Response to Reviewer #2

This paper introduces a novel precipitation dataset for the Tibetan Plateau. Given that the Himalaya–Tibetan Plateau region continues to stand out in global precipitation datasets (both in satellite retrievals as well as in high-resolution model simulations) continued efforts to improve precipitation quantification by combining different data sources are highly valuable. Such datasets can serve both practical hydrological applications and as useful references for model evaluation. However, I think there are still a few issues that should be addressed before the paper is ready for publication. In particular, the manuscript would benefit from a clearer explanation of the added value of the new dataset and from additional evidence demonstrating that it provides improvements over existing gridded precipitation datasets, as this appears to be one of the main motivations and objectives of the study. I also strongly recommend clarifying several aspects of the methodology and being more cautious in some instances where the results or novelty of the dataset may currently be overstated. Overall, I see value in the work, and I believe that addressing these points would substantially strengthen the manuscript. Please see my detailed comments below.

Response:

We sincerely thank the reviewer for the thoughtful and constructive evaluation of our manuscript. We appreciate the recognition that improving precipitation estimates over the Tibetan Plateau region is important for both practical hydrological applications and the evaluation of climate and Earth system models. We agree that the revised manuscript should provide a clearer explanation of the added value of CRISP and stronger evidence demonstrating its performance relative to existing gridded precipitation datasets. We also acknowledge the need to clarify several methodological details and to use more cautious language when describing the novelty and scientific interpretation of the dataset.

In response, we will revise the Introduction to explain the CRISP framework and its added value more clearly at an earlier stage. We will expand the comparison with existing precipitation products and provide a more systematic description. We will also strengthen the discussion of the uncertainties associated with station observations, wind-induced precipitation undercatch, ERA5 near-surface winds, and sub-grid-scale precipitation processes.

Furthermore, we will provide additional clarification and, where possible, supplementary analyses concerning the independent validation data, the contribution of the GTS observations, and the role of the different input datasets. We will revise statements that may overstate the novelty of CRISP and describe it more precisely as a regional, multi-source, uncertainty-aware precipitation dataset that integrates existing observations, satellite retrievals, and reanalysis information while addressing precipitation biases and phase-related uncertainties over the Third Pole region.

Finally, we will improve the accessibility of the dataset and providing easier way for data access and use. We thank the reviewer for these valuable comments, which have helped us identify important areas for improving the transparency, rigor, and practical usefulness of the manuscript.

Major comments

Data availability: First of all, the link to the dataset in the abstract is not working, which constitutes a major obstacle to properly assessing the dataset and, consequently, reviewing the publication. There is another link in the data section of the paper but I could not succeed to download any files. It says that I should use the FTP client tools but no proper documentation on how to download the data is given. Given that the dataset is only 5 GB (Is that really true? I would have expected at least a few hundred GB for 30 years of daily data), there should be ways to easily download the files with one click. Alternatively, please provide instructions on how to access the data. Since this is a dataset-based publication, I would prefer to reserve my final assessment until the dataset is made available and can be evaluated directly. In my view, the final decision regarding the publishability of the manuscript should take into account the quality, completeness, and usability of the underlying dataset.

Response:

Thank you for pointing this out. **We have created a new version of the existing Zenodo record and uploaded all of the original data as a backup (https://zenodo.org/records/22045710).** The revised Zenodo record is now accessible through the link provided in the Data Availability section. The files can be downloaded directly through the web interface without installing an FTP client or following any additional technical procedures. We have tested the revised link and confirmed that the files are accessible and downloadable.

For completeness, the original FTP distribution method remains available. To access the data

through FTP, users should install an FTP client, such as FTP Rush, WinSCP, FileZilla, or an equivalent application, and enter the following connection information:

Host: ftp2.tpdc.ac.cn or ftp3.tpdc.ac.cn

Port: 6201

Username: download_52680702

Password: 49294969

We have retested the FTP connection and confirmed that the data can be accessed using these settings.

Regarding the relatively small total size of the dataset, the approximately 5 GB size refers to the compressed distributed files and is a consequence of both the dataset design and the storage format. First, the variables were stored using appropriate data types: the precipitation-related variables (Prcp_Best, Prcp_5th, Prcp_95th, Prcp_Clim, and Prob_Rain) were stored as 32-bit floating-point values, which are sufficient for preserving the relevant precision of millimeter-scale precipitation estimates, while the phase variable, whose values are limited to four categories (0, 1, 2, and 3), was stored as an 8-bit integer. Second, the files were saved in NetCDF4 format with lossless zlib compression at compression level 4. Finally, precipitation fields were hard-gated so that non-precipitating grid cells were assigned exactly zero rather than retaining small numerical noise from the model output. This produces large spatially coherent zero-value areas and substantially improves lossless compression without removing precipitation information. In addition, the spatial extent of the study area and the daily temporal resolution at 0.1° grid spacing are consistent with a compressed dataset of this size.

Added value of the new precipitation estimates: While there is definitely value in merging different data sources, adding user-friendly information and making the data accessible in a nice way, I think it is overstated that new information is really created here, rather than bias correction of the dry bias in precipitation in the satellite data that becomes as the new product uses gauge measurements as the target data. I acknowledge that it is hard to fully validate the correction against ordinary uncorrected station precipitation because the reference gauge data contain the same systematic bias in wind-induced undercatch. Instead the validation can focus on whether the correction improves consistency with independent hydrological and atmospheric constraints which is partly done here

in Fig. 12 by looking at the water balance. However, I think the science problem behind this validation is not discussed in sufficient detail and references to how others have addressed this issue could be important (e.g. Kochendorfer, J., Rasmussen, R., Wolff, M., Baker, B., Hall, M. E., Meyers, T., ... & Leeper, R. (2017). The quantification and correction of wind-induced precipitation measurement errors. Hydrology and Earth System Sciences, 21(4), 1973-1989.) In addition, it is not explained where the SWE measurements come from and what biases they may contain. Can the SWE from ERA5 be added to show that the novel estimates are closer to the observed SWE than what ERA5 precipitation would be to the ERA5-derived SWE?

Response:

We thank the reviewer for this important comment. We agree that the scientific contribution of CRISP should be described more carefully. CRISP is not intended to represent the creation of completely independent precipitation information. Rather, its main contribution is the integration of multiple data sources to provide precipitation estimates together with two additional types of information that are generally unavailable in conventional precipitation products: **the uncertainty range of the estimates and the diagnosed precipitation phase**. In addition, the integration and bias-adjustment framework improve the spatial and temporal consistency of the precipitation estimates over the Tibetan Plateau. We have therefore revised the relevant statements throughout the manuscript to avoid overstating the generation of "new" precipitation information and to more accurately describe CRISP as a multi-source precipitation estimation and uncertainty-quantification product.

We also agree that the use of gauge observations as reference data is subject to uncertainties, particularly the wind-induced undercatch of solid and mixed precipitation. To address this issue, we have added a discussion of gauge-measurement errors and their implications for precipitation evaluation. **We now cite Kochendorfer et al. (2017) and explicitly acknowledge that gauge observations should not be regarded as an unbiased representation of true precipitation, especially under windy conditions and for snowfall.** Consequently, the gauge-based evaluation in this study is interpreted as an assessment of consistency with the available observations rather than as an evaluation against an error-free ground truth.

We agree that comparing the SWE evolution associated with CRISP precipitation and ERA5 precipitation would provide an informative additional assessment. However, ERA5 variables were

used as environmental predictors in the development of CRISP. Therefore, ERA5-derived SWE would not constitute a fully independent reference for evaluating CRISP. Moreover, a comparison between ERA5 precipitation and ERA5-derived SWE could partly reflect the internal consistency of the ERA5 system, rather than the ability of the precipitation product to reproduce independently observed snow accumulation and ablation. For this reason, we used the most reliable and representative station-based SWE observations available in the study area as the primary cryospheric constraint. The stations were selected according to the quality-control and representativeness criteria described in Miao et al. (2024). These station observations were not used as environmental predictors in CRISP and therefore provide a more appropriate basis for assessing the consistency of the precipitation estimates with observed snow-water storage. Meanwhile, we have clarified the source, processing background, and intended use of these SWE data in the revised manuscript. Since the uncertainty information and a complete error characterization of the SWE dataset are not independently available to us, we do not assign a quantitative uncertainty range to these data in the present study. **Instead, we now explicitly state that the SWE data may contain uncertainties associated with measurement, spatial representativeness, depending on the original data source and methodology described by Miao et al. (2024).** The SWE-based analysis is therefore used as an additional consistency constraint rather than as an absolute benchmark.

