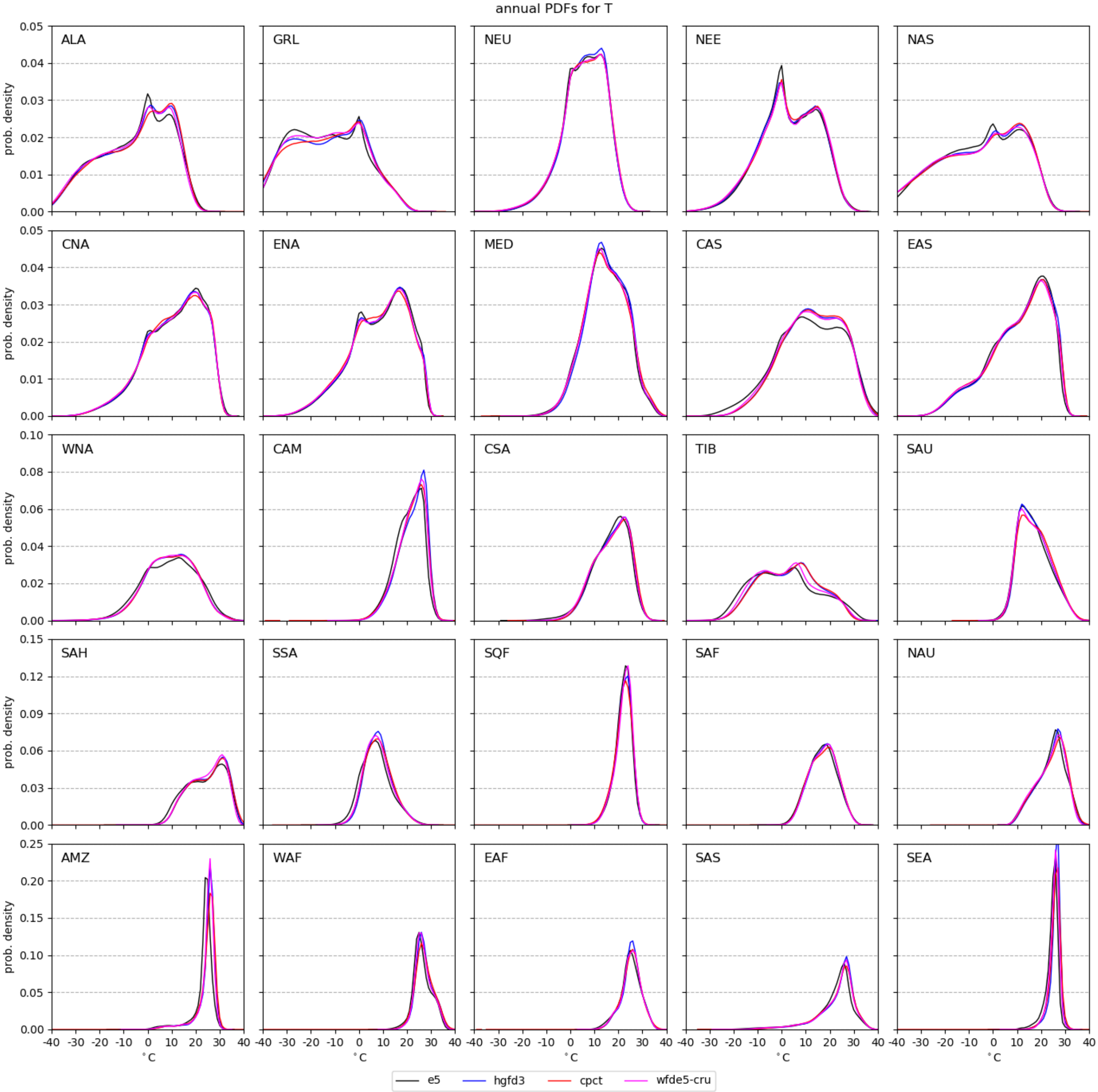

Figure 8. T PDFs of each Giorgi region for the data sets with daily output data in the period 1980–2009.

systems used in e.g. the MSWEP data set (Beck et al., 2017). The forecasts produced by these hydrological models are primarily using the ECMWF deterministic medium-range forecasts or the probabilistic SEAS5 seasonal forecasts, which both use the same model as e5. The priority order of the different redundancy options, i.e. the Tiers 1–3, is based on experience with using the different data sources for our forecasts, with impact from both availability for a given month and experienced longer interruptions.

