GloH2O

MSWEP V3.15 Documentation

Important: Please review the Version History section. The daily correction was not applied, resulting in lower-than-intended accuracy in regions with dense gauge coverage.

Version: 3.15

October 16, 2025

GloH2O LLC

Contents

GloH2O LLC
100 Overlook Ctr, 2nd floor
Princeton, NJ 08540, USA

1 Introduction

This document provides an overview of Multi-Source Weighted-Ensemble Precipitation (MSWEP), the only fully global, machine learning-based gridded precipitation product. MSWEP has an hourly 0.1° resolution and spans the period from 1979 to approximately two hours from real-time. Unique features of MSWEP include its unique blending of (re)analysis outputs, satellite estimates, and gauge observations, ensuring reliable precipitation estimates across different climatic regions, from frontal-dominated, to convection-dominated, and densely gauged regions. Additionally, it is one of the few precipitation products incorporating daily gauge observations and accounting for gauge reporting times. The satellite-(re)analysis baseline precipitation estimation algorithm is described in Wang et al. (2025) and Wang and Beck (2025), while the gauge correction scheme is described in Wang and Beck (2025). The dataset has been rigorously evaluated in multiple studies, including Beck et al. (2017b), Beck et al. (2019a), Abbas et al. (2025), and Wang and Beck (2025).

2 Data license

MSWEP is distributed under the Creative Commons Attribution-NonCommercial (CC BY-NC) 4.0 license (`https://creativecommons.org/licenses/by-nc/4.0/deed.en`). This license permits the use, sharing, and adaptation of the dataset for noncommercial purposes, provided that appropriate credit is given in the form of citations to Wang et al. (2025) and Wang and Beck (2025). For commercial use, a separate license must be obtained by contacting us. Since the definition of "commercial" can sometimes be ambiguous, we provide the following clarifications.

Noncommercial use

The following types of usage are classified as noncommercial and do not require additional permissions:

- **Academic research**: Usage by university researchers, students, and educators for research, teaching, or publication, provided it is not part of a paid contract or consulting service.
- **Nonprofit scientific research**: Research conducted by nonprofit institutions, excluding cases involving commercial partnerships, consulting services, or operational applications.
- **Personal use**: Downloading or analyzing the dataset for personal learning, hobbyist exploration, or individual, noncommercial projects.
- **Government and NGO use**: Utilization by government agencies and non-governmental organizations (NGOs) for non-revenue-generating projects (e.g., public research reports), provided no commercial activities, commercial partnerships, or operational applications are involved.

If there is any uncertainty, please contact us for clarification.

GloH2O LLC
100 Overlook Ctr, 2nd floor
Princeton, NJ 08540, USA

Commercial use

The following types of usage are classified as commercial and require a separate license. Please contact us for further details.

- **For-profit company use**: Any use within a for-profit company or startup, even if limited to internal research.
- **Consulting, advisory services, or contract research**: Use of the data in reports or analyses for paying clients.
- **Commercial integration and utilization**: Any use or incorporation of the dataset in the development, operation, or provision of commercial products, tools, or software that are sold, licensed, or otherwise monetized.
- **Paywalled services**: Usage as part of a paywalled service, subscription model, or data platform.

If you are uncertain whether your intended use qualifies as noncommercial, please contact us for clarification.

3 Data attribution

Users of MSWEP must cite the following references in reports, publications, presentations, and all other forms of dissemination:

- Wang, X., Alharbi, R. S., Baez-Villanueva, O. M., Green, A., McCabe, M. F., Wada, Y., Van Dijk, A. I. J. M., Abid, M. A., & Beck, H. E. (2025). Saudi Rainfall (SaRa): Hourly 0.1° Gridded Rainfall (1979–Present) for Saudi Arabia via Machine Learning Fusion of Satellite and Model Data. *Hydrology and Earth System Sciences* 29, 4983–5003.
- Wang, X., & Beck, H. E. (2025). MSWEP V3: Machine Learning-Powered Global Precipitation Estimates at 0.1° Hourly Resolution (1979–Present). In preparation.