To further examine the added value of CRISP, we have expanded the comparison with other precipitation products. In addition to the comparisons presented in the original manuscript, we now include AERA5-Asia, AIMERG, and GMCP. These products provide complementary references based on different data sources and retrieval or assimilation strategies. **The expanded comparison shows that CRISP provides competitive precipitation estimates while additionally supplying uncertainty intervals and precipitation-phase information. The results also provide further evidence that the value of CRISP lies not only in its precipitation amount estimates**, but also in the additional diagnostic information and the improved usability of the dataset for hydrological and cryospheric applications.

****

Please see Introduction in detail:

However, extreme topography, harsh climatic conditions, and the sparse distribution of in situ observations make reliable precipitation quantification a persistent challenge (Gao & Liu, 2013).

To address this challenge, we develop the Conformal Regression and Integrated Stacking for Precipitation (CRISP) framework. CRISP uses a machine-learning-based data integration strategy to combine complementary information from ground-based precipitation observations, satellite and reanalysis precipitation products, static topographic variables, and dynamic atmospheric predictors. The observation-based data provide local precipitation constraints, while satellite and reanalysis products provide spatially continuous information in areas with limited gauge coverage. Topographic and atmospheric variables are used to describe the spatial and temporal controls on precipitation, including the effects of elevation, terrain structure, atmospheric circulation, and near-surface meteorological conditions. The resulting framework produces daily precipitation estimates at a 0.1° spatial resolution over the TP. In addition to the best precipitation estimate, CRISP provides climatology-corrected precipitation, precipitation occurrence probability, calibrated prediction intervals, and precipitation-phase classifications. These additional variables provide information about the reliability and physical state of precipitation that is generally not available in conventional merged precipitation products.

Numerous global precipitation products based on satellite retrievals and atmospheric reanalysis have been developed to compensate for the limited gauge coverage over the TP (Lyu et al., 2024). Such datasets can provide broad spatial coverage and high temporal resolution, showing high potential for regional hydrological studies (Tong et al., 2014).

…

Our previous work introduced a double machine-learning framework for precipitation estimation over the TP and its subregions (Lyu & Yong, 2024, 2025). The framework combined an initial precipitation estimation step with a subsequent gauge-based error-correction step, thereby improving the representation of topographic and atmospheric effects on precipitation. Although this approach outperformed several benchmark products under specific conditions, including MSP and TPHiPr, it primarily focused on deterministic precipitation accuracy. CRISP extends this line of work in three ways. First, it integrates multiple precipitation, topographic, and atmospheric data sources within a unified stacking framework, allowing the model to use spatially continuous environmental information in areas with limited observations. Second, it applies distribution-free conformalized quantile regression to produce calibrated prediction intervals and precipitation probabilities rather than a single deterministic estimate. The intervals reflect the predictive uncertainty learned from the available predictors and are not

intended to explicitly decompose measurement errors, sampling biases, retrieval uncertainties, or other individual sources of uncertainty. Third, it includes a thermodynamic phase-discrimination module to classify precipitation events as dry, snow, mixed, or rain. The framework also accounts for measurement-related effects, particularly potential wind-induced undercatch, by incorporating atmospheric and surface predictors that influence gauge exposure and precipitation measurements. These predictors help the models learn systematic variations in catch efficiency associated with near-surface wind conditions, precipitation phase, and the surrounding surface environment, thereby reducing the potential bias in precipitation estimates caused by wind-driven undercatch.

…

Third, it incorporates a thermodynamic phase discrimination module to identify the physical state of precipitation. The CRISP dataset can meet the requirements of different users. The deterministic layers provide the best estimate (P_Best) and climatology-corrected precipitation (P_Clim) for generalized weather analysis and hydrological modeling.

…

The primary contribution of CRISP is therefore not the generation of completely independent precipitation information, but the integration of multiple complementary data sources within a unified framework and the provision of information that is not commonly available in existing products. In particular, CRISP provides precipitation amount together with calibrated uncertainty information, precipitation occurrence probability, and precipitation-phase classification. These variables are important for hydrological, cryospheric, and atmospheric applications, especially over regions where precipitation phase, measurement uncertainty, and unresolved small-scale processes strongly affect the interpretation of precipitation estimates.

Please see Section 2.3 in detail:

We acknowledge that GTS measurements, particularly under snowfall and complex terrain conditions, may also be affected by undercatch and representativeness errors; thus, they were not treated as an error-free reference.

Prior to serving as the training target, the raw gauge observations were corrected by the authors for potential wind-induced undercatch. The correction factor, referred to as the catch ratio (CR), was empirically calculated from concurrent near-surface air temperature (T, °C) and 10-m wind speed (WS). These variables were obtained from the NMIC station streams when available and were supplemented

with co-located ERA5 reanalysis data when station observations were missing. The dataset included 484 stations, although the number of available stations varied among years; approximately 30% of the station-day records required ERA5 supplementation. We also derived the gauge-height wind speed (WS_g, constrained to a minimum of 0.1m/s) via an empirical scaling factor of 0.7. The aerodynamic CR was dynamically calculated based on precipitation phase thresholds:

$$CR_{rain} = \exp(-0.03 \cdot WS_g), T > 2°C \quad (5)$$

$$CR_{snow} = \exp(-0.20 \cdot WS_g), T < -2°C \quad (6)$$

For mixed-phase precipitation (-2°C≤T≤2°C), *CR* was determined through linear interpolation:

$$CR_{mixed} = CR_{snow} + \frac{T-(-2)}{4} \cdot (CR_{rain} - CR_{snow}) \quad (7)$$

To prevent unrealistically large precipitation corrections under extreme wind conditions, a strict lower bound of 0.2 was enforced for all catch ratios. The actual precipitation mass (P_c) was then reconstructed as $P_c = P_{obs} / CR$. Subsequently, to align these point-based observations with the 0.1° gridded framework, a nearest-neighbor approach was employed to map each station to its corresponding grid cell. In instances where multiple stations fell within the same 0.1° grid cell, their corrected precipitation values were arithmetically averaged to generate a single, unique grid-level training target. The corrected gauge observations were then used as the training targets for the CRISP framework, allowing the model to learn precipitation patterns and residual uncertainties associated with solid precipitation in alpine environments.

Please see Section 4.2 in detail:

Ground-based precipitation gauges are useful references for precipitation evaluation, but they are not error-free observations. In particular, wind-induced undercatch can lead to substantial underestimation of solid and mixed precipitation, with the magnitude of the error depending on precipitation type, wind speed, gauge design, and site exposure. These issues are especially relevant over the Tibetan Plateau, where snowfall and strong winds are common. Following the discussions in Kochendorfer et al. (2017), gauge observations used in this study are therefore regarded as available observational references rather than unbiased measurements of true precipitation.

This limitation means that agreement with gauge observations alone cannot fully demonstrate the absolute accuracy of a precipitation product. The gauge-based results are consequently interpreted together with the comparisons against other precipitation products and the independent hydrological and

cryospheric consistency analyses. Thus, we compared it with AIMERG, AERA5-Asia and GMCP, in addition to IMERG, GSMaP, ERA5, and TPHiPr. The main characteristics of these datasets are summarized in Table 3. AIMERG, AERA5-Asia, and GMCP provide hourly or sub-hourly precipitation estimates at 0.1° resolution, whereas CRISP provides daily estimates over 1998–2024. The products also differ in their input datasets and calibration strategies. AIMERG and AERA5-Asia use APHRODITE-based information for regional calibration, whereas GMCP is produced through global multi-source precipitation merging and calibration. CRISP integrates satellite retrievals, reanalysis variables, station observations, and terrain information, and additionally provides precipitation-phase classification and prediction intervals.

To ensure a consistent comparison, all datasets were converted to daily precipitation totals using the same UTC-based daily definition and regridded to the 0.1° CRISP grid before evaluation. The spatial comparison was conducted within the geographic boundary of the Tibetan Plateau rather than over the full rectangular input domain. For each product, only grid cells with valid data within this common TP mask were included in the statistical analysis. Therefore, areas outside the TP boundary and regions not covered by a specific dataset were excluded from the corresponding evaluation rather than treated as zero precipitation. TPHiPr has a different effective spatial coverage near the southern margin of the study domain; this difference was considered by applying the common TP mask and by calculating the statistics only over the overlapping valid area.

