Spatial partitioning of terrestrial precipitation reveals varying dataset agreement across different environments

Yannis Markonis[1*], Mijael Rodrigo Vargas Godoy[1], Rajani Kumar Pradhan[1], Shailendra Pratap[1], Johanna Ruth Tomson[1], Martin Hanel[1], Athanasios Paschalis[2], Efthymios Nikolopoulos[3], Simon Michael Papalexiou[1, 4]

[1]Faculty of Environmental Sciences, Czech University of Life Sciences Prague, Kamýcká 129, Praha – Suchdol, Czech Republic.
[2]Department of Civil and Environmental Engineering, Imperial College of London, London, United Kingdom.
[3]Department of Civil and Environmental Engineering, Rutgers University, Piscataway, NJ 08854, USA.
[4]Department of Civil Engineering, University of Calgary, Calgary, Canada.

*Corresponding author(s). E-mail(s): markonis@fzp.czu.cz;

Abstract

The study of the water cycle at planetary scale is crucial for our understanding of large-scale climatic processes. However, very little is known about how terrestrial precipitation is distributed across different environments. In this study, we address this gap by employing a 17-dataset ensemble to provide, for the first time, precipitation estimates over a suite of land cover types, biomes, elevation zones, and precipitation intensity classes. We estimate annual terrestrial precipitation at approximately 114 000 ± 9 400 km$^{\mathbf{3}}$, with about 70% falling over tropical, subtropical and temperate regions. Our results highlight substantial inconsistencies, mainly, over the arid and the mountainous areas. To quantify the overall discrepancies, we utilize the concept of dataset agreement and then explore the pairwise relationships among the datasets in terms of "genealogy", concurrency, and distance. The resulting uncertainty-based partitioning demonstrates how precipitation is distributed over a wide range of environments and improves our understanding on how their conditions influence observational fidelity.

Keywords: Water cycle, Terrestrial precipitation, Dataset agreement, Precipitation uncertainty

Introduction

In the last 100 years, more than 40 studies have attempted to quantify the global water cycle budget [1][1]. This is no surprise because, despite the challenges in robustly estimating the amount of water that is exchanged between the atmosphere, lithosphere, and hydrosphere, the role of water is pivotal in many abiotic and biotic processes. The role of water does not only affect the energy cycle through the latent heat release, but it is also closely related to the Earth's biogeochemical cycles, which are crucial factors for ecosystem functioning. Thus, the assessment of the global water cycle budget and its variability is critical for understanding how the Earth system works. Having accurate estimates of its fluxes is a vital first step to achieve it.

Among the water cycle fluxes, precipitation, which includes all the forms of water that is condensed in the atmosphere and then transferred to the ground, is one of the major components and certainly the most measured one. In the last decades, its estimation has come a long way as more accurate instruments became available and rain-gauge networks have been established at global scale, like for example the Global Historical Climatology Network [2]. At the same period, the rise of the internet and open data policies allowed for easy and quick exchange of precipitation records, which resulted in the development of gridded global datasets. The availability of data products became exponential with the beginning of the satellite era, marked by the launch of the Tropical Rainfall Measuring Mission [3], offering coverage over previously inaccessible or unmonitored regions. In a parallel attempt to further improve the spatio-temporal resolution of the measurements, reanalysis data products such as NASA/DAO, NCEP/NCAR, and ERA-15 rose to the avant-garde [4–6]. Once again, reanalyses implied a further increase in the number of available datasets because now we can permute different combinations of models, observations, and assimilation schemes. Nowadays, we are in the propitious position to have increasingly accurate precipitation estimates coming from these three categories; gridded station-based observations, satellite measurements, and reanalysis simulations.

The unprecedented data wealth had a direct effect on the quantification of global water cycle budget and its constituent fluxes. In their milestone study, Trenberth et al. [7] were the first to exploit the observational and model simulation data availability (GPCP v2, CRU TS 2.1, PREC/L, CLM3, ERA-40) to report the global water cycle mean state during the 1979–2000 period. Their multi-source approach became the norm for the studies that followed, and in the last decade the focus of research shifted to the application of consistent data fusion techniques between the various data products [8]. Still, although all the studies of global water cycle budget provide estimates of precipitation, exploring how precipitation is partitioned over land has received quite less attention. Despite the progress that has been made, we still find it hard to answer

[1] The original excerpts from all referenced works (excluding dataset studies) can be found in Section S1 of the Supporting Information.

simple questions about how precipitation is distributed over land, for example *"How much does it precipitate over the boreal forests?"*.

So far, there has been only one study of the global water cycle budget that effectively mapped the distribution of water over various land cover types [9]. Being itself a review of earlier works [10–13], the study of Oki and Kanae reports that out of the 111 thousand km^3 of water that annually falls over land, almost half of it (54 thousand km^3) falls over forests, less than a third (31 thousand km^3) over grassland, 11.6 thousand km^3 over cropland, 2.4 thousand km^3 over lakes, and the remaining 12 thousand km^3 are distributed over other smaller fractions of land cover types. A similar, but rather simpler, approach can be found in the study of global transpiration by Schlesinger and Jasechko [14]. In this meta-analysis of the global transpiration/evapotranspiration ratio, the precipitation estimates were calculated by simply multiplying the total biome area to the average precipitation that is known to correspond to each biome [14]. This kind of partitioning is missing from modern water cycle budget studies, which at most report how precipitation is separated over ocean and land [1].

In this work, we use a large ensemble of global precipitation datasets to revisit the prior estimates and extend them to elevation zones and precipitation intensity classes. To quantify the uncertainty in the estimation of the spatial partitioning for each category we introduce the approach of dataset agreement, assuming that there is no observable "ground truth". In this manner, we determine the regions and categories with high observational fidelity among the 17 datasets, and discuss their impact on the overall partitioning. The pattern of differences between the gridded station observations, the satellite measurements, and the reanalysis simulations can be easily observed, helping us pave the way to future improvements and better estimates of terrestrial precipitation. Still, despite their differences, the state-of-the-art precipitation data products are able to provide a clear overview of the distribution of precipitation over land in the first two decades of the 21st century.

Results

The ensemble mean of the annual terrestrial precipitation is estimated at 111 650 $\pm$ 9 445 km^3 (Tables 1 & S2). In this estimate the precipitation over Antarctica is not included due to poor station coverage. If we add to the global annual volume the Antarctica precipitation estimates reported by Rodell et al. [15] and Bromwich et al. [16], then the annual terrestrial precipitation reaches 114 thousand km^3 (see Methods). As expected, almost half of terrestrial precipitation falls over the tropical climates, with temperate regions coming second ($\approx$21%). Together, these two regions account for slightly more than two thirds of the terrestrial precipitation while covering only one third of global land. On the contrary, the arid regions that have a similar areal extent, receive only 10% of the precipitation. The polar regions, which in this study include only the arctic and high mountainous domains, receive a very small fraction of the total precipitation.