We encourage users to specify the version number of MSWEP when citing the dataset, to ensure reproducibility.

4 Historical MSWEP

The historical MSWEP dataset — covering the period from 1979 to a few months before the present — was produced using one reanalysis precipitation product (ERA5), three satellite-based precipitation products (GSMaP, IMERG, and PERSIANN-CCS-CDR), and gauge observations from multiple sources (GSOD, GHCN-D, GPCC, and various national databases). Two variants of the historical MSWEP dataset are available in the subfolders `Past` and `Past_nogauge`, representing the satellite-reanalysis baseline with and without gauge corrections, respectively. The uncorrected variant is recommended only for:

1. Evaluating the baseline MSWEP against independent gauge observations.
2. Computing near-real-time (NRT) anomalies such as event return periods.

The corrected variant is suitable for a broader range of applications, including:

1. Hydrological and water resource modeling.
2. Calibration and validation of hydrological and land surface models.
3. Climate variability and change assessments.
4. Agricultural and crop yield simulations.

5 MSWEP-NRT

MSWEP-NRT extends the historical MSWEP record to near real-time with a latency of approximately two hours, using the ERA5 reanalysis, GDAS analysis, and satellite-based GSMaP, IMERG, and PDIR-Now products. MSWEP-NRT data are stored in the `NRT` subfolder. To account for latency differences among data sources and potential data disruptions, MSWEP-NRT files for the past ten days are progressively updated as new data become available. Users should redownload upgraded files to ensure the highest accuracy.

The initial hourly MSWEP-NRT files rely solely on PDIR-Now, which has the shortest latency (approximately two hours) but also the lowest accuracy and no high-latitude coverage (>50/60°N/S). After roughly four hours, they are updated with preliminary, low-latency variants of GSMaP and IMERG (called "NRT" and "Early," respectively), along with GDAS data. After about one day, they are further updated using more reliable, delayed versions of GSMaP and IMERG ("Standard" and "Late," respectively). After five days, ERA5 replaces GDAS, ensuring consistency with the historical MSWEP. Similar to the hourly files, the 3-hourly, daily, and monthly files are initially preliminary and subsequently updated as the hourly files are updated.

6 Data format

Gauge-corrected historical MSWEP data are stored in the `Past` subfolder, uncorrected historical MSWEP data in the `Past_nogauge` subfolder, and near real-time MSWEP data in the `NRT` subfolder. The historical dataset covers 1979–2024, while the near real-time dataset extends from 2025 to approximately two hours from real-time. The data are available at hourly, 3-hourly, daily, and monthly temporal resolutions, with a spatial resolution of 0.1° (approximately 11 km at the equator), in the self-describing netCDF-4 format (`www.unidata.ucar.edu/software/netcdf/`).

The file naming convention follows `YYYYDOY.HH` for hourly and 3-hourly files, `YYYYDOY` for daily files, and `YYYYMM` for monthly files. Here, `YYYY` represents the year, `MM` the month, `DOY` the day of the year, and `HH` the starting hour for the hourly or 3-hourly accumulation. Precipitation estimates are stored in the `precipitation` netCDF field (dimensions 1800×3600) in units of mm/hour, mm/3-hour, mm/day, or mm/month, depending on the temporal resolution. For instance, the file `MSWEP_V315/Past/Hourly/2020116.18.nc` contains hourly precipitation data from 18:00 to 19:00 UTC on April 25, 2020.