Table 3. Recent developed precipitation datasets.

Dataset	Coverage	Resolution	Main calibration/source	Phase information	Uncertainty information
CRISP	1998–2024	0.1°; Daily	Satellite, reanalysis, gauges, terrain	Yes	Yes
AIMERG	2000–2015	0.1°; Half-hourly	IMERG calibrated with APHRODITE	No	No
AERA5-Asia	1951–2015	0.1°; Hourly	ERA5-Land constrained by APHRODITE	No	No
GMCP	2000–present	0.1°; Hourly	Global multi-source merging and calibration	No	No
TPHiPr	1979–2020	1/30°; Daily	ERA5_CNN and gauge observations	No	No

The higher temporal resolution of AIMERG, AERA5-Asia, and GMCP is advantageous for applications that depend on the timing and intensity of short-duration precipitation events, such as flash-flood forecasting, event-scale rainfall-runoff simulation, urban or small-catchment hydrological modeling, and sub-daily extreme precipitation analysis.

…

Table 4. Evaluation results of recent developed precipitation datasets during the period of 2011-2015.

Dataset	CC	RB (%)	RMSE (mm/day)	POD	FAR	CSI
CRISP	0.75	11.49	8.53	0.90	0.36	0.60
AIMERG	0.49	-10.75	11.31	0.78	0.40	0.51
AERA5-ASIA	0.46	2.66	11.55	0.86	0.49	0.46
GMCP	0.68	-0.37	9.71	0.89	0.45	0.52
TPHiPr	0.60	-4.45	12.86	0.79	0.50	0.43

Insert Table 4 near here

As shown in Table 4, CRISP achieved the best overall performance among the evaluated datasets. It obtained the highest correlation coefficient (CC = 0.75), the lowest RMSE (8.53 mm/day), and the highest CSI (0.60), indicating better agreement with the gauge observations and a more balanced representation of precipitation occurrence. CRISP also produced the highest POD (0.90) and the lowest FAR (0.36) among the evaluated datasets. GMCP showed the smallest relative bias (RB = −0.37%) and relatively high POD (0.89), but its higher FAR resulted in a lower CSI (0.52). AIMERG had a moderate CSI (0.51) and the second-lowest FAR, but its lower CC and negative bias reduced its overall performance. AERA5-Asia and TPHiPr showed comparatively larger errors, with lower CC and higher RMSE, FAR, or both. These results suggest that CRISP provided the most consistent daily precipitation estimates during 2011–2015, while its positive relative bias (11.49%) indicates that the improvement in event detection and overall agreement was accompanied by a tendency toward precipitation overestimation.

Please see Section 4.7 in detail:

Several limitations should be considered when interpreting the validation results. First, the gauge observations used as references may underestimate solid and mixed precipitation because of wind-induced undercatch. Second, the SWE data from Miao et al. (2024) also contain uncertainties, for which

a complete independent error characterization is not available in this study. Third, some precipitation products may share common input data or systematic errors. Accordingly, the results should not be interpreted as demonstrating absolute error-free performance. Instead, the combined evaluation indicates the relative consistency and practical value of CRISP across multiple precipitation, hydrological, and cryospheric constraints.

Please see Conclusion in detail:

Finally, CRISP provides a daily precipitation dataset that can be evaluated using complementary hydrological and cryospheric constraints. Through climatological calibration by CDF mapping, the dataset aligns with the regional climatological baseline. When evaluated against the physical constraints of the Third Pole precipitation paradox, CRISP produces estimates that are generally higher than the uncorrected gauge measurements and more consistent with the broad PET- and SWE-based plausibility constraints examined here, while avoiding the overestimations frequently observed in reanalysis models. By offering these key variables, the CRISP dataset provides observation-constrained precipitation estimates with uncertainty information for future cryospheric and hydrological research.

Please see Code and data availability in detail:

The CRISP precipitation dataset in NetCDF format is available at the National Tibetan Plateau Data Center, which can be accessed at https://doi.org/10.11888/Atmos.tpdc.303469 (Yong & Lyu, 2026). The codes used for producing this dataset are available from https://doi.org/10.5281/zenodo.19567996. The input data can be found in https://disc.gsfc.nasa.gov/datasets/GPM_3IMERGDF_07 ; https://www.ecmwf.int/en/forecasts/dataset/ecmwf-reanalysis-v5 ; http://sharaku.eorc.jaxa.jp/GSMaP/index.htm ; https://www.earthdata.nasa.gov/data/instruments/srtm. The gauge data is available at https://data.cma.cn/ai/#/detail?id=12. The PET and SWE data used in this study were obtained from Miao et al. (2024). These data were used to provide an additional cryospheric constraint for evaluating the temporal and spatial consistency of the precipitation estimates. The SWE dataset may contain uncertainties related to measurement error, spatial representativeness, and the processing procedures described in the original study.

****

ERA5 uncertainty for small-scale processes: A crucial part of this dataset is based on ERA5 but its uncertainties in both near-surface winds and precipitation are discussed in an insufficient manner.

I suggest first to add some references and a discussion of what the expected uncertainty in ERA5 near-surface winds is (e.g. Minola, L., Zhang, G., Ou, T., Kukulies, J., Curio, J., Guijarro, J. A., ... & Chen, D. (2024). Climatology of near-surface wind speed from observational, reanalysis and high-resolution regional climate model data over the Tibetan Plateau. Climate Dynamics, 62(2), 933-953). Second, the fact that ERA5 uses a convective parametrization is not discussed at all. A valuable addition would be 1) to quantify how much of the precipitation in ERA5 is sub-grid scale. The authors briefly talk about surface pressure as a proxy for vertical motion, but why do the authors choose to not include any variables that are more directly linked to convection, e.g. vertical velocity or any instability parameters such as CAPE and CIN?

Response:

Done. We have revised the manuscript in three ways. **First, we added a discussion of the uncertainty and spatial representativeness of ERA5 near-surface wind speed over the Tibetan Plateau, with reference to Minola et al. (2024).** We now clarify that ERA5 near-surface winds may contain substantial regional and seasonal biases due to the coarse model resolution, complex terrain, unresolved surface heterogeneity, and the difficulty of representing boundary-layer processes over high mountains. Therefore, ERA5 wind variables are used in CRISP as large-scale environmental predictors rather than as error-free local observations.

We also added a discussion of the uncertainty in ERA5 precipitation. ERA5 precipitation is produced by both the parameterized convective precipitation scheme and the resolved large-scale precipitation scheme. Neither component can fully represent the spatial intermittency and intensity of small-scale precipitation over complex terrain. In particular, parameterized convection may have limitations in terms of the timing, location, and intensity of convective precipitation. We therefore do not interpret ERA5 precipitation as a local precipitation reference, but as one of the large-scale information sources used by CRISP.

Second, following the reviewer's suggestion, we quantified the contribution of ERA5 parameterized convective precipitation to ERA5 total precipitation. Specifically, we separated ERA5 total precipitation into convective precipitation and large-scale precipitation and calculated the convective precipitation fraction as:

The spatial distribution of this fraction during 2011-2015 is presented in Fig. X. The unweighted mean of the grid-cell convective precipitation fractions is 59.11% over the valid grid cells within

the Tibetan Plateau mask. The contribution is spatially heterogeneous, with relatively high values over the western, southwestern, and central Tibetan Plateau and lower values in parts of the southern margin and eastern plateau. These results show that the parameterized convective component represents a substantial

fraction of ERA5 precipitation and should be considered when interpreting ERA5 precipitation-based predictors over the Tibetan Plateau.

Third, we have clarified the rationale for not initially including variables that are more directly related to convection, such as pressure-level vertical velocity, CAPE, and CIN. Surface pressure was selected as a relatively robust and widely available descriptor of synoptic-scale pressure variability and large-scale atmospheric forcing. By contrast, vertical velocity, CAPE, and CIN are more directly related to convective development but are also strongly affected by model resolution, parameterization, vertical interpolation, and the complex topography of the Tibetan Plateau. In addition, these variables may be highly correlated with other atmospheric predictors and may introduce redundancy into the model.

We did not add these variables to the final predictor set because the objective was to retain a parsimonious predictor configuration and avoid introducing additional model-dependent variables without demonstrating a clear improvement in predictive performance. We have added this limitation to the revised manuscript and identify a systematic assessment of convection-related predictors as an important direction for future work.