7 Data availability

For HydroGFD3, a historical period, ranging from February 1979 to December 2019, is available as open source from the Zenodo repository at https://doi.org/10.5281/zenodo.3871707 (Berg et al., 2020). For years prior to 2017, cru and gpccm are used as reference data for T and P, respectively. The following years use instead cpct and gpccm reference data. Real-time updates of the data set are available for a processing charge via subscriptions. Please make a request here at https:

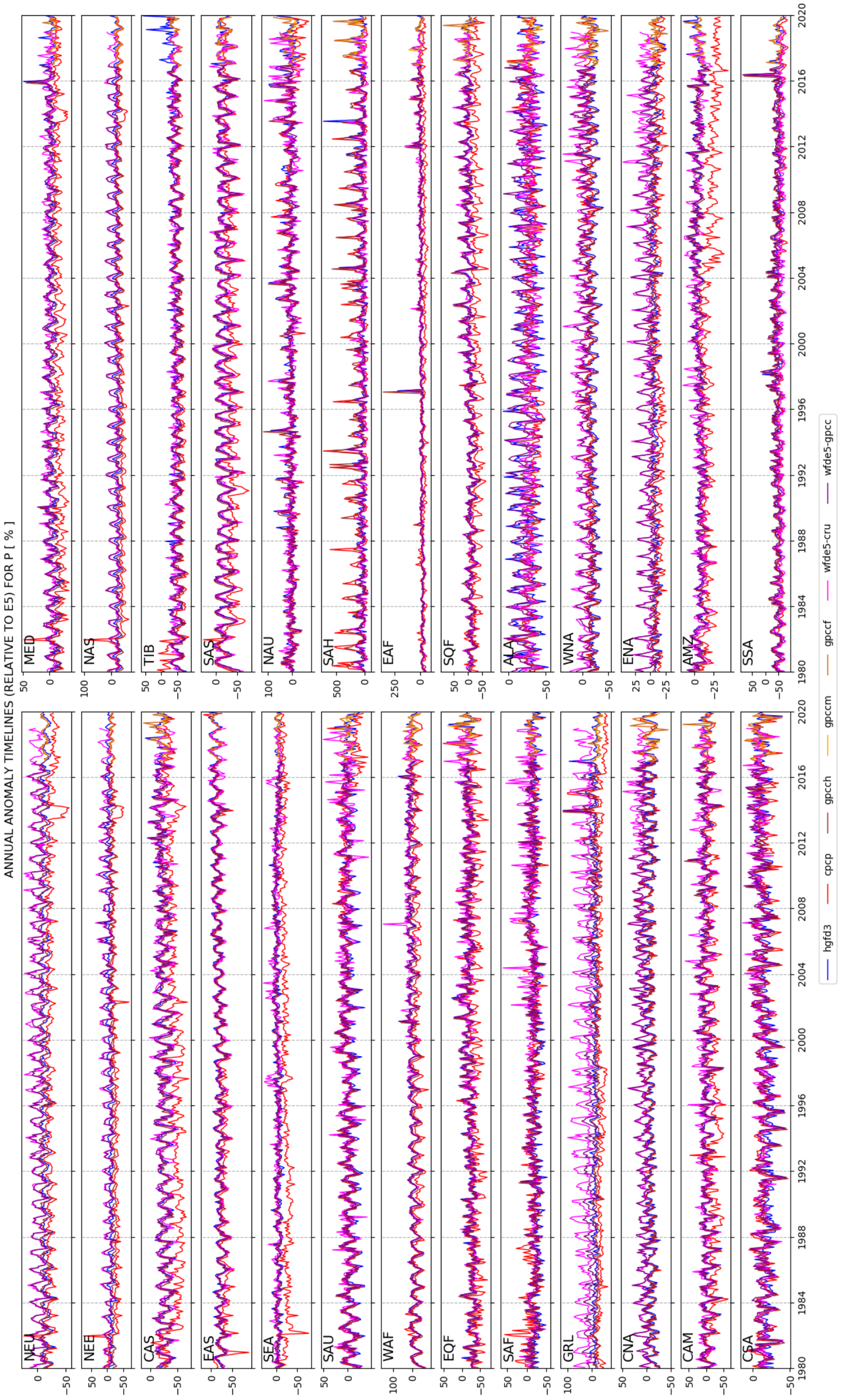

Figure 9. Monthly P anomalies for all data sets, averaged over the Giorgi regions for all valid land data points. The anomalies are relative to the e5 data set and are evaluated for each single month.