Hourly MSWEP-NRT files include an additional field, `model_id`, with dimensions 1800×3600. This field contains values from 1 to 18, indicating the combination of dynamic predictors used for each

GloH2O LLC
100 Overlook Ctr, 2nd floor
Princeton, NJ 08540, USA

Table 1: The dynamic predictors incorporated in the different ML model stacks.

model_	ERA5 *P*	ERA5 *T*	GDAS *P*	GDAS *T*	IMERG-L V07	GSMaP-MVK V8	P-CCS-CDR	PDIR-Now
01	✓	✓	✗	✗	✓	✓	✗	✗
02	✓	✓	✗	✗	✓	✗	✗	✗
03	✓	✓	✗	✗	✗	✓	✗	✗
04	✓	✓	✗	✗	✗	✗	✓	✗
05	✓	✓	✗	✗	✗	✗	✗	✗
06	✓	✓	✓	✓	✓	✓	✗	✗
07	✗	✗	✓	✓	✓	✓	✗	✗
08	✓	✓	✓	✓	✓	✗	✗	✗
09	✓	✓	✓	✓	✗	✓	✗	✗
10	✗	✗	✓	✓	✓	✗	✗	✗
11	✗	✗	✓	✓	✗	✓	✗	✗
12	✗	✗	✗	✗	✓	✓	✗	✗
13	✗	✗	✗	✗	✓	✗	✗	✗
14	✗	✗	✗	✗	✗	✓	✗	✗
15	✗	✗	✓	✓	✗	✗	✗	✓
16	✓	✓	✓	✓	✗	✗	✗	✗
17	✗	✗	✓	✓	✗	✗	✗	✗
18	✗	✗	✗	✗	✗	✗	✗	✓

grid cell (see Table 1). At low and mid latitudes, when `model_id` is 1, the precipitation estimates can be considered 'final' and are consistent with the historical MSWEP. At high latitudes, this is the case when `model_id` is 5.

7 Gauge metadata

A set of auxiliary CSV files is included in the `Gauge_metadata` folder. These files contain metadata for the daily and monthly gauges used in the MSWEP gauge correction procedure:

- The files `daily_station_locations.csv` and `monthly_station_locations.csv` provide the latitude and longitude coordinates of the daily and monthly gauges, respectively. Only daily gauges that passed quality control and deduplication are included. Daily gauge identifiers (e.g., `GHCND_GME00010350`) correspond to station codes from the original data sources. Monthly gauge identifiers (e.g., `gridcell_00042230`) refer to 0.25° GPCC grid cells with available data. The numeric suffix indicates the grid cell's flattened index in a global 720×1440 grid using row-major order, starting from the top-left of the grid.
- The files `daily_station_date_ranges.csv` and `monthly_station_date_ranges.csv` indi-

cate the first and last valid observations for each gauge. These dates define the temporal span of available data but do not imply continuous coverage. Many gauges contain gaps, including missing days, months, or years.

- The file `daily_station_reporting_times.csv` contains the inferred reporting time for each daily gauge, expressed as an integer offset from 00:00 UTC. A value of –5, for example, indicates that the gauge likely reports precipitation totals ending at 19:00 UTC (i.e., 5 hours before midnight UTC). A value of 0 implies alignment with the UTC calendar day, while a value of +3 indicates reporting at 03:00 UTC on the following day.

8 Common questions

For answers to frequently asked questions regarding MSWEP usage, methodology, data access, and troubleshooting, refer to the FAQ section on the official website: `www.gloh2o.org/mswep`.

9 Version history

Version 3.15 (July 29, 2025)

Important note (October 15, 2025): A technical issue in version 3.15 caused the monthly correction to overwrite the daily gauge correction. As a result, day-to-day variability may be less accurate than expected in regions with dense gauge coverage (notably Europe, the US, and Australia). This issue affects only the historical dataset (in the `Past` folder). Users requiring the highest-accuracy historical data are advised to use version 2.8 until the issue is resolved in the near future.

This significant update reflects the improvements detailed in Wang et al. (2025) and Wang and Beck (2025). The key improvements are:

1. Introduction of a machine learning-based algorithm for bias correction and seamless merging of multiple data sources.
2. Replacement of GridSat with PERSIANN-CCS-CDR as the primary infrared data source for periods prior to 2000.
3. Integration of PDIR-Now to provide near real-time precipitation estimates before GSMaP and IMERG data are available, reducing the latency to approximately 2 hours.
4. After incorporating ERA5, MSWEP-NRT estimates are fully consistent with the historical MSWEP dataset.
5. Enhanced daily gauge correction scheme that accounts for spatial variability in precipitation correlation length.
6. GPCC data used for monthly corrections.