****

Please see Section 4.7 in detail:

Additional uncertainties are associated with the ERA5-based predictors used in CRISP. ERA5 near-surface wind speed may have regional and seasonal biases over the Tibetan Plateau because of its relatively coarse spatial resolution, complex topography, unresolved surface heterogeneity, and uncertainties in the representation of boundary-layer processes over high mountains (Minola et al., 2024). Therefore, ERA5 wind fields should be interpreted as estimates of large-scale environmental conditions rather than as error-free representations of local near-surface winds. This limitation is particularly relevant to solid and mixed precipitation, for which wind conditions can affect both precipitation-phase classification and the spatial redistribution of snowfall.

ERA5 precipitation also has limitations in representing small-scale precipitation processes. Total precipitation in ERA5 consists of large-scale precipitation and parameterized convective precipitation. The convective component may not accurately represent the timing, location, intensity, and spatial intermittency of localized convection over the complex terrain of the Tibetan Plateau. As described in Appendix C, we separated the two precipitation components and calculated the contribution of parameterized convective precipitation to total precipitation during 2011-2015. The unweighted spatial mean of the convective precipitation fraction was 59.11% over the valid grid cells, with substantial spatial heterogeneity across the Tibetan Plateau (Fig. A2). This result indicates that parameterized convection accounts for a substantial proportion of the ERA5 precipitation used as an input to CRISP. However, this value represents the mean of grid-cell precipitation fractions and should not be interpreted as an area-weighted regional precipitation ratio or as the observed fraction of actual sub-grid-scale convective precipitation. The result is affected by the ERA5 convection scheme, model resolution, and the representation of complex topography. Therefore, ERA5 precipitation should be regarded as a source of large-scale precipitation information rather than a local ground-truth reference.

The current predictor design also involves a trade-off between physical completeness and model simplicity. Variables such as pressure-level vertical velocity, convective available potential energy (CAPE), and convective inhibition (CIN) may provide more direct information about convective development. However, these variables are strongly affected by model resolution, parameterization, vertical interpolation, and the complex topography of the Tibetan Plateau. They may also be correlated with other atmospheric predictors and introduce redundancy into the model. For these reasons, they were not included in the final predictor set, which was designed to remain parsimonious and avoid adding model-dependent variables without clear evidence of improved predictive performance. Surface pressure is therefore used to represent synoptic-scale pressure variability and large-scale atmospheric forcing, rather than as a direct proxy for vertical motion. A systematic assessment of vertical velocity, CAPE, CIN, and other convection-related predictors will be considered in future work.

Overall, these limitations indicate that the performance of CRISP is partly constrained by the quality and physical representation of its input datasets. CRISP can correct and redistribute errors based on the available information, but it cannot fully recover precipitation processes that are absent from or poorly represented in the input fields. The results should therefore be interpreted as evidence of improved

consistency and practical utility, rather than as proof that all small-scale precipitation processes are explicitly resolved.

Please see Appendix C in detail:

Appendix C: Contribution of Parameterized Convective Precipitation in ERA5

ERA5 precipitation is generated by two physically distinct components: large-scale precipitation produced by the resolved or stratiform precipitation scheme, and convective precipitation produced by the parameterized convection scheme. The latter represents precipitation associated with convective processes that are not explicitly resolved at the spatial resolution of the reanalysis. Over the Tibetan Plateau, the representation of these processes is particularly challenging because of the complex topography, strong elevation gradients, heterogeneous land-surface conditions, and frequent interaction between large-scale circulation and local convection.

Therefore, ERA5 total precipitation should not be interpreted as a direct observation of local precipitation, especially at the sub-grid scale. The parameterized convective component may have uncertainties in its timing, location, intensity, and spatial intermittency. Nevertheless, separating the convective and total precipitation components provides useful information about the role of parameterized convection in the ERA5 precipitation input used by CRISP.

To quantify this contribution, we calculated the fraction of total ERA5 precipitation associated with the parameterized convective precipitation scheme over the Tibetan Plateau during 2011-2015. This analysis provides quantitative context for assessing the extent to which the ERA5 precipitation predictor depends on parameterized convection.

Daily ERA5 precipitation data for 2011-2015 were used, including total precipitation (tp) and convective precipitation (cp).

The accumulated precipitation for each grid cell was calculated as:

$$TP_{\mathrm{sum}}(i,j) = \sum_{t=1}^{N} T\,P(i,j,t) \tag{7}$$

$$CP_{\mathrm{sum}}(i,j) = \sum_{t=1}^{N} C\,P(i,j,t) \tag{8}$$

where $i$and $j$denote longitude and latitude grid cells, respectively, $t$denotes the daily time step, and $N$is the total number of daily observations during 2011-2015. ERA5 precipitation is provided in metres.

Since total and convective precipitation have the same units, no unit conversion is required for calculating their ratio.

The contribution of parameterized convective precipitation was then calculated as:

$$F_{\mathrm{conv}}(i,j) = 100 \times \frac{CP_{\mathrm{sum}}(i,j)}{TP_{\mathrm{sum}}(i,j)} \tag{9}$$

Here, F_{conv} represents the percentage contribution of the ERA5 parameterized convective precipitation component to total precipitation at each grid cell. It does not represent the observed fraction of actual sub-grid-scale convective precipitation.

Non-finite values were excluded from the calculation. Grid cells with non-positive accumulated total precipitation were treated as invalid. The resulting values were restricted to the physically meaningful range of 0-100%. The multi-year spatial mean was calculated as the arithmetic mean of all valid grid-cell values:

$$\bar{F}_{\mathrm{conv}} = \frac{1}{M}\sum_{k=1}^{M} F_{\mathrm{conv},k} \tag{10}$$

where M is the number of valid grid cells. Thus, the reported spatial mean is an unweighted grid-cell mean. The corresponding grid-summed precipitation ratio would instead be calculated as:

$$F_{\mathrm{regional}} = 100 \times \frac{\sum C\,P_{\mathrm{sum}}}{\sum T\,P_{\mathrm{sum}}} \tag{11}$$

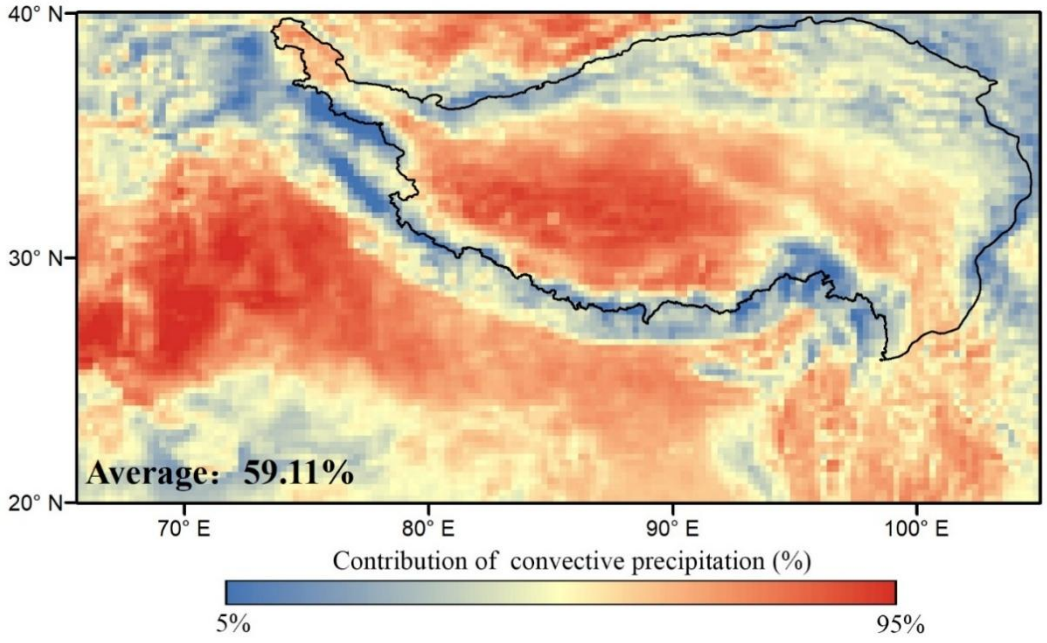

Figure A2. Spatial distribution of the contribution of parameterized convective precipitation to total precipitation in ERA5 over the Tibetan Plateau during 2011-2015. The black line denotes the

boundary of the Tibetan Plateau. The inset value denotes the unweighted spatial mean over all valid grid cells within the plotted domain (59.11%).