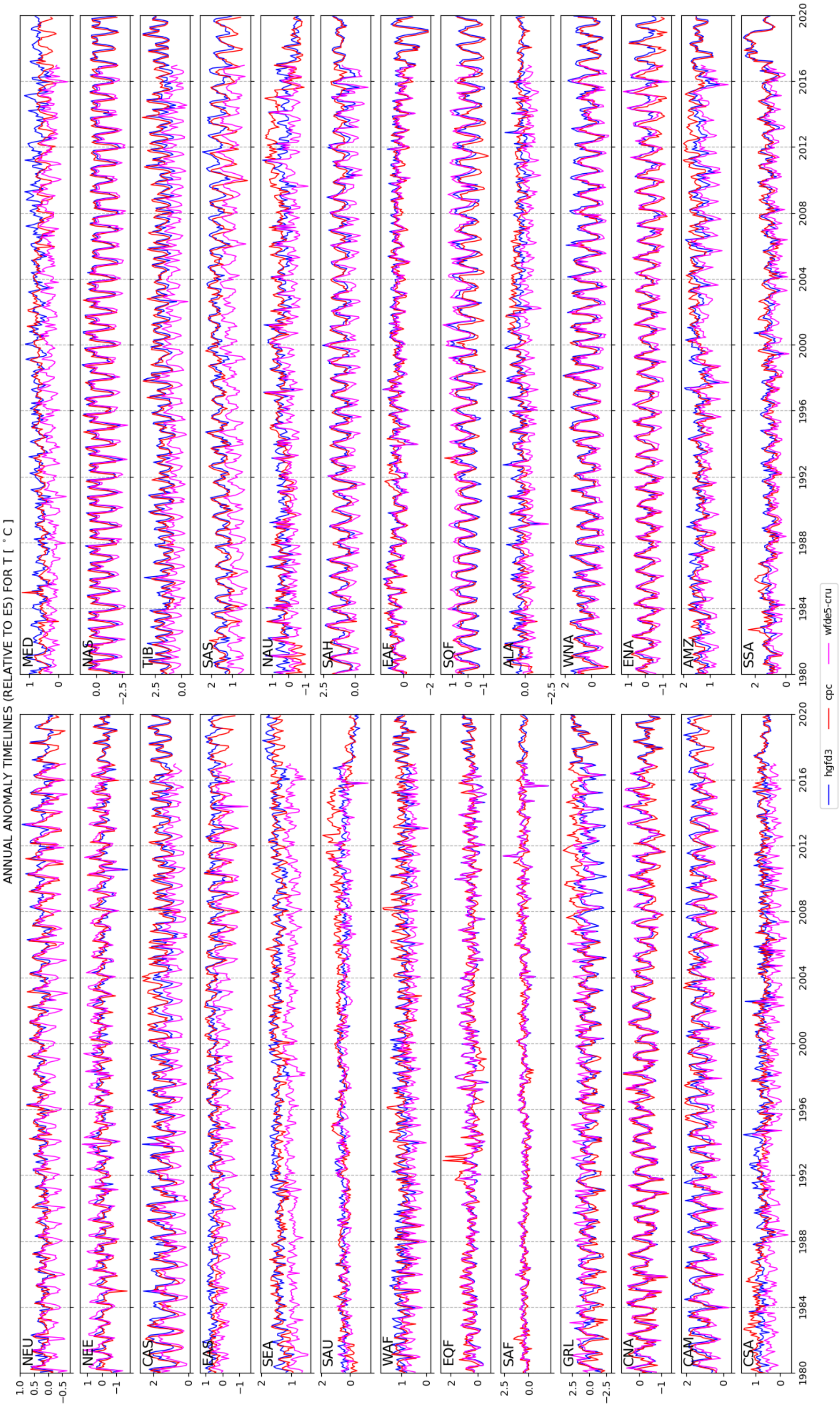

Figure 10. Monthly T anomalies for all data sets, averaged over the Giorgi regions for all valid land data points. The anomalies are relative to the e5 data set and are evaluated for each single month.

//hypeweb.smhi.se/buy-water-services/data-subscription/ (last access: 22 March 2021) and make sure to mention the data set name "HydroGFD3". All data sets listed in Table 1 are available through the provided references.

8 Conclusions

The HydroGFD3 methodology of correcting the e5 reanalysis model toward an observational reference, along with the resulting data sets, was presented. We conclude that the data sets compare well with existing similar data sets.

The main new features of HydroGFD3 are

- higher spatial resolution of 0.25°
- near-real-time corrected data until 5 d from now, i.e. following the continuously updated e5 + e5t time period
- temporal coverage from 1979, which will be extended back to 1950 along with the extended e5 data expected during 2021
- multiple redundancy options to avoid halting production when single data sets are delayed.

Supplement. The supplement related to this article is available online at: https://doi.org/10.5194/essd-13-1531-2021-supplement.

Author contributions. PB developed the methodology and designed the experiments. FA developed the code and infrastructure for the operational system. DB performed analysis and prepared figures for the manuscript. PB prepared the manuscript with contributions from all co-authors.

Competing interests. The authors declare that they have no conflict of interest.