GloH2O LLC
100 Overlook Ctr, 2nd floor
Princeton, NJ 08540, USA

Version 2.8 (February 14, 2021)

Major update featuring new data sources, a longer data record, near real-time estimates, improved weight maps, less peaky precipitation estimates, and compatibility with the Multi-Source Weather (MSWX) product. Full list of changes:

1. The five datasets incorporated in the previous version (CMORPH, TMPA 3B42RT, GSMaP-MVK, ERA-Interim, and JRA-55) were replaced with just two datasets (ERA5 and IMERG) with superior performance. GridSat is still used prior to 2000.
2. The record was extended to December 31, 2020. Near real-time estimates (based on GSMaP, IMERG, ERA5, and GDAS) are available with a latency of ~3 hours (see Section 5).
3. The merging was performed for four air temperature zones (<0°C, 0°C to 10°C, 10°C to 20°C, and >20°C) as opposed to two (<5°C and >5°C), to better account for the influence of precipitation regime (convective versus frontal) on the performance of the datasets.
4. Weight maps for the datasets were produced, for each temperature zone, using random forest regression based on seven climate- and topography-related predictors. In previous versions, the weight maps were derived using spatial interpolation and therefore lacked detail in sparsely gauged regions.
5. The 3-hourly GridSat infrared data represents images acquired at 00:00, 03:00, …, 21:00 UTC, whereas the 3-hourly MSWEP data represent accumulations during 00:00–02:59, 03:00–05:59, …, 21:00–23:59 UTC. To resolve this inconsistency, we downscaled the 3-hourly GridSat infrared data to 30 minutes using cubic interpolation, and calculated means for the intervals 00:00–02:59, 03:00–05:59, …, 21:00–23:59 UTC.
6. Our algorithm to estimate precipitation from GridSat infrared data was changed. In this version, we CDF-matched the 3-hourly infrared brightness temperature data to CMORPH V1.0 for the aforementioned four air temperature zones.
7. The MSWX meteorological product was used as reference cumulative distribution function (CDF) to correct the merged satellite-reanalysis estimates. MSWX forecasts are therefore consistent with MSWEP. Since MSWX is based on ERA5, the new CDF of MSWEP is significantly less peaky.
8. Unlike in previous versions, we did not correct precipitation underestimation over mountainous and snow-dominated regions, because users generally give preference to precipitation estimates that match rain gauge observations as closely as possible. To obtain a bias-corrected version of MSWEP — important for hydrological modeling — we recommend using the PBCOR dataset (`www.gloh2o.org/pbcor`; Beck et al., 2020).
9. Applied monthly climatological corrections using CHELSA (Karger et al., 2017) instead of WorldClim (Fick and Hijmans, 2017).
10. Improved quality control of daily gauge observations, by discarding months with <24 observations and discarding gauges with implausible annual precipitation trends (defined as > ±30 %/year).
11. This version does not include monthly gauge corrections as the GPCC dataset was not yet available for 2020.
12. Completely rewrote the code to be more efficient.

GloH2O LLC
100 Overlook Ctr, 2nd floor
Princeton, NJ 08540, USA

Version 2.2 (March 2, 2018)

This version corresponds to the description in Beck et al. (2019b). This version includes three changes:

1. The dataset was extended to October 31, 2017 (reflecting the availability of ERA-Interim).
2. We conducted some additional quality control of the GridSat infrared archive and now incorporate GridSat-based rainfall estimates from 1980 (rather than 1983) onwards.
3. In version 2.1, the long-term precipitation trends were amplified due to the CDF corrections, particularly in the tropics during the pre-TRMM era. This issue was resolved in version 2.2 by rescaling the estimates to match the trends of the non-CDF corrected data.