The largest portion of terrestrial precipitation falls over forested regions, and most forest precipitation is concentrated over tropical forests specifically (Fig. 1a, Table S4). Depending on the subset criterion, the total precipitation volume ranges between 47.39

(land cover) and 66.25 (biome) thousand km^3 per year. Land cover refers to the physical characteristics of the Earth's surface, such as forests, wetlands, and water bodies, while the biome refers to a large geographic area with similar climate, vegetation, and animal life. Therefore, the reason for the above discrepancy is that savannas are regarded as a different land cover than forests, while they are considered part of the forest biome (Fig. 1b, Table S5). In total, forests, savannas, and croplands receive 73% of the terrestrial precipitation, with the remaining 27% consisting of shrublands (mainly desert and tundra), grasslands, barren, and water/snow/ice-covered regions. A similar fraction (75%) of the terrestrial precipitation falls over the 0 – 800 m elevation zone, with only 7.8% falling over 1 500 meters (Fig. 1c, Table S6). The shape of the elevation distribution depends on the elevation zone selection and the different climatic classes are well-distributed among them. Overall, 30% of the global land area receives the 70% of terrestrial precipitation, laying within the three highest precipitation intensity classes (Fig. 1d, Table S7).

In general there is good agreement between the various data sources over the regions of high precipitation and low in the more arid ones (Fig. 2a). The Sahara and Arabian deserts, the Tibetan plateau, the Andes and the Rocky Mountains, as well as the high latitude areas, show large disagreement between the datasets. Water-scarce ecosystems, such as deserts, tundras, and montane grasslands, portray the largest discrepancies among the datasets (Fig. S5). These ecosystems are dominated by shrublands or non-vegetated land cover types such as permanent snow and barren regions. Additionally, the higher elevation zones have lower observational fidelity with regions above three thousand meters demonstrating low and below average dataset agreement close to 75% of the grid cells (Fig. S5c). However, due to the low amounts of precipitation that these regions receive, the uncertainty stemming from the dataset disagreement doesn't affect the global total much. We estimate that the grid cells with low and below average dataset agreement cover only about 13% of the total precipitation (circa 14.5 thousand km^3 per year with a standard deviation around 2.5 thousand km^3). This has a rather small impact to the spatial partitioning, which doesn't change significantly if the grid cells with below average dataset agreement are omitted from its estimation (Figs. S6 – S9).

Conversely, regions with high precipitation show stronger consistency among the datasets, which is partially caused by the estimation of the standardized inter-quantile range used to determine the dataset agreement. This is because the absolute differences in many low precipitation regions remain relatively high when compared to their means. Thus, if we use the absolute inter-quantile range then the high precipitation regions will have lower agreement (Fig. S10). To remedy this effect, we also estimated dataset agreement per precipitation intensity class (Fig. 2b). This representation provides some extra information about the uncertainty across regions with similar climatic properties. For example, the western half of the Sahara desert has lower spread among the datasets than its eastern counterpart. Also the tropics and other regions of higher dataset agreement appear less homogeneous with emerging hotspots of uncertainty. The most likely cause for the heterogeneity is the (non-) existence of operational ground stations (Fig. S11).

Looking at each data source category, i.e., gauge-based, remote sensing, and reanalyses, there are distinct differences per climate class (Table 1) and land cover type. The mean of reanalyses show consistently higher values compared to the station data across all climate classes, ranging from 4% for tropical to a tenfold 42% for polar climate, and resulting to 11% globally. On the contrary, the estimates of remote sensing data appear closer to the ground stations, even in regions with scarce gauge coverage such as the polar or the tropical ones. The highest divergence between them is encountered over the continental climate. These differences occur irrespective of the land type classification used examined in this study (Fig. 3, and Figs. S12 – S14). In addition, the probability distribution of grid average precipitation per land use is significantly different in terms of overall shape. For example, in forests and grasslands, station datasets appear to cover half of the total data spread and mainly overlap with remote sensing data. On the contrary, the remote sensing datasets overlap with reanalysis datasets over croplands, where the station datasets show an even narrower spread. The highest similarity appears over barren land, where all three data products share a common empirical distribution. In general, despite their differences, we see that on average the ground stations provide the lowest estimates, the reanalyses the highest, while the remote sensing data products are in between them.

By further examining the overall uncertainty across individual datasets, we observe that their variance is more than four times higher than the average inter-annual variability of the dataset ensemble. The range of the global twenty-year means spans from 92.6 (CPC) to 126.6 (NCEP-DOE) thousand km^3 per year (Table S3), with a standard deviation of about 11 thousand km^3 per year. The mean of the ensemble standard deviation of the annual global precipitation values is slightly less, but still quite higher than the mean of the inter-annual standard deviation, which is approximately 2.2 thousand km^3. The dataset with the lowest inter-annual variability is CRU-TS, whereas on the other extreme lies NCEP-DOE with a value almost 3.5 times higher (Fig. 4). CPC appears to report the lowest amount of precipitation in all climate classes. Other remarkable negative deviations from the dataset mean manifest in MERRA2 for tropical, in CMAP for temperate and continental, MSWEP for arid, and GPCC for polar climate. On the contrary, the highest estimates of precipitation can be found in NCEP-NCAR for tropical, in ERA5 for temperate, in NCEP-DOE and JRA55 for dry and continental, and in EM-Earth for polar climate. The datasets closest to the ensemble mean are CRU-TS and GPCP, followed by EM-Earth and MSWEP. Based on these findings CRU and GPCP, can be regarded as the most representative choices for large-scale climatologic studies of the terrestrial precipitation, when a multi-source approach is not available.

Discussion

Spatial partitioning of terrestrial precipitation

Understanding how precipitation is distributed over different land types and their corresponding climatic properties is crucial for progressing the study of the global water cycle. Our results can be used either as a reference for attributing past and future changes in the global water cycle functioning or to evaluate its representation in climatic models. We also expect future research to apply similar partitioning in the other water cycle components, such as evaporation and runoff. When these variables will have also been analyzed, we will have a more consistent picture of the moisture exchange between the land and the atmosphere, as well as its storage across land. Terrestrial precipitation is a good place to start, due to the increasing data availability which has also been exploited in this study.

Following the same principle, all the global water cycle studies use terrestrial precipitation as the most reliable component for estimating the global mass budget. Our results of 114 thousand km^3 per year show a good match with the pioneering studies of Oki and Kanae [9] and Trenberth et al. [7], where the total terrestrial precipitation was reported at 111 and 113 thousand km^3 per year, respectively. In addition, looking into the global estimates of terrestrial precipitation in more recent studies, our global estimate appears to be very close to their median. In their chronological literature review on global water budget studies, Vargas et al. [1] show that the 11 studies which have been published since 2009 have a median of terrestrial precipitation at 113 thousand km^3 per year (range 110 to 126 thousand km^3). All these results advocate that in the last two decades we have increased our confidence about the estimate of total terrestrial precipitation by significantly constraining its uncertainty.

If we look at the spatial partitioning by Oki and Kanae [9], we observe small deviations in the three land cover types presented there. Forests appear to receive 54 thousand km^3 per year versus 47 thousand in our study, grassland 31 versus 28 thousand km^3 per year, and cropland 11 versus 18 thousand km^3 per year. These differences could be attributed to the satellite advancements in land type characterization, but also to the land cover changes that occurred in the last 15 years. Nevertheless, the adjacency of the results is encouraging and supports the distribution among the other land cover types. When compared with the results of Schlesinger and Jasechko [14], we also see some agreement in the relative partitioning over biomes. The two dominant biomes, i.e., tropical rainforests and grasslands, appear to receive a larger fraction of precipitation in our study, i.e., 42% vs. 35% and 18% vs. 14%, respectively. On the contrary, there is up to 1% difference on temperate forests (14% of total precipitation in our analysis), boreal forests (8%), temperate grasslands (5%), deserts (4%), steppes (2%), Mediterranean biomes (1%). The most likely reason for the discrepancy could be found in the fact that Schlesinger and Jasechko [14] omit the estimation for subtropical forests and grasslands, which if taken into account would result to comparable values to our findings. An interesting implication of this match is the potential to use the biomes with high dataset agreement as predictors in the extrapolation schemes for generating gridded datasets.