Acknowledgements. The authors would like to thank the EU and the Swedish Research Council Formas for funding within the framework of the GlobalHydroPressure project, financed under the 2017 Joint Call of Water JPI (IC4WATER).

Review statement. This paper was edited by Martin Schultz and reviewed by Graham Weedon and one anonymous referee.

References

Andersson, J. C., Ali, A., Arheimer, B., Gustafsson, D., and Minoungou, B.: Providing peak river flow statistics and forecasting in the Niger River basin, Phys. Chem. Earth, 100, 3–12, https://doi.org/10.1016/j.pce.2017.02.010, 2017.

Arheimer, B., Pimentel, R., Isberg, K., Crochemore, L., Andersson, J. C. M., Hasan, A., and Pineda, L.: Global catchment modelling using World-Wide HYPE (WWH), open data, and stepwise parameter estimation, Hydrol. Earth Syst. Sci., 24, 535–559, https://doi.org/10.5194/hess-24-535-2020, 2020.

Ashouri, H., Hsu, K.-L., Sorooshian, S., Braithwaite, D. K., Knapp, K. R., Cecil, L. D., Nelson, B. R., and Prat, O. P.: PERSIANN-CDR: Daily Precipitation Climate Data Record from Multisatellite Observations for Hydrological and Climate Studies, B. Am. Meteorol. Soc., 96, 69–83, https://doi.org/10.1175/bams-d-13-00068.1, 2015.

Beck, H. E., van Dijk, A. I. J. M., Levizzani, V., Schellekens, J., Miralles, D. G., Martens, B., and de Roo, A.: MSWEP: 3-hourly 0.25° global gridded precipitation (1979–2015) by merging gauge, satellite, and reanalysis data, Hydrol. Earth Syst. Sci., 21, 589–615, https://doi.org/10.5194/hess-21-589-2017, 2017.

Berg, P., Donnelly, C., and Gustafsson, D.: Near-real-time adjusted reanalysis forcing data for hydrology, Hydrol. Earth Syst. Sci., 22, 989–1000, https://doi.org/10.5194/hess-22-989-2018, 2018.

Berg, P., Almén, F., and Bozhinova, D.: HydroGFD3.0, Zenodo, https://doi.org/10.5281/ZENODO.3871707, 2020.