Version 2.1 (November 20, 2017)

Two changes. First, NCEP-CFSR precipitation data were removed due to the presence of spurious trends (NCEP-CFSR was added in version 2.0). Second, corrections using the monthly gauge-based GPCC precipitation dataset were reintroduced (these were removed in version 2.0).

Version 2.01 (October 8, 2017)

Fixed a mistake in the netCDF metadata. For all netCDF files, the units for the time variable were changed from "days since 1900-01-01 00:00:00" to "days since 1899-12-31 00:00:00". The actual time data have not been changed. In the case of the 3-hourly data, a (fictional) time value of, for example, 1.125 would indicate that the data represent the period 03:00Z to 05:59Z on January 1, 1900.

Version 2.0 (July 23, 2017)

Major upgrade containing numerous changes in the data sources as well as the merging algorithm, and as a result markedly different precipitation estimates, especially at the daily and 3-hourly time scales (notably less drizzle and higher peaks). The most important changes in version 2 include:

1. The correction of distributional precipitation biases to account for the spurious drizzle and attenuated peaks evident in version 1.
2. Increasing the spatial resolution from 0.25° to 0.1° to increase the local relevance of the precipitation estimates.
3. The inclusion of ocean areas, to enable oceanic studies and avoid missing data in coastal areas.
4. The addition of precipitation estimates derived from GridSat thermal infrared imagery for the pre-TRMM era to supplement the reanalysis and gauge data.
5. Daily gauge corrections for each grid-cell based on the five nearest gauges, to replace the coarse 0.5° CPC Unified.
6. The use of a daily gauge correction scheme that accounts for differences in gauge reporting times, to minimize timing mismatches when merging the daily gauge estimates with the satellite and reanalysis data.

7. Extension of the data record to 2017.

Version 1.2 (November 28, 2016)

Despite the many changes in version 1.2, the terrestrial precipitation estimates have not changed considerably since the previous version. The full list of changes is as follows:

1. MSWEP now also provides experimental precipitation estimates for ocean areas. When using MSWEP ocean data it should be kept in mind that: (i) the weights used for the temporal dynamics are almost entirely based on land stations which are not necessarily representative of ocean areas; (ii) the gauge-based data sources (CPC Unified and GPCC) are unavailable over ocean areas; and (iii) the estimates have not been validated (this is work in progress).
   For determining the long-term mean over ocean areas we could not use CHPclim as it only covers land areas. The long-term mean over ocean areas was therefore derived by weighting the long-term means of CMORPH, TMPA 3B42RT, GSMaP-MVK, ERA-Interim, and JRA-55. The weights for CMORPH, TMPA 3B42RT, and GSMaP-MVK were set to 1 for latitudes < 25˚ and to 0 for latitudes > 35˚. The weights decrease linearly from 1 at 25˚ to 0 at 35˚. The weights for ERA-Interim and JRA-55 were set to 1. In the future, we intend to refine the weight estimates over ocean areas to obtain more reliable long-term means.
2. The record has been extended from 2014 to 2015.
3. ERA-Interim data were mistakenly offset by +3 hours in previous versions.
4. The threshold temperature for the inclusion of satellite data has been increased to 5°C for all time scales, to minimize the probability of incorporating potentially erroneous satellite data.
5. Satellite data prior to the year 2000 have been excluded. Among the satellite sources, only CMORPH provides data prior to 2000. However, the data were in lesser agreement with the gauge and reanalysis estimates and have therefore been excluded.
6. For generating the weight maps used for determining the temporal dynamics, the GHCN-D and GSOD station data are now normalized prior to the computation of grid-cell average time series.
7. The weight maps have been produced at 0.25° rather than 0.5°.
8. For each grid-cell, rather than normalizing the satellite data, we rescaled the satellite data to match the reanalysis data for the period of overlap, to ensure retainment of the long-term trends.