The merits of the dataset agreement approach

All the precipitation estimates are dependent to each other. There is a large degree of overlap in the source data, i.e., the gauge station networks, that go into the different observational data products, as well as the use of some datasets by some other (Fig. 5a). Thus, it is no surprise that the majority of the cross-correlation coefficients of global annual precipitation lies above 0.8 for the annual precipitation time series (Fig. 5b). This is a result of the different methodological approaches applied to the same raw data records. Either it is the calibration process of the satellite sensors, the assimilation schemes of the reanalyses, or the extrapolation method of the gridded station products, in principle each method uses a transfer function to predict the areal precipitation sum for each grid cell. If datasets use similar methods and/or sources which result in high cross-correlation, the mean estimates will be inevitably affected because in our study all observations are considered equally plausible estimates. This would imply that there is some sort of "observational democracy", which dampens any strongly opposing "opinion" or outlier.

A similar issue has risen in the case of climate model simulations. It soon became apparent that the "model democracy" assumption can result to significant biases in the estimates of the ensemble statistics [17]. In the same study, it is also argued that taking the "model democracy" approach of the large model ensembles, could be a more robust method compared to weighting or sub-sampling approaches without out-of-sample testing. In the case of gridded observations, an objective out of-sample testing or any other form of evaluation is not possible as there is no ground truth. There are very few regions with high-resolution (< 10 km) gauge networks, for different climatologies, elevations, etc. to make them suitable for global scale evaluation. Therefore, despite the on-going research in the data fusion techniques or the climate model ensemble validation, there is no straightforward way to tackle this challenge, because the true value of each grid cell remains unknown [18].

Is there a way to distinguish whether high correlation (Fig. 5b) and similar mean values (Fig. 5c) are due to structural similarities between the datasets (same sources/methods) and not a confirmation of lower uncertainty? By simply using the cross-correlation or mean distance metrics, it is hard to say. However, if we look in the "genealogic" information among the datasets (Fig. 5a), we can disentangle if what we see is a robust or a biased estimate (Fig. 5d). If two datasets have a direct structural relationship and share high correlation and low mean distance, they can be regarded as alternative versions of the same dataset. This is, for example, the case of GPCC and MSWEP. On the contrary, in most cases data products from the same family do not agree in terms of cross-correlation and mean distance, e.g., ERA5-Land and EM-Earth. Here, we can assume that the datasets offer extra insight to the dataset ensemble with far less structural overlap.

By applying this methodology, "observational democracy" can provide reasonable results by keeping the datasets that appear to significantly diverge from the ensemble mean. Hence, we propose to first present the whole range of data source variability, and then address the observational fidelity in terms of quantifying the dataset agreement. In this manner, we enhance the explanatory capability of the results at a cost of predictability strength due to increased uncertainty. Inevitably, this approach is prone to the threshold selection that determines which datasets are considered similar and which not. Despite that, it can be very insightful in determining the influence of these relationships to our global estimates as we will see below.

The impact of dataset disagreement in the global precipitation fluxes

Even if we cannot be absolutely confident about the dataset dependencies and overlap, the dataset agreement framework can function as an indicator of the most plausible bias sources. In our case, it is easy to see that MSWEP is very similar to GPCC, and GPCP to GPM-IMERG (Fig. 5d and Table S3). In addition, all four of them are linked with numerous other datasets (Fig. 5a), implying that their estimates could be repeatedly diffused to the other data products. To explore the impact of the potential overlapping, we repeated our global estimations, excluding these four datasets in multiple combinations. In all cases, the differences were not higher than 1% for the mean global precipitation volume and 3% for climatic means. This is because their estimates are so close to the ensemble mean that it makes the estimation of the mean insensitive to their removal. Correspondingly, we can investigate the consequences of removing some obvious outliers, i.e., CPC and the NCEP family (NCAR, DOE, and CMAP; Fig. 4 and Fig. 5b, c). Again, the results remain below 1%, most likely due to the high number of datasets and the symmetry of the outliers, as two of them underestimate and two overestimate the global mean. Therefore, by keeping all the datasets, we preserve the maximum information, with no severe consequences to the estimation of global or climatic means.

The other side of the coin is the uncertainty due to dataset disagreement. Since it is strongly dependent to precipitation intensity, reaching its top over arid and mountainous regions, its impact in our results is quite low (Fig. 3, and Figs. S6 – S9). However, looking more into the regions with high dataset disagreement should be one of the cornerstones of future research. Even though the grid cells with the low dataset agreement receive a small fraction of the global precipitation total, they can be found in regions of high environmental and socioeconomic significance. We see that the strongest inconsistencies lie over arid zones covering approximately 41% of the Earth's land surface with a population above two billion, mainly engaged in agricultural and pastoral activities that are sensitive to water availability [19]. Similarly, mountains or high elevation zones that also show high discrepancies, play an important role in the formation of glaciers, snowfields, and aquifers that store water over extended periods. An exception to this is barren land, where there is enhanced agreement between reanalyses and the other data sources. This could mean that the reanalyses land surface schemes are not ideal and overestimate transpiration and water flux to the atmosphere and thus higher local recycling of rainfall. Finally, future changes in precipitation patterns and amounts may have critical impacts on water availability and ecological functioning over arid or mountainous areas. Thus, improving our estimation of the water cycle components, particularly in regions with low observational fidelity, is crucial for better managing water resources and mitigating the impacts of extreme climatic fluctuations.

The best way to increase observational fidelity is by extending the in-situ monitoring networks. A simplified example for the importance of ground stations to dataset fidelity can be demonstrated if we consider the stations from GHCN network (Figure S11). Although, each data product uses a slightly different station network for interpolation, validation or assimilation, examining the relationship between GHCN stations locations and grid cell dataset agreement is quite informative. Approximately 60%

grid cells with at least one station of the GHCN network have above-average and high dataset agreement. Unfortunately, this covers only 5% of the grid. In the rest 95% of the grid cells with no stations, only 30% show above-average or high dataset agreement. If this is the case for annual values at 0.25° resolution, then we should expect even stronger disagreement at higher spatio-temporal resolutions. Increasing the number of precipitation stations world-wide is the only tangible approach to remedy this issue and improve observational fidelity.

Conclusions

In this study, a detailed estimation of the spatial partitioning of precipitation over land is presented for the first time. The partitioning is supported by, a conceptual framework based on dataset agreement to determine the impact of the uncertainty in the precipitation fluxes. We see that despite the progress in precipitation measurement the global estimate of total terrestrial precipitation remains very close to the values reported at earlier studies [1]. Hence, we can be quite confident that the mean global terrestrial precipitation lies close to 114 000 ± 9 400 km^3. However, the rise in the number of precipitation datasets also revealed the uncertainties at regional scale. The reason that the local precipitation variability does not affect the global mean much, is that it largely appears over arid regions. As a rule of thumb the lower the precipitation, the higher the uncertainty.