C3S (Copernicus Climate Change Service): Near surface meteorological variables from 1979 to 2018 derived from bias-corrected reanalysis, https://doi.org/10.24381/cds.20d54e34, 2020.

C3S (Copernicus Climate Change Service): ERA5 hourly data on single levels from 1979 to present, https://doi.org/10.24381/cds.adbb2d47, 2020.

Chen, M., Shi, W., Xie, P., Silva, V. B. S., Kousky, V. E., Wayne Higgins, R., and Janowiak, J. E.: Assessing objective techniques for gauge-based analyses of global daily precipitation, J. Geophys. Res.-Atmos., 113, D04110, https://doi.org/10.1029/2007JD009132, 2008.

CHP (Climate Hazards Group): Monthly quasi-global satellite and observation based precipitation climatology, available at: https://data.chc.ucsb.edu/products/CHPclim/netcdf/ (last access: 18 September 2020), 2015.

Climate Hazards Group (CHP): Monthly quasi-global satellite and observation based precipitation climatology, available at: https://data.chc.ucsb.edu/products/CHPclim/netcdf/ (18 September 2020), 2015.

Climate Prediction Center (CPC): CPC Unified gauge-based analysis of global daily precipitation, available at: https://ftp.cpc.ncep.noaa.gov/precip/CPC_UNI_PRCP/GAUGE_GLB/ (18 September 2020), 2021.

Copernicus Climate Change Service (C3S): Near surface meteorological variables from 1979 to 2018 derived from bias-corrected reanalysis, https://doi.org/10.24381/cds.20d54e34, 2020a.

Copernicus Climate Change Service (C3S): ERA5 hourly data on single levels from 1979 to present [dataset], https://doi.org/10.24381/cds.adbb2d47, 2020b.

CPC (Climate Prediction Center): https://www.esrl.noaa.gov/psd/data/gridded/data.cpc.globaltemp.html (last access: 18 September 2020), 2017.

CPC (Climate Prediction Center): CPC Unified gauge-based analysis of global daily precipitation, available at: https://ftp.cpc.ncep.noaa.gov/precip/CPC_UNI_PRCP/GAUGE_GLB/ (last access: 18 September 2020), 2020.

CPCtemp: https://www.esrl.noaa.gov/psd/data/gridded/data.cpc.globaltemp.html (last access: 18 September 2020), 2017.

Cucchi, M., Weedon, G. P., Amici, A., Bellouin, N., Lange, S., Müller Schmied, H., Hersbach, H., and Buontempo, C.: WFDE5: bias-adjusted ERA5 reanalysis data for impact studies, Earth Syst. Sci. Data, 12, 2097–2120, https://doi.org/10.5194/essd-12-2097-2020, 2020.

Fallah, A., Rakhshandehroo, G. R., Berg, P. O. S., and Orth, R.: Evaluation of precipitation datasets against local observations in southwestern Iran, Int. J. Climatol., 40, 4102–4116, https://doi.org/10.1002/joc.6445, 2020.

Fan, Y. and Van den Dool, H.: A global monthly land surface air temperature analysis for 1948–present, J. Geophys. Res.-Atmos., 113, D01103, https://doi.org/10.1029/2007JD008470, 2008.

Funk, C., Peterson, P., Landsfeld, M., Pedreros, D., Verdin, J., Shukla, S., Husak, G., Rowland, J., Harrison, L., Hoell, A., and Michaelsen, J.: The climate hazards infrared precipitation with stations – a new environmental record for monitoring extremes, Sci. Data, 2, 150066, https://doi.org/10.1038/sdata.2015.66, 2015a.

Funk, C., Verdin, A., Michaelsen, J., Peterson, P., Pedreros, D., and Husak, G.: A global satellite-assisted precipitation climatology, Earth Syst. Sci. Data, 7, 275–287, https://doi.org/10.5194/essd-7-275-2015, 2015b.

Giorgi, F. and Bi, X.: Updated regional precipitation and temperature changes for the 21st century from ensembles of recent AOGCM simulations, Geophys. Res. Lett., 32, L21715, https://doi.org/10.1029/2005gl024288, 2005.