The spatial distribution of the parameterized convective precipitation fraction in ERA5 is shown in Figure A2. The unweighted mean of the grid-cell convective precipitation fractions is 59.11% over the valid grid cells, indicating that the parameterized convective component accounts for a substantial fraction of the ERA5 precipitation in the study region. The contribution is spatially heterogeneous, with relatively high values over the western, southwestern, and central parts of the Tibetan Plateau and lower values in parts of the southern margin and eastern plateau. This pattern should be interpreted as the spatial distribution of the ERA5 parameterized convective precipitation component rather than as a direct observation of the actual sub-grid-scale convective precipitation. The result is affected by the ERA5 convection scheme, model resolution, and the representation of complex Tibetan Plateau topography. It nevertheless provides useful quantitative context for assessing the role of parameterized convection in the ERA5 precipitation input used by CRISP.

****

Comment on hourly data: Can you comment on what applications might need hourly data and if your framework could somehow be used to correct hourly precipitation estimates in the future?

Response:

Done. We added a discussion of the application value of hourly precipitation data. Hourly or sub-hourly data are particularly useful for flash-flood forecasting, event-scale rainfall-runoff simulation, small-catchment hydrological modeling, short-duration extreme precipitation analysis, and the representation of snow accumulation and melt timing. However, the higher temporal resolution of these datasets may also increase sensitivity to satellite sampling errors, sparse gauge calibration, precipitation intermittency, and phase-dependent measurement errors over the Tibetan Plateau.

We also clarified that an hourly extension of CRISP is technically feasible because the current framework already uses hourly atmospheric variables and high-frequency satellite information. However, reliable hourly production would require hourly training observations, an hourly-specific calibration strategy, and independent validation of precipitation timing, intensity, occurrence, and

phase. Direct disaggregation of daily estimates would not adequately represent short-duration precipitation events. We therefore identify hourly CRISP development as a future direction rather than claiming that the current daily product can directly provide validated hourly estimates.

****

Please see Section 4.2 in detail:

The higher temporal resolution of AIMERG, AERA5-Asia, and GMCP is advantageous for applications that depend on the timing and intensity of short-duration precipitation events, such as flash-flood forecasting, event-scale rainfall-runoff simulation, urban or small-catchment hydrological modeling, and sub-daily extreme precipitation analysis. Hourly precipitation information may also improve the representation of snow accumulation and melt timing in cryospheric models, particularly when combined with an energy-balance model or an hourly snowpack model. However, higher temporal resolution does not necessarily ensure higher accuracy or temporal stability over the Tibetan Plateau. At hourly and sub-hourly scales, the estimates are more sensitive to satellite sampling limitations, retrieval errors, sparse gauge calibration, precipitation intermittency, and wind-induced undercatch during snowfall. These uncertainties are particularly important for weak, solid, and mixed-phase precipitation events.

The current daily resolution of CRISP was selected because the primary training observations are daily precipitation records, while reliable and spatially extensive hourly observations remain limited over the Third Pole. Daily aggregation provides a relatively stable basis for multi-source calibration and is appropriate for long-term climatological analysis, daily water-balance assessment, snow accumulation, glacier mass-balance studies, and daily or sub-daily hydrological applications in which daily precipitation is used as the forcing input. The CRISP framework could be extended to hourly resolution because it already uses hourly atmospheric forcing and high-frequency satellite information. Nevertheless, an hourly version would require hourly training observations, an hourly-specific bias-correction strategy, and independent validation of precipitation occurrence, timing, intensity, and phase. Direct temporal disaggregation of daily CRISP estimates would not adequately reproduce the actual timing or short-duration intensity of precipitation events. Therefore, an hourly CRISP product is technically feasible but is left for future development.

****

Validation with independent stations: Throughout the manuscript, the new precipitation estimates

are validated against a set of independent observations which I believe belong to the red circle stations in Fig. 2. Can the authors comment on the expected deviations from those measurements compared to the rest of the station network given that the red stations seem to represent a quite different environment and terrain compared to all blue stations?

Response:

We thank the reviewer for raising this important point. **The independent stations shown in red in Figure 2 differ from the blue meteorological stations in their deployment purpose, instrumentation, temporal coverage, and environmental representation.** Most of the independent stations were established by hydrological agencies to monitor key catchments or specific river sections, whereas the meteorological stations were primarily designed for routine weather and climate monitoring. The independent network therefore samples some high-elevation, remote, and topographically complex environments that are relatively underrepresented by the meteorological station network.

The instrumentation and observation periods also differ between the two networks. For example, the hydrological stations commonly use HOBO tipping-bucket rain gauges with a 0.2 mm resolution, whereas the meteorological stations generally use gauges with a 0.1 mm resolution. In addition, many hydrological stations operate only during the flood or warm season, while the meteorological stations generally provide more continuous observations. The specialized 53-station high-altitude network also consists of spatial transects equipped with tipping-bucket rain gauges with a 0.2 mm resolution and mainly provides warm-season hourly observations from 2017 to 2020. Given these differences, we expect the deviations at the independent stations to differ from those obtained for the meteorological network. In particular, the more complex terrain and higher elevations represented by the independent stations can increase grid-to-point representativeness errors because local orographic effects may not be fully resolved by the gridded predictors and precipitation estimates. Differences in gauge resolution, site exposure, wind-induced undercatch, and temporal coverage may introduce additional discrepancies. Because most hydrological observations are restricted to the warm or flood season, their validation statistics also characterize a different precipitation regime and should not be interpreted as directly equivalent to annual statistics from the meteorological network.

Nevertheless, this distinction is also the main purpose of the independent validation. These

observations were completely excluded from model training and provide a more stringent assessment of the spatial and environmental transferability of the precipitation estimates to locations and terrain conditions that are not well represented by the training network. We have clarified the deployment purpose, instrumentation, temporal coverage, and representativeness of the independent stations in the revised data description.

****

Please see Section 2.3 in detail:

For independent validation, we incorporated supplementary datasets that were completely excluded from model training because of their limited temporal coverage. These include hydrological observations from the Yellow River Conservancy Commission (Lyu & Yong, 2025) and observations from a specialized 53-station high-altitude network (Zhan et al., 2022). The independent stations differ from the meteorological stations in their deployment purpose, instrumentation, temporal coverage, and environmental representation. The hydrological stations were primarily established to monitor key catchments or specific river sections rather than for routine meteorological monitoring, and they generally operate only during the flood or warm season. The specialized high-altitude network consists of spatial transects equipped with tipping-bucket rain gauges with a 0.2 mm resolution and provides warm-season hourly observations from 2017 to 2020. Hydrological observations also commonly have a 0.2 mm gauge resolution, whereas the meteorological stations generally use gauges with a 0.1 mm resolution. The locations of the independent stations are shown in Figure 2. These stations sample high-elevation, remote, and topographically complex environments that are relatively underrepresented by the meteorological network. Consequently, their validation statistics are not expected to be directly equivalent to those derived from the meteorological stations. Instead, they provide an independent and relatively stringent assessment of model transferability beyond the environments represented by the training network.

****

Comparison between datasets: It is not entirely clear if for the validation part of the paper, the datasets are compared at the same resolution (have they been all regridded to the same grid?) or not. Similarly, Fig. 4 shows a slightly different domain covering less of the wet parts of Northern India for TPHiPr compared to the other sets. Has this been taken into account in the analyses that

follow?

Response:

Thank you for pointing this out. We have clarified the procedures used for inter-product evaluation. All hourly and half-hourly datasets were accumulated into daily precipitation totals using the same UTC-based daily definition. **The datasets were then transformed to the common 0.1° CRISP grid before comparison. TPHiPr was converted from its native 1/30° grid to the target grid, while ERA5 was transformed from its native 0.25° grid** using the bilinear interpolation method. All products were clipped using the same Tibetan Plateau boundary, and the statistical analyses used only valid overlapping grid cells. Missing regions outside the effective coverage of a product, including the southern-margin coverage difference of TPHiPr, were excluded from the calculations and were not treated as zero precipitation. The caption of Figure 4 has been revised to make this treatment explicit.

****

Please see Section 2.3 in detail:

Following the completion of the validation protocols, the CRISP framework was re-trained utilizing the full integrated long-term observational dataset to generate the final production version. This comprehensive dataset was employed for the 27-year multidimensional data cube and served as the foundation for the subsequent independent station evaluations and physical consistency tests.