By utilizing the concept of dataset agreement, we mapped the global uncertainty not by comparing the precipitation datasets to the "ground truth", but to their ensemble spread. In this manner, we assume that dataset agreement can be regarded as the quantification of the current research status quo in the estimation the total precipitation over land. If the majority of the research is close to the true value of precipitation then our results will be unintentionally skill-weighted by the inclusion of multiple versions of datasets which are closer to the reality. In addition, looking deeper into the reasons of dataset disagreement over regions with different geographical features can result in improvements for the next generation of data products. Correspondingly, areas of strong dataset agreement can be used for evaluating the performance of climate model simulations, and benchmark precipitation shifts as seen in the climate projections that can be of paramount importance for climate resilience studies.

Future research could further explore these directions and as well determine the partitioning and dataset agreement of the other components of terrestrial water cycle. In addition, even though the suggested methodological framework is applied here at global scale, it can be easily downscaled up to regional or catchment scale in order to map the local atmospheric moisture recycling. Finally, a plausible followup will be to investigate the partitioning of the current terrestrial precipitation dynamics and its change across the globe over the last decades. All these future steps can offer new insights in the study of global water cycle and the quantification of its budget.

Going back to our initial question about how much water precipitates over the boreal forests, our results show that it is still difficult to give an accurate estimate. Nevertheless, our study offers an entry point to the answer with an estimate of the annual mean between 8 219 – 10 650 km^3 or 535 – 693 mm. Station observations would report an annual average at 8 219 km^3, satellite estimates would be around 8 760 km^3, while reanalyses would show a quite higher value (10 650 km^3). This example highlights that a lot remains to be done to narrow down the uncertainty of the estimates between the data products at regional scale, but we hope that this study can provide a solid starting point to resolve the challenges that lay ahead.

Methods

Data

To quantify the global terrestrial precipitation we have used a homogenized inventory of 17 precipitation datasets that cover the period 1/2000 – 12/2019. These include:

- Five gauge-based products: CPC-Global [20], CRU TS v4.06 [21], EM-EARTH [22], GPCC v2020 [23], and PREC/L [24]
- Seven satellite-based products: CHIRPS v2.0 [25], CMAP [26], CMORPH [27], GPCP v2.3 [28], GPM IMERGM v06 [29], MSWEP v2.8 [30], and PERSIANN-CDR [31].
- Five reanalysis products: ERA5 [32], JRA55 [33], MERRA2 [34], NCEP/NCAR R1 [5], and NCEP/DOE R2 [35].

A detailed description of the datasets used can be found in Supporting Information (Text S2 and Table S1).

The analysis was performed at annual time step and 0.25° resolution. To achieve this, data homogenization was performed over four stages that address the variable type, measuring units, time step/period, and spatial resolution, respectively. First, data products containing precipitation rates were transformed into total precipitation, and the measuring units were converted initially to mm and then to km^3/grid cell to address the differences in grid cell area. The datasets with daily time steps were aggregated to annual and subset for the selected period which maximizes the number of datasets (1/2000 – 12/2019). In the last step, spatial remapping was performed using Climate Data Operators (CDO) [36]. Datasets with resolutions coarser than 0.25° were regridded by repeating the values over the finer resolutions (i.e., nearest neighbor remapping), while datasets with resolutions finer than 0.25° were upscaled through area-weighted averages and remapped (using the same procedure as for the coarser datasets) in the case when 0.25° was not divisible by the original resolution of a given dataset. The annual mass budget of the regridded datasets were approximately 0.01% lower than the original data. Additionally, we filtered out all the grid cells covered by less than 10 datasets to remove the dissimilarities found in the coastal boundaries of the datasets. Antarctica was not included in the analysis, due to extremely low station coverage. Instead, the estimate of 2.3 thousand km^3 by Rodell et al. [15] and Bromwich et al. [16] was added only to the global volume to have a complete estimate of the terrestrial precipitation. Three out of 17 datasets do not have global coverage (CHIRPS, CMORPH, PERSIANN), and hence were not used for the estimation of the global precipitation sum. The annual records were then uploaded to zenodo repository (https://zenodo.org/records/7078097) and are freely available for download through the *pRecipe* package [37].

Partition categories

The terrestrial precipitation means were estimated globally, as well as per the Köppen-Geiger climate classes, land cover types, biome categories, elevation zones, and precipitation intensity classes. For the climate partitioning, we use the main five Köppen-Geiger classes (A: Tropical, B: Dry, C: Temperate, D: Continental, E: Polar) of the recent classification of Beck et al. [38]. The 14 land cover types of the "MODIS MOD12C1 0.25 Degree Land Cover" data product [39] were aggregated to nine by merging together the different forest types (e.g., broadleaf and conifer; (Fig. S1)). We have also aggregated the 14 biome categories as identified by Dinerstein et al. [40] to 10 by merging together open and closed shrublands, permanent ice and snow, water and wetlands, and by removing the urban and unclassified categories as they covered a negligible fraction of the total area (Fig. S2). The elevation zones were determined using the topography of ERA5 reanalysis [41] (Fig. S3). Finally, we partitioned the grid cells into 10 precipitation intensity classes, based on the deciles of the distribution of the total annual precipitation over all grid cells (Fig. S4).

Dataset agreement

It is well-known that each data product comes with its strengths and weaknesses. At grid scale all of them depend on either an extrapolation scheme (observational datasets), either to a physical model combined to an assimilation framework (reanalysis simulations), or to some transfer function and a calibration approach (satellite data products). Hence, none of them can be considered as "ground truth".

As an alternative approach we propose the concept of "dataset agreement". To quantify the consensus between the available datasets we calculated the standardized interquartile range of the dataset 20-year precipitation means at each grid cell $D = \frac{Q^P_{0.75} - Q^P_{0.25}}{\overline{m}}$, where $(Q^P_{0.25})$ and $(Q^P_{0.75})$, are respectively the first and third quartile, and $\overline{m}$ the mean value of all datasets.

We then classified the standardized interquartile range to five subsets of agreement ranging from "High" to "Low", according to its own quantiles (Q^D) over all grid cells, i.e., "High" $D < Q^D_{0.1}$; "Above average" $Q^D_{0.1} \leq D < Q^D_{0.3}$; "Average" $Q^D_{0.3} \leq D < Q^D_{0.7}$; "Below average" $Q^D_{0.7} \leq D < Q^D_{0.9}$; "Low" $D \geq Q^D_{0.9}$ (Fig. 2a). Hence, "High dataset agreement" corresponds to the lowest 10% of the dataset standardized interquartile ranges among all grid cells (low dataset spread).

In our study, dataset agreement depends on precipitation intensity. Therefore, to compare the dataset agreement for each precipitation intensity (e.g., dataset agreement over heavy precipitation areas), we separately estimated the dataset agreement for each of the ten precipitation intensity classes. In this alternative approach, "High dataset agreement" will represent the 10% of the grid cells with the lowest spread of each intensity class (Fig. 2b).