Version 1.1 (August 2, 2016)

No changes to the actual data, the only changes are in the netCDF formatting. First, we changed the order of the variables from “lon, time, lat” to “time, lat, lon”, which should solve some of the problems people have encountered when reading the data. Second, we corrected the time variable for the daily data, which was mistakenly offset by 1 day.

Version 1.0 (May 30, 2016)

Initial release corresponding to the description in Beck et al. (2017a).

GloH2O LLC
100 Overlook Ctr, 2nd floor
Princeton, NJ 08540, USA

References

Abbas, A., Yang, Y., Pan, M., Tramblay, Y., Shen, C., Ji, H., Gebrechorkos, S. H., Pappenberger, F., Pyo, J., Feng, D., Huffman, G., Nguyen, P. D., Massari, C., Brocca, L., and Beck, H. E. (2025). Comprehensive global assessment of 23 gridded precipitation datasets across 16,295 catchments using hydrological modeling. *Hydrology and Earth System Sciences Discussions*. Submitted.

Beck, H. E., McVicar, T. R., Zambrano-Bigiarini, M., Alvarez-Garret, C., Baez-Villanueva, O. M., Sheffield, J., Karger, D., and Wood, E. F. (2020). Bias correction of global precipitation climatologies using discharge observations from 9372 catchments. *Journal of Climate*, 33:1299–1315.

Beck, H. E., Pan, M., Roy, T., Weedon, G. P., Pappenberger, F., van Dijk, A. I. J. M., Huffman, G. J., Adler, R. F., and Wood, E. F. (2019a). Daily evaluation of 26 precipitation datasets using Stage-IV gauge-radar data for the CONUS. *Hydrology and Earth System Sciences*, 23:207–224.

Beck, H. E., van Dijk, A. I. J. M., Levizzani, V., Schellekens, J., Miralles, D. G., Martens, B., and de Roo, A. (2017a). MSWEP: 3-hourly 0.25° global gridded precipitation (1979–2015) by merging gauge, satellite, and reanalysis data. *Hydrology and Earth System Sciences*, 21(1):589–615.

Beck, H. E., Vergopolan, N., Pan, M., Levizzani, V., van Dijk, A. I. J. M., Weedon, G. P., Brocca, L., Pappenberger, F., Huffman, G. J., and Wood, E. F. (2017b). Global-scale evaluation of 22 precipitation datasets using gauge observations and hydrological modeling. *Hydrology and Earth System Sciences*, 21(12):6201–6217.

Beck, H. E., Wood, E. F., Pan, M., Fisher, C. K., Miralles, D. G., van Dijk, A. I. J. M., McVicar, T. R., and Adler, R. F. (2019b). MSWEP V2 global 3-hourly 0.1° precipitation: methodology and quantitative assessment. *Bulletin of the American Meteorological Society*, 100(3):473–500.

Fick, S. E. and Hijmans, R. J. (2017). WorldClim 2: new 1-km spatial resolution climate surfaces for global land areas. *International Journal of Climatology*, 37(12):4302–4315.

Karger, D. N., Conrad, O., Böhner, J., Kawohl, T., Kreft, H., Soria-Auza, R. W., Zimmermann, N. E., Linder, H. P., and Kessler, M. (2017). Climatologies at high resolution for the earth's land surface areas. *Scientific Data*, 5:170122.

Wang, X., Alharbi, R. S., Baez-Villanueva, O. M., Green, A., McCabe, M. F., Wada, Y., Van Dijk, A. I. J. M., Abid, M. A., and Beck, H. E. (2025). Saudi Rainfall (SaRa): hourly 0.1° gridded rainfall (1979–present) for Saudi Arabia via machine learning fusion of satellite and model data. *Hydrology and Earth System Sciences*, 29(19):4983–5003.

Wang, X. and Beck, H. E. (2025). MSWEP V3: Machine learning-powered global precipitation estimates at 0.1° hourly resolution (1979–present. In prep.