For the inter-product evaluation, all precipitation datasets were first converted to a common temporal and spatial framework. Hourly and half-hourly precipitation estimates were accumulated into daily totals using UTC days. Products with a native resolution of 0.1° were retained at their original grid, whereas products with finer or coarser spatial resolution were transformed to the 0.1° CRISP grid. TPHiPr was spatially aggregated from 1/30° to 0.1°, and ERA5 was interpolated from 0.25° to the target grid using the bilinear interpolation method. After spatial transformation, all datasets were clipped using the same geographic boundary of the Tibetan Plateau. Statistical metrics were calculated only for grid cells and dates with valid values in the compared datasets. Missing areas outside a product's valid coverage were excluded from the calculation and were not assigned zero precipitation.

Please see Figure 4 in detail:

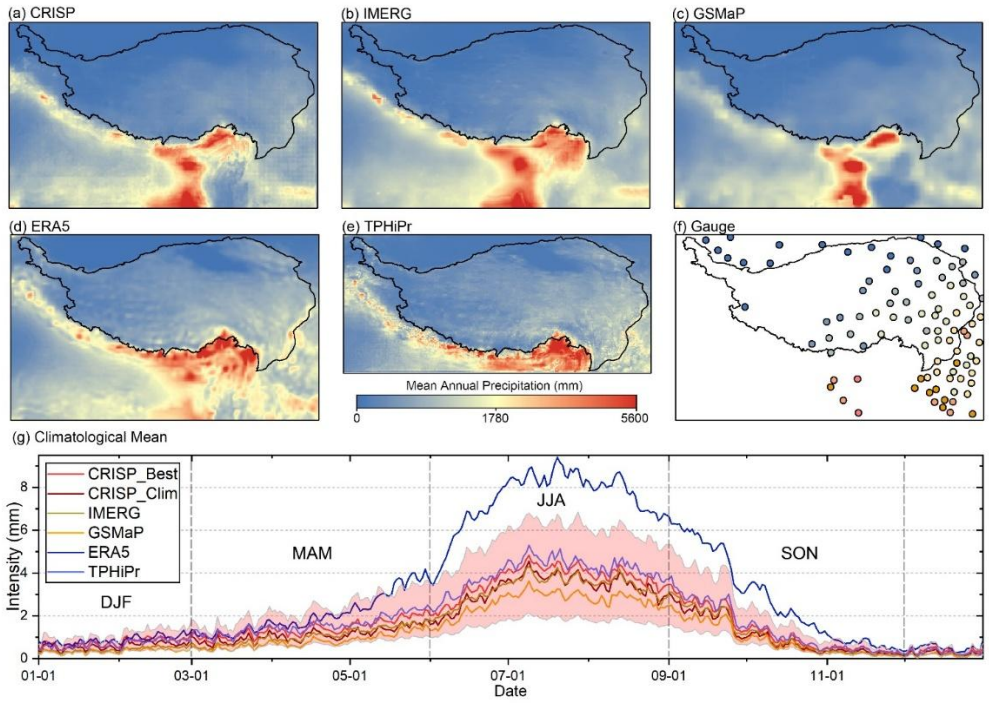

Figure 4. Spatiotemporal distribution of mean annual precipitation and daily climatological precipitation across the Tibetan Plateau during 1998–2024. (a–f) show the spatial distribution of multi-year mean annual precipitation from CRISP, IMERG, GSMaP, ERA5, TPHiPr, and NMIC Gauge, respectively. TPHiPr is shown for 1998–2020 and only within its valid coverage area. **(g)** presents the climatological mean daily precipitation over the annual cycle. Solid lines denote the deterministic best estimates, and the red shaded area indicates the dynamically estimated 5th–95th percentile interval. All datasets were regridded to a common grid before comparison, and dataset-specific missing values were excluded using the corresponding valid-data masks.

****

Detailed comments

The machine learning method is explained quite late in the introduction and it does not immediately become clear how the authors utilize the input datasets to provide novel information. I suggest revising the introduction and moving a simple explanation of the developed framework relatively close to the start of the paper.

Response:

Done.

****

Please see Introduction in detail:

The Tibetan Plateau (TP), widely known as the “Third Pole” and the “Asian Water Tower”, sustains the livelihoods of nearly two billion people through its wide river systems (Shukla & Sen, 2021; Yao et al., 2022). Because of its extreme elevation and strong land-atmosphere coupling, the TP is highly sensitive to global climate change (You et al., 2021). This sensitivity is evident in rapid glacier retreat, expanding alpine lakes, and pronounced shifts in the regional water balance (Zhang et al., 2015; Su et al., 2022). Understanding these hydroclimatic transformations requires accurate and long-term precipitation estimates, as precipitation is the driver of the energy and water cycles on the TP (Zhang et al., 2017). However, extreme topography, harsh climatic conditions, and the sparse distribution of in situ observations make reliable precipitation quantification a persistent challenge (Gao & Liu, 2013).

To address this challenge, we develop the Conformal Regression and Integrated Stacking for Precipitation (CRISP) framework. CRISP uses a machine-learning-based data integration strategy to combine complementary information from ground-based precipitation observations, satellite and reanalysis precipitation products, static topographic variables, and dynamic atmospheric predictors. The observation-based data provide local precipitation constraints, while satellite and reanalysis products provide spatially continuous information in areas with limited gauge coverage. Topographic and atmospheric variables are used to describe the spatial and temporal controls on precipitation, including the effects of elevation, terrain structure, atmospheric circulation, and near-surface meteorological conditions. The resulting framework produces daily precipitation estimates at a 0.1° spatial resolution over the TP. In addition to the best precipitation estimate, CRISP provides climatology-corrected precipitation, precipitation occurrence probability, calibrated prediction intervals, and precipitation-phase classifications. These additional variables provide information about the reliability and physical state of precipitation that is generally not available in conventional merged precipitation products.

Numerous global precipitation products based on satellite retrievals and atmospheric reanalysis have been developed to compensate for the limited gauge coverage over the TP (Lyu et al., 2024). Such datasets can provide broad spatial coverage and high temporal resolution, showing high potential for regional hydrological studies (Tong et al., 2014). While the reliability of these products is limited by sensor sensitivity and retrieval algorithms, a more critical constraint lies in their heavy dependence on ground-based calibration (Kidd & Levizzani, 2011). These reference observations suffer from systematic

biases, as the sparse measurement network fails to capture the precipitation variability over the TP (Tang et al., 2018).

Regional precipitation datasets have subsequently been developed by integrating available gauge observations with high-resolution atmospheric simulations or gridded background fields. For example, the Tibetan Plateau High-resolution Precipitation (TPHiPr) dataset improves precipitation estimates through dense observation-based correction (Jiang et al., 2023). Several recent developed products, including AIMERG, AERA5-Asia, and GMCP, have further improved precipitation estimation over Asia and the TP (Ma et al., 2020, 2022, 2025). These products provide enhanced temporal resolution and regional coverage, but they differ in their calibration strategies and in whether they provide information on precipitation phase and estimation uncertainty. Moreover, many regional products rely on global reanalysis datasets as background fields. This dependence can propagate systematic errors from the parent models, including excessive light-precipitation frequency and insufficient representation of extreme rainfall intensity (Ward et al., 2011).

Our previous work introduced a double machine-learning framework for precipitation estimation over the TP and its subregions (Lyu & Yong, 2024, 2025). The framework combined an initial wet or dry estimation step with a subsequent gauge-based error-correction step, thereby improving the representation of topographic and atmospheric effects on precipitation. Although this approach outperformed several benchmark products under specific conditions, including MSP and TPHiPr, it primarily focused on deterministic precipitation accuracy. CRISP extends this line of work in three ways. First, it integrates multiple precipitation, topographic, and atmospheric data sources within a unified stacking framework, allowing the model to use spatially continuous environmental information in areas with limited observations. Second, it applies distribution-free conformalized quantile regression to produce calibrated prediction intervals and precipitation probabilities rather than a single deterministic estimate. The intervals reflect the predictive uncertainty learned from the available predictors and are not intended to explicitly decompose measurement errors, sampling biases, retrieval uncertainties, or other individual sources of uncertainty. Third, it includes a thermodynamic phase-discrimination module to classify precipitation events as dry, snow, mixed, or rain. The framework also accounts for measurement-related effects, particularly potential wind-induced undercatch, by incorporating atmospheric and surface predictors that influence gauge exposure and precipitation measurements. These predictors help the models learn systematic variations in catch efficiency associated with near-surface wind conditions,

precipitation phase, and the surrounding surface environment, thereby reducing the potential bias in precipitation estimates caused by wind-driven undercatch.