To understand the contribution of each dataset to dataset (dis-)agreement, we performed two additional steps. Firstly, we estimated the ratio of each dataset global and climatic mean to the ensemble mean of all datasets (Fig. 4). In this manner, we have pinpointed the most/least representative datasets, i.e., the ones that are closest/furthest to the ensemble mean. Then, we used the complex network method

[42], to visualize the relationships between the datasets in terms of their usage by each other, their correlation, and their distance to their means (Fig. 5). As a threshold for the network edges, the highest one third of correlation values and the lowest one third for mean distance values was chosen.

Code and data availability. All source data used are are freely available for download through the *pRecipe* package [37] or at the zenodo repository (https://zenodo.org/records/7078097). All code used in the analysis can be found at https://github.com/imarkonis/ithaca/tree/main/projects/partition_prec and the data relevant to the study outcomes at https://zenodo.org/records/10836849.

Author contributions. Y.M. designed the study and wrote the manuscript, Y.M. and M.R.V.G., performed the analyses, Y.M., M.R.V.G., R.K.P., and S.P. prepared the figures and tables, all authors contributed to the study discussion and to the manuscript editing.

Supplementary information. Additional material can be found in the Supplementary Information.

Acknowledgments. The authors would like to thank K. Trenberth for performing a friendly review and providing constructive comments, as well as the associate editor and the three anonymous reviewers who provided useful feedback and suggestions during the review process. This work was carried out within the project "Investigation of Terrestrial HydrologicAl Cycle Acceleration (ITHACA)" funded by the Czech Science Foundation (Grant 22-33266M).

References

[1] Vargas Godoy, M.R., Markonis, Y., Hanel, M., Kyselỳ, J., Papalexiou, S.M.: The Global Water Cycle Budget: A Chronological Review. Surveys in Geophysics **42**(5), 1075–1107 (2021). Publisher: Springer

[2] Vose, R.S., Schmoyer, R.L., Steurer, P.M., Peterson, T.C., Heim, R., Karl, T.R., Eischeid, J.K.: The global historical climatology network: Long-term monthly temperature, precipitation, sea level pressure, and station pressure data. Technical report, Oak Ridge National Lab., TN (United States). Carbon Dioxide Information . . . (1992)

[3] Huffman, G., Bolvin, D., Nelkin, E., Wolff, D., Adler, R., Gu, G., Hong, Y., Bowman, K., Stocker, E.: The trmm multisatellite precipitation analysis (tmpa): Quasi-global, multiyear, combined-sensor precipitation estimates at fine scales. Journal of hydrometeorology **8**(1), 38–55 (2007)

[4] Schubert, S.D., Rood, R.B., Pfaendtner, J.: An assimilated dataset for earth science applications. Bulletin of the American meteorological Society **74**(12), 2331–2342 (1993)

[5] Kalnay, E., Kanamitsu, M., Kistler, R., Collins, W., Deaven, D., Gandin, L., Iredell, M., Saha, S., White, G., Woollen, J.: The NCEP/NCAR 40-year reanalysis project. Bulletin of the American meteorological Society **77**(3), 437–472 (1996)

[6] Gibson, J., Kållberg, P., Uppala, S., Nomura, A., Hernandez, A., Serrano, E.: Era description, era-15 rep. series 1, ecmwf. Reading, UK (1997)

[7] Trenberth, K.E., Smith, L., Qian, T., Dai, A., Fasullo, J.: Estimates of the global water budget and its annual cycle using observational and model data. Journal of Hydrometeorology **8**(4), 758–769 (2007)

[8] Bhuiyan, M.A.E., Nikolopoulos, E.I., Anagnostou, E.N.: Machine learning–based blending of satellite and reanalysis precipitation datasets: A multiregional tropical complex terrain evaluation. Journal of Hydrometeorology **20**(11), 2147–2161 (2019)

[9] Oki, T., Kanae, S.: Global hydrological cycles and world water resources. science **313**(5790), 1068–1072 (2006)

[10] Korzoun, V.I.: World water balance and water resources of the earth. Studies and Reports in Hydrology **25** (1978). Publisher: UNESCO

[11] Shiklomanov, I.A.: World water resources: A new appraisal and assessment for the 21st century (1998)

[12] Dirmeyer, P.A., Gao, X., Zhao, M., Guo, Z., Oki, T., Hanasaki, N.: GSWP-2: Multimodel analysis and implications for our perception of the land surface **87**(10), 1381–1398 (2006). Publisher: American Meteorological Society

[13] Oki, T.: The hydrologic cycles and global circulation, 13–22 (2006). Publisher: Wiley Online Library

[14] Schlesinger, W.H., Jasechko, S.: Transpiration in the global water cycle. Agricultural and Forest Meteorology **189**, 115–117 (2014)

[15] Rodell, M., Beaudoing, H.K., L'ecuyer, T., Olson, W.S., Famiglietti, J.S., Houser, P.R., Adler, R., Bosilovich, M.G., Clayson, C.A., Chambers, D., *et al.*: The observed state of the water cycle in the early twenty-first century. Journal of Climate **28**(21), 8289–8318 (2015)

[16] Bromwich, D.H., Nicolas, J.P., Monaghan, A.J.: An assessment of precipitation changes over antarctica and the southern ocean since 1989 in contemporary global reanalyses. Journal of Climate **24**(16), 4189–4209 (2011)

[17] Abramowitz, G., Herger, N., Gutmann, E., Hammerling, D., Knutti, R., Leduc, M., Lorenz, R., Pincus, R., Schmidt, G.A.: Esd reviews: Model dependence in multi-model climate ensembles: Weighting, sub-selection and out-of-sample

testing. Earth System Dynamics **10**(1), 91–105 (2019)

[18] Daly, C.: Guidelines for assessing the suitability of spatial climate data sets. International Journal of Climatology: A Journal of the Royal Meteorological Society **26**(6), 707–721 (2006)

[19] Prăvălie, R.: Drylands extent and environmental issues. a global approach. Earth-Science Reviews **161**, 259–278 (2016)

[20] Xie, P., Chen, M., Shi, W.: CPC global unified gauge-based analysis of daily precipitation. In: Preprints, 24th Conf. on Hydrology, Atlanta, GA, Amer. Metero. Soc, vol. 2 (2010)

[21] Harris, I., Osborn, T.J., Jones, P., Lister, D.: Version 4 of the CRU TS monthly high-resolution gridded multivariate climate dataset. Scientific data **7**(1), 1–18 (2020). Publisher: Nature Publishing Group

[22] Tang, G., Clark, M.P., Papalexiou, S.M.: EM-Earth: The Ensemble Meteorological Dataset for Planet Earth. Bulletin of the American Meteorological Society **103**(4), 996–1018 (2022) https://doi.org/10.1175/BAMS-D-21-0106.1 . Publisher: American Meteorological Society Section: Bulletin of the American Meteorological Society

[23] Schneider, U., Becker, A., Finger, P., Meyer-Christoffer, A., Rudolf, B., Ziese, M.: GPCC full data reanalysis version 6.0 at 0.5: monthly land-surface precipitation from rain-gauges built on GTS-based and historic data. GPCC Data Rep., doi **10** (2011)

[24] Chen, M., Xie, P., Janowiak, J.E., Arkin, P.A.: Global land precipitation: A 50-yr monthly analysis based on gauge observations. Journal of Hydrometeorology **3**(3), 249–266 (2002)