Harris, I. C. and Jones, P. D.: CRU TS4.03: Climatic Research Unit (CRU) Time-Series (TS) version 4.03 of high-resolution gridded data of month-by-month variation in climate (Jan. 1901–Dec. 2018), CEDA Archive, https://doi.org/10.5285/10D3E3640F004C578403419AAC167D82, 2019.

Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz-Sabater, J., Nicolas, J., Peubey, C., Radu, R., Schepers, D., Simmons, A., Soci, C., Abdalla, S., Abellan, X., Balsamo, G., Bechtold, P., Biavati, G., Bidlot, J., Bonavita, M., Chiara, G., Dahlgren, P., Dee, D., Diamantakis, M., Dragani, R., Flemming, J., Forbes, R., Fuentes, M., Geer, A., Haimberger, L., Healy, S., Hogan, R. J., Hólm, E., Janisková, M., Keeley, S., Laloyaux, P., Lopez, P., Lupu, C., Radnoti, G., Rosnay, P., Rozum, I., Vamborg, F., Villaume, S., and Thépaut, J.-N.: The ERA5 global reanalysis, Q. J. Roy. Meteor. Soc., 146, 1999–2049, https://doi.org/10.1002/qj.3803, 2020.

Huebener, H. and Körper, J.: Changes in Regional Potential Vegetation in Response to an Ambitious Mitigation Scenario, J. Environ. Prot., 4, 16–26, https://doi.org/10.4236/jep.2013.48a2003, 2013.

Hundecha, Y., Arheimer, B., Donnelly, C., and Pechlivanidis, I.: A regional parameter estimation scheme for a pan-European multi-basin model, J. Hydrol. Regional Studies, 6, 90–111, https://doi.org/10.1016/j.ejrh.2016.04.002, 2016.

Joyce, R. J., Janowiak, J. E., Arkin, P. A., and Xie, P.: CMORPH: A Method that Produces Global Precipitation Estimates from Passive Microwave and Infrared Data at High Spatial and Temporal Resolution, J. Hydrometeorol., 5, 487–503, https://doi.org/10.1175/1525-7541(2004)005<0487:camtpg>2.0.co;2, 2004.

Lindström, G., Pers, C., Rosberg, R., Strömqvist, J., and Arheimer, B.: Development and test of the HYPE (Hydrological Predictions for the Environment) model – A water quality model for different spatial scales, Hydrol. Res., 41, 295–319, https://doi.org/10.2166/nh.2010.007, 2010.

Schneider, U., Becker, A., Finger, P., Meyer-Christoffer, A., and Ziese, M.: GPCC Full Data Monthly Version 2018.0 at 0.25: Monthly Land-Surface Precipitation from Rain-Gauges built on GTS-based and Historic Data, Global Precipitation Climatology Centre, https://doi.org/10.5676/DWD_GPCC/FD_M_V2018_025, 2018a.

Schneider, U., Becker, A., Finger, P., Meyer-Christoffer, A., and Ziese, M.: GPCC Monitoring Product Version 6.0 at 1.0: Near Real-Time Monthly Land-Surface Precipitation from Rain-Gauges based on SYNOP and CLIMAT Data, Global Precipitation Climatology Centre, https://doi.org/10.5676/DWD_GPCC/MP_M_V6_100, 2018b.

Stillman, S. and Zeng, X.: Development of a 0.5° global monthly raining day product from 1901 to 2010, Geophys. Res. Lett., 43, 9704–9711, https://doi.org/10.1002/2016gl070244, 2016.

Weedon, G., Gomes, S., Viterbo, P., Shuttleworth, W., Blyth, E., Österle, H., Adam, C., Bellouin, N., Boucher, O., and Best, M.: Creation of the watch forcing data and its use to assess global and regional reference crop evaporation over land during the twentieth century, J. Hydrometeor., 12, 823–848, https://doi.org/10.1175/2011JHM1369.1, 2011.

Weedon, G. P., Balsamo, G., Bellouin, N., Gomes, S., Best, M. J., and Viterbo, P.: The WFDEI meteorological forcing data set: WATCH Forcing Data methodology applied to ERA-Interim reanalysis data, Water Resour. Res., 50, 7505–7514, 2014.