As outlined in Figure 1, CRISP is provided as a multidimensional NetCDF4 data cube and extends conventional statistical merging through three main components. First, it employs physics-aware corrections by integrating multi-source data with static topography and dynamic atmospheric variables, compensating for wind-induced undercatch in complex terrain. Second, it utilizes distribution-free conformalized quantile regression to generate uncertainty bounds. Third, it incorporates a thermodynamic phase discrimination module to identify the physical state of precipitation. The CRISP dataset can meet the requirements of different users. The deterministic layers provide the best estimate (P_Best) and climatology-corrected precipitation (P_Clim) for generalized weather analysis and hydrological modeling. The probabilistic layers offer precipitation probability (Prob) and calibrated 5th and 95th prediction intervals (P_5th, P_95th) for extreme event detection and risk assessment. Finally, the physical state layer (Phase) classifies events into dry, snow, mixed, or rain conditions to support cryospheric studies.

Insert Figure 1 near here

The primary contribution of CRISP is therefore not the generation of completely independent precipitation information, but the integration of multiple complementary data sources within a unified framework and the provision of information that is not commonly available in existing products. In particular, CRISP provides precipitation amount together with calibrated uncertainty information, precipitation occurrence probability, and precipitation-phase classification. These variables are important for hydrological, cryospheric, and atmospheric applications, especially over regions where precipitation phase, measurement uncertainty, and unresolved small-scale processes strongly affect the interpretation of precipitation estimates.

****

L. 45: What exactly was improved in the double machine learning approach and why was it called a double machine learning approach? This information would be useful to have here.

Response:

Done.

****

Please see Introduction in detail:

Our previous work introduced a double machine-learning framework for precipitation estimation over the TP and its subregions (Lyu & Yong, 2024, 2025). The framework combined an initial wet or dry estimation step with a subsequent gauge-based error-correction step, thereby improving the representation of topographic and atmospheric effects on precipitation. Although this approach outperformed several benchmark products under specific conditions, including MSP and TPHiPr, it primarily focused on deterministic precipitation accuracy.

****

L. 54: I suggest to be a bit more conservative in the language and not overemphasize the novelty of the used data formats and technologies. For example, I believe "comprehensive NetCDF4 data cube" simply refers to NetCDF4 files with multiple dimensions which has been the state-of-the-art data format for climate records for a long time. In fact, new data formats such as zarr stores that are equally practical but easier to compress have emerged. Have the authors considered storing the data as zarr files?

Response:

Done. We agree that NetCDF4 is a well-established format for multidimensional climate data and have revised the text to avoid overstating its novelty. We selected NetCDF4 because of its broad compatibility with commonly used climate-data tools. We acknowledge that Zarr is also a practical alternative, particularly for cloud-based storage and access, and will consider it in future data releases.

****

Please see Introduction in detail:

As outlined in Figure 1, CRISP is provided as a multidimensional NetCDF4 data cube and extends conventional statistical merging through three main components. First, it employs physics-aware corrections by integrating multi-source data with static topography and dynamic atmospheric variables, compensating for wind-induced undercatch in complex terrain.

****

Is the ERA5 regridded onto the same grid?

Response:

Yes. ERA5 was regridded to the same spatial grid as the other datasets before being used in the analysis.

Fig. 4: The colors don't match the legend so it is hard to map those (please correct that), but is it right that iMERG and CRISP_Clim are the two curves that are very well aligned? Then what bias is really reduced here?

Response:

Thank you for pointing this out. We rechecked and revised the color assignments and legend, and increased the line widths to improve distinguishability. However, reduced figure resolution may have made the curves difficult to distinguish, so we increased the line widths in the revised figure. IMERG and CRISP_Clim have similar mean values, but their local variations differ substantially: **IMERG shows more smoothed peak precipitation, whereas CRISP_Clim better preserves sharper peaks.** The apparent agreement in the mean therefore does not imply identical precipitation characteristics. In addition, IMERG-Final includes climatological adjustment, so its similarity to CRISP_Clim in the mean is not unexpected. The bias reduction discussed here refers to the improved agreement of CRISP with gauge observations, particularly in the representation of precipitation variability and extremes.

L. 47: Could you please clarify what kind of uncertainty you refer to when you discuss the "uncertainty interval"? It would be helpful to add a brief discussion here what the sources of different types of uncertainties are (e.g. uncertainties in the measurements, sampling biases, model/retrieval uncertainties, etc)

Response:

Done. The intervals reflect the predictive uncertainty learned from the available predictors and are not intended to explicitly decompose measurement errors, sampling biases, retrieval uncertainties, or other individual sources of uncertainty.

****

Please see Introduction in detail:

First, it integrates multiple precipitation, topographic, and atmospheric data sources within a unified stacking framework, allowing the model to use spatially continuous environmental information in areas with limited observations. Second, it applies distribution-free conformalized quantile regression to produce calibrated prediction intervals and precipitation probabilities rather than a single deterministic estimate. The intervals reflect the predictive uncertainty learned from the available predictors and are not intended to explicitly decompose measurement errors, sampling biases, retrieval uncertainties, or other individual sources of uncertainty. Third, it includes a thermodynamic phase-discrimination module to classify precipitation events as dry, snow, mixed, or rain

****

L. 49: Since much of this paper is about wind-induced undercatch, I think it would be worthwhile to explain in more detail here.

Response:

Done.

****

Please see Introduction in detail:

Third, it includes a thermodynamic phase-discrimination module to classify precipitation events as dry, snow, mixed, or rain. The framework also accounts for measurement-related effects, particularly potential wind-induced undercatch, by incorporating atmospheric and surface predictors that influence gauge exposure and precipitation measurements. These predictors help the models learn systematic variations in catch efficiency associated with near-surface wind conditions, precipitation phase, and the surrounding surface environment, thereby reducing the potential bias in precipitation estimates caused by wind-driven undercatch.

****

L. 58: can can meet → can meet

Response:

Done.

L. 78: Do you mean that you use an additional calibration based on the radar and microwave

measurements of GPM or do you simply mean that IMERG includes these measurements? This should be clarified because the sentence sounds like the existing retrievals are further enhanced.

Response:

We did not perform an additional calibration using DPR or GMI. The sentence referred to calibration and retrieval procedures already implemented within the original IMERG and GSMaP production systems. We have revised the text to avoid implying that CRISP applies a separate DPR/GMI calibration.

****

Please see Section 2.1 in detail:

The accuracy of these satellite retrievals is enhanced through calibration using Dual-frequency Precipitation Radar (DPR) and GPM Microwave Imager (GMI) data (Tang et al., 2020).

****

L. 83: Please correct this statement as GSMap does not solely rely on DPR measurements, but also includes the information from multiple satellite sensors.

Response:

We agree that GSMaP incorporates information from multiple satellite sensors and does not rely solely on DPR measurements. However, our statement refers specifically to the calibration standard for its microwave retrievals, rather than to all of the satellite observations used by GSMaP.

****

Please see Section 2.1 in detail:

Although GSMaP has a spatial resolution comparable to that of IMERG, it uses different retrieval and calibration procedures and employs DPR precipitation estimates as the calibration standard for **its passive microwave retrievals**, while integrating observations from multiple satellite sensors (Kubota et al., 2020).

****

Related to the comment above, it would be useful to clearly state what the difference between GSMap and IMERG is and why it is practical to combine those two datasets. It should also be clarified to what extent both datasets are built on the same information which means that the satellite

observations are given more weight in the created precipitation product. How would the authors expect results to change if only one of these two datasets were included?

Response:

Thank you for this important comment. **We have clarified the differences between IMERG and GSMaP and the rationale for using both products.** Although the two products use partly overlapping satellite observations, they employ different retrieval and calibration procedures and therefore have distinct error characteristics. **Our previous study (Lyu et al., 2024) compared their performance over the Tibetan Plateau and showed that each product has its own strengths and limitations**, with neither consistently outperforming the other. Thus, using both products as separate predictors allows the framework to exploit their complementary information. Using only one product would retain much of the shared satellite-derived precipitation signal but could lose information associated with the different retrieval characteristics of the other product.