[25] Funk, C., Peterson, P., Landsfeld, M., Pedreros, D., Verdin, J., Shukla, S., Husak, G., Rowland, J., Harrison, L., Hoell, A., Michaelsen, J.: The climate hazards infrared precipitation with stations—a new environmental record for monitoring extremes. Scientific Data **2**(1), 150066 (2015) https://doi.org/10.1038/sdata.2015.66 . Number: 1 Publisher: Nature Publishing Group

[26] Xie, P., Arkin, P.A.: Global precipitation: A 17-year monthly analysis based on gauge observations, satellite estimates, and numerical model outputs. Bulletin of the American Meteorological Society **78**(11), 2539–2558 (1997)

[27] Joyce, R.J., Janowiak, J.E., Arkin, P.A., Xie, P.: CMORPH: A method that produces global precipitation estimates from passive microwave and infrared data at high spatial and temporal resolution. Journal of hydrometeorology **5**(3), 487–503 (2004)

[28] Adler, R.F., Sapiano, M.R.P., Huffman, G.J., Wang, J.-J., Gu, G., Bolvin, D., Chiu, L., Schneider, U., Becker, A., Nelkin, E., Xie, P., Ferraro, R., Shin, D.-B.: The Global Precipitation Climatology Project (GPCP) Monthly Analysis (New Version 2.3) and a Review of 2017 Global Precipitation. Atmosphere **9**(4), 138 (2018) https://doi.org/10.3390/atmos9040138 . Number: 4 Publisher: Multidisciplinary Digital Publishing Institute

[29] Huffman, G.J., Stocker, E.F., Bolvin, D.T., Nelkin, E.J., Tan, J.: GPM IMERG Final Precipitation L3 1 Month 0.1 Degree X 0.1 Degree V06, Greenbelt, MD, Goddard Earth Sciences Data and Information Services Center (GES DISC), (2019)

[30] Beck, H.E., Pan, M., Roy, T., Weedon, G.P., Pappenberger, F., Dijk, A.I.J.M., Huffman, G.J., Adler, R.F., Wood, E.F.: Daily evaluation of 26 precipitation datasets using Stage-IV gauge-radar data for the CONUS. Hydrology and Earth System Sciences **23**(1), 207–224 (2019) https://doi.org/10.5194/hess-23-207-2019 . Publisher: Copernicus GmbH

[31] Ashouri, H., Hsu, K.-L., Sorooshian, S., Braithwaite, D.K., Knapp, K.R., Cecil, L.D., Nelson, B.R., Prat, O.P.: PERSIANN-CDR: Daily precipitation climate data record from multisatellite observations for hydrological and climate studies. Bulletin of the American Meteorological Society **96**(1), 69–83 (2015)

[32] Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz-Sabater, J., Nicolas, J., Peubey, C., Radu, R., Schepers, D., *et al.*: The ERA5 global reanalysis. Quarterly Journal of the Royal Meteorological Society **146**(730), 1999–2049 (2020). Publisher: Wiley Online Library

[33] JMA, Japan: JRA-55: Japanese 55-year Reanalysis, Monthly Means and Variances. Research Data Archive at the National Center for Atmospheric Research, Computational and Information Systems Laboratory, Japan Meteorological Agency, Boulder CO (2013). "https://doi.org/10.5065/D60G3H5B"

[34] Bosilovich, M., Lucchesi, R., Suarez, M.: Merra-2: File specification. gmao office note no. 9 (version 1.1) **19**, 73 (2016)

[35] Kanamitsu, M., Ebisuzaki, W., Woollen, J., Yang, S.-K., Hnilo, J.J., Fiorino, M., Potter, G.L.: Ncep–doe amip-ii reanalysis (r-2). Bulletin of the American Meteorological Society **83**(11), 1631–1644 (2002)

[36] Schulzweida, U.: CDO User Guide (2022) https://doi.org/10.5281/zenodo.7112925 . Publisher: Zenodo

[37] Vargas Godoy, M.R., Markonis, Y.: precipe: A global precipitation climatology toolbox and database. Environmental Modelling & Software **165**, 105711 (2023)

[38] Beck, H.E., Zimmermann, N.E., McVicar, T.R., Vergopolan, N., Berg, A., Wood,

E.F.: Present and future köppen-geiger climate classification maps at 1-km resolution. Scientific data **5**(1), 1–12 (2018)

[39] Friedl, M., Sulla-Menashe, D.: Land Cover Type Yearly L3 Global 0.05Deg CMG [Data set]. NASA EOSDIS Land Processes DAAC. ORNL Distributed Active Archive Center (2010). https://webmap.ornl.gov/ogc/dataset.jsp?dg_id=10011_1

[40] Dinerstein, E., Olson, D., Joshi, A., Vynne, C., Burgess, N.D., Wikramanayake, E., Hahn, N., Palminteri, S., Hedao, P., Noss, R., *et al.*: An ecoregion-based approach to protecting half the terrestrial realm. BioScience **67**(6), 534–545 (2017)

[41] Hersbach, H., Bell, B., Berrisford, P., Biavati, G., Horányi, A., et al.: ERA5 hourly data on single levels from 1959 to present. Copernicus Climate Change Service (C3S) Climate Data Store (CDS) (2018)

[42] Tsonis, A.A., Swanson, K.L., Roebber, P.J.: What do networks have to do with climate? Bulletin of the American Meteorological Society **87**(5), 585–596 (2006)

Figures & Tables

Table 1 Mean annual precipitation volume (km^3) for the main Köppen-Geiger climatic classes per dataset type and their terrestrial sum. The standard deviation of each value can be found in Table S2, while the individual values for each data product are presented in Table S3.

Source	Tropical	Arid	Temperate	Continental	Polar	Global
All	51 259	11 528	22 966	20 129	4 415	111 650
Stations	49 596	10 583	22 637	18 198	4 113	105 721
Reanalyses	53 668	12 630	24 227	2 2658	5 036	119 006
Remote Sensing	50 726	11 417	22 300	19 383	4 017	109 932

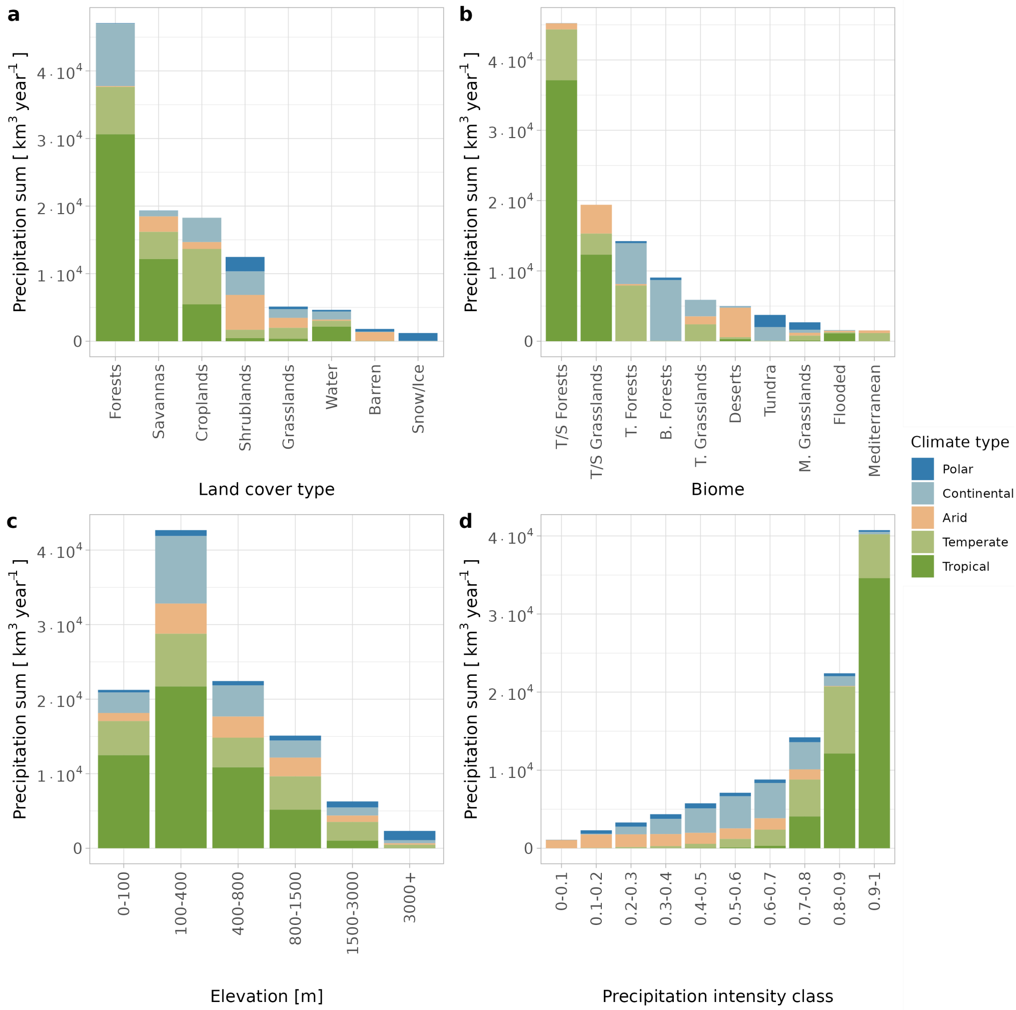

Fig. 1 Global precipitation volume per year and the main Köppen-Geiger classification classes (A: Tropical, B: Dry, C: Temperate, D: Continental, E: Polar) partitioned by (a) land cover types, (b) biomes (T/S Forests: Tropical & Sub-tropical Forests, T/S Grasslands: Tropical & Subtropical Grasslands, Savannas & Shrublands, T. Forests: Temperate Forests, B. Forests: Boreal Forests/Taiga, T. Grasslands: Temperate Grasslands, Savannas & Shrublands, Savannas & Shrublands, Deserts & Xeric Shrublands, Tundra, M. Grasslands: Montane Grasslands & Shrublands, Flooded: Mangroves & Flooded Grasslands/Savannas, Mediterranean: Mediterranean Forests, Woodlands & Scrublands), (c) elevation zones, and (d) precipitation intensity classes

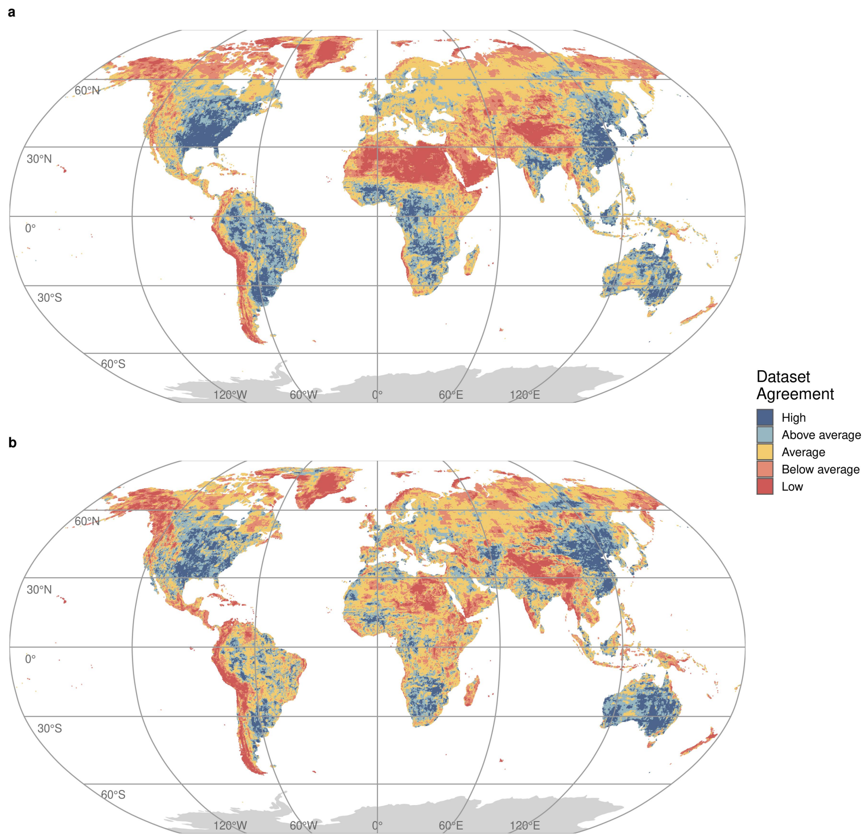

Fig. 2 Maps of dataset agreement derived by the standardized interquartile range of (a) all grid cells, (b) conditioned over corresponding precipitation intensity class.

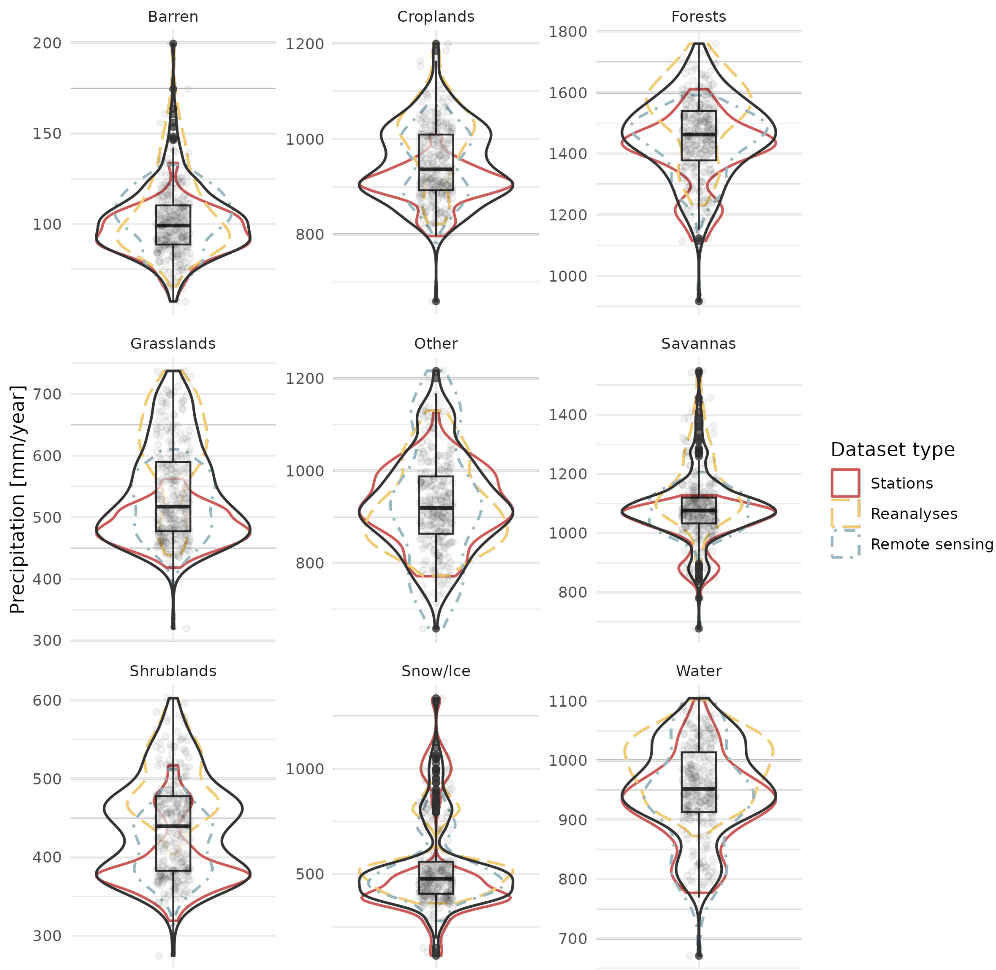

Fig. 3 Mean annual precipitation of all datasets for each land cover and data set type. The black line and the box plot correspond to all three sources. Points represent annual values from individual data sets.

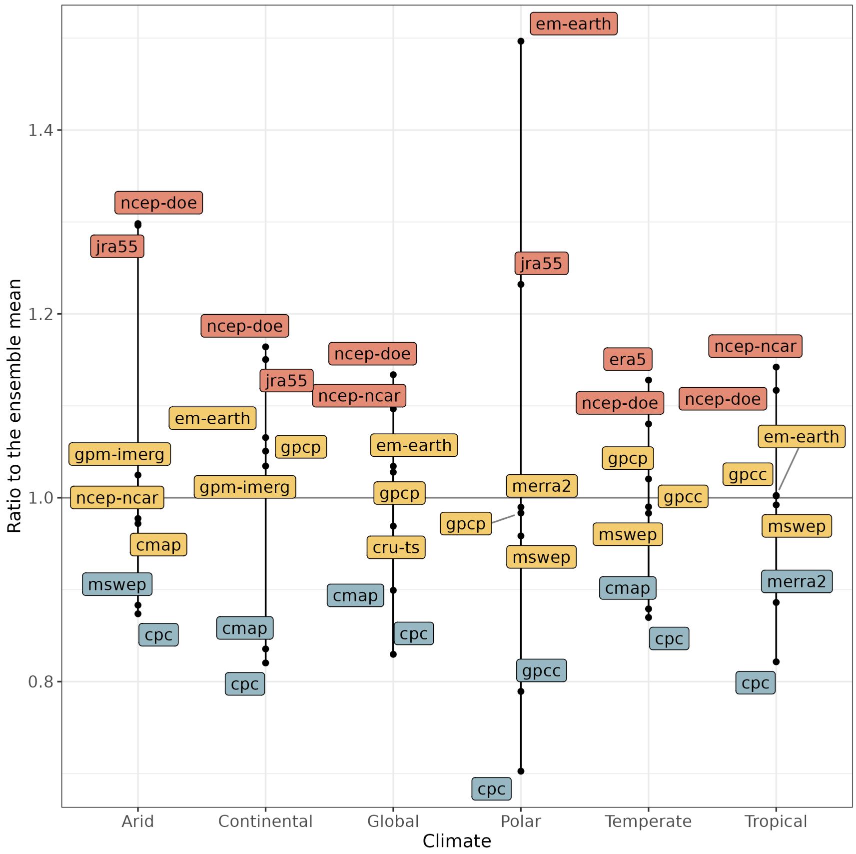

Fig. 4 Dataset (dis-)agreement of individual data products per climate class. The three datasets with annual estimates closest to the ensemble mean and the two with the lowest/highest means.

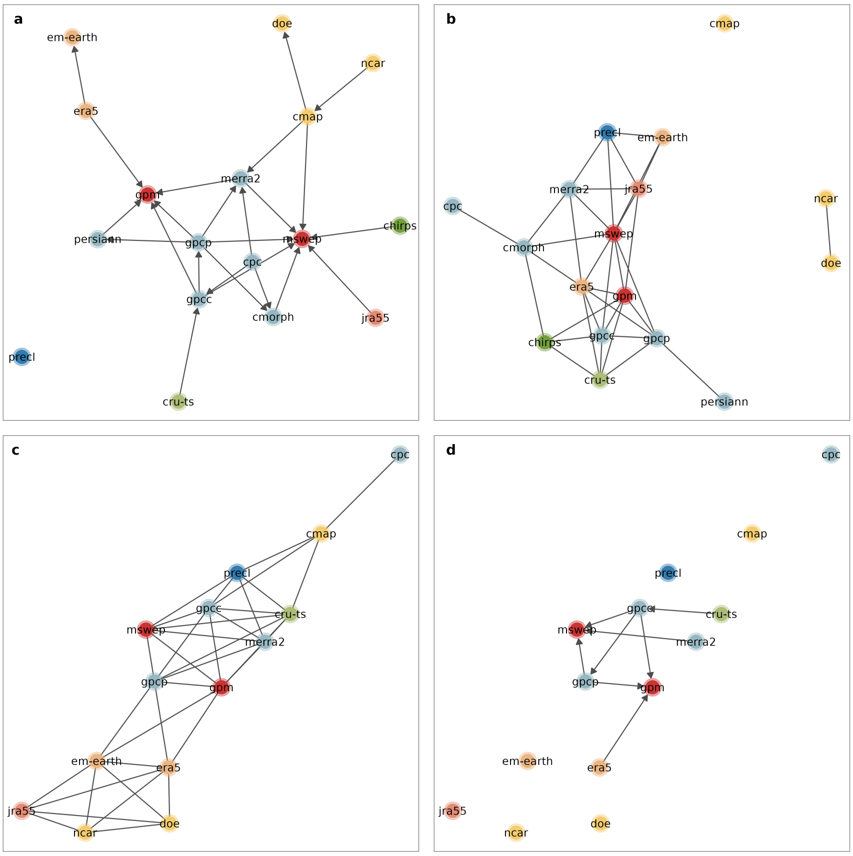

Fig. 5 (a) Dataset generation relationships (dataset "genealogies"). The arrows show the direction of data application (e.g., GPCC employs CRU-TS). Same color suggest a data product family that share sources. GPM-IMERG and MSWEP are considered an individual family as they only employ data from five or more sources but are not used in any other data product. (b) Dataset cross-correlation network. The network edges represent the highest one-third of the correlated pairs among the datasets. (c) Dataset mean distance network. The network edges represent the smallest one-third of the mean distance among each dataset pair. CMORPH, CHIRPS and PERSIANN not included due to the limitation on global coverage. (d) Dataset generation relationships after keeping only the cross-correlation and mean distance network edges that appear in Figures (b) and (c).