****

Please see Section 2.1 in detail:

Although GSMaP has a spatial resolution comparable to that of IMERG, it uses different retrieval and calibration procedures and employs DPR precipitation estimates as the calibration standard for its passive microwave retrievals, while integrating observations from multiple satellite sensors (Kubota et al., 2020). Consequently, IMERG and GSMaP are not fully independent products, but they can provide complementary information because of their different algorithms and error characteristics. This complementarity is particularly relevant over the Tibetan Plateau, where each product has shown distinct strengths and limitations (Lyu et al., 2024). We use both products as separate predictors to take advantage of this complementary information.

****

L. 84: There are two Kubota references: Kubota et al., 2017 and Kubota et al., 2020 but only Kubota et al., 2020 exist in the reference list. I believe the year is wrong and the authors intention was to refer to Kubota et al., 2020 in the instances where it is mistakenly written Kubota et al., 2017.

Response:

Done.

****

Please see Section 2.1 in detail:

Although GSMaP has a spatial resolution comparable to that of IMERG, it uses different retrieval and calibration procedures and employs DPR precipitation estimates as the calibration standard for its passive microwave retrievals, while integrating observations from multiple satellite sensors (Kubota et al., 2020).

****

L. 121: Can you clarify here which model is trained with which dataset and how the different data are combined? Do you compute daily values for the satellite and ERA5 data first? Is ERA5 regridded? Are there any other preprocessing steps of the existing datasets at this point that are not made clear?

Response:

Thank you for raising this point. Lines 121 provide a brief overview of the datasets used in this study, whereas the detailed procedures for model–dataset correspondence, temporal aggregation, spatial regridding, and other preprocessing steps are described in Section 3.1. To make this organization clearer, **we have added a cross-reference to Section 3.1 at this point and revised the text** to clarify that the subsequent section provides the complete data-processing workflow.

****

Please see Section 2.1 in detail:

To capture the spatiotemporal pattern of precipitation across the TP, we integrated state-of-the-art satellite retrievals and atmospheric reanalysis. The dynamic input datasets were subsequently preprocessed to ensure consistency in temporal and spatial dimensions, with the detailed procedures described in Section 3.1. For the satellite-based inputs, we selected two high-resolution, gauge-adjusted products at a 0.1° daily resolution: the Integrated Multi-satellitE Retrievals for GPM (IMERG) Final Run V07 (Huffman et al., 2015) and the Global Satellite Mapping of Precipitation (GSMaP) gauge-adjusted V08 (Kubota et al., 2020).

****

L. 130: Can you clarify what the added value of the GTS dataset is given that both IMERG and GSMaP already contain the station data? It is stated here that "traditional large-scale satellite calibration algorithms frequently neglect local wind-induced solid precipitation undercatch and

complex subgrid orographic lifting". This is more of an explanation why it is useful to include the information of gauge measurements in satellite retrievals and not an explanation why the GTS dataset alone adds important information. Moreover, snow undercatch and potential missing precipitation associated with complex orographic lifting over small spatial scales are also inherent uncertainties of gauge measurements rather than biases that are introduced through in algorithms for satellite retrievals like IMERG? Why would the station data used for IMERG exhibit a different degree of undercatch than the additional stations used?

Response:

Done. We have revised this statement. Although IMERG and GSMaP use gauge observations in their calibration procedures, the GTS dataset provides additional station-based observations that are not necessarily identical to those used in the satellite products. These additional observations offer an independent and more spatially detailed reference for model training, while also preserving the local precipitation information that may be smoothed or modified during the large-scale calibration and merging processes. We acknowledge that GTS observations may also suffer from snowfall undercatch and representativeness errors; therefore, they are used as an additional observational constraint rather than as an error-free reference.

****

Please see Section 2.3 in detail:

To ensure data integrity, a stringent quality control (QC) protocol was implemented prior to model training. Consequently, as illustrated in Figure 2b, the number of active and highly reliable daily station records available for training dynamically fluctuated over the study period, expanding from 40,100 in 2001 to over 100,000 in recent years. Although the gauge-adjusted satellite products evaluated in this study (e.g., IMERG calibrated via GPCC and GSMaP adjusted via CPC) also incorporate gauge observations, the GTS dataset provides additional station-based measurements with local information that may not be fully retained after large-scale satellite calibration and merging. These observations were therefore included as an additional, spatially detailed constraint for model training. We acknowledge that GTS measurements, particularly under snowfall and complex terrain conditions, may also be affected by undercatch and representativeness errors; thus, they were not treated as an error-free reference.

****

L. 133 - 149: As I understand it, this is a catch ratio added to the dataset by the authors to quantify snow undercatch.

Response:

Thank you for pointing this out. We clarify that the catch ratio (CR) is not an additional observational dataset. **It is an empirical correction factor calculated by the authors using the observed temperature and wind speed to estimate and correct wind-induced precipitation undercatch in the gauge observations before model training.** The corrected precipitation was then used as the training target. We have revised the text to make this procedure clearer.

****

Please see Section 2.3 in detail:

Prior to serving as the training target, the raw gauge observations were corrected by the authors for potential wind-induced undercatch. The correction factor, referred to as the catch ratio (CR), was empirically calculated from concurrent near-surface air temperature (T, °C) and 10-m wind speed (WS). These variables were obtained from the NMIC station streams when available and were supplemented with co-located ERA5 reanalysis data when station observations were missing. The dataset included 484 stations, although the number of available stations varied among years; approximately 30% of the station-day records required ERA5 supplementation. We also derived the gauge-height wind speed (WS_g, constrained to a minimum of 0.1m/s) via an empirical scaling factor of 0.7. The aerodynamic CR was dynamically calculated based on precipitation phase thresholds:

$$CR_{rain} = \exp\left(-0.03 \cdot WS_g\right), T > 2°C \quad (5)$$

$$CR_{snow} = \exp(-0.20 \cdot WS_g), T < -2°C \quad (6)$$

For mixed-phase precipitation (-2°C≤T≤2°C), *CR* was determined through linear interpolation:

$$CR_{mixed} = CR_{snow} + \frac{T-(-2)}{4} \cdot (CR_{rain} - CR_{snow}) \quad (7)$$

To prevent unrealistically large precipitation corrections under extreme wind conditions, a strict lower bound of 0.2 was enforced for all catch ratios. The actual precipitation mass (P_c) was then reconstructed as $P_c = P_{obs} / CR$. Subsequently, to align these point-based observations with the 0.1° gridded framework, a nearest-neighbor approach was employed to map each station to its corresponding grid cell. In instances where multiple stations fell within the same 0.1° grid cell, their corrected precipitation values were arithmetically averaged to generate a single, unique grid-level training target. The corrected gauge

observations were then used as the training targets for the CRISP framework, allowing the model to learn precipitation patterns and residual uncertainties associated with solid precipitation in alpine environments.

****

L. 136: How often is the missing data replaced with ERA5 data? An appropriate addition to the supplementary material would be 1) a quantification of which portion of the dataset has this step based on ERA5 data and 2) how well the ERA5 data compares to the point measurements of the stations for these variables.

Response:

Thank you for this helpful suggestion. We have clarified the procedure used to supplement missing meteorological observations and quantified its approximate extent. The dataset included approximately 450 stations, although station availability varied among years, and approximately 30% of the station-day records required ERA5 supplementation. **We also added a comparison between ERA5 and concurrent station observations over the Tibetan Plateau using 2011-2015 data as a case study.** We did not include it in the supplementary material because the relevant results are similar to the currently known assessment results. The comparison reports the number of paired records, CC, and RMSE for 2-m air temperature and 10-m wind speed. **The results show substantially better agreement for air temperature than for wind speed, which is consistent with the known difficulty of gridded reanalysis products in representing small-scale wind variability over the complex terrain of the Tibetan Plateau.**

Table X. The evaluation of 2-m air temperature and 10-m wind speed between ERA5 and gauge observations in 2011-2015

Variable	**N**	**CC**	**RMSE**
2-m air temperature	585,821	0.76	2.81 °C
10-m wind speed	513,912	0.53	2.57 m/s

L. 138: What exactly are the precipitation thresholds?

Response:

These are precipitation-phase thresholds based on near-surface air temperature, rather than precipitation-amount thresholds, as shown in the formula:

$$CR_{rain} = \exp(-0.03 \cdot WS_g), T > 2°C$$
$$CR_{snow} = \exp(-0.20 \cdot WS_g), T < -2°C$$

L. 143: Replace the term "numerical instability" as this usually implies an instability in numerical modeling due to the small numerical errors growing uncontrollably during calculations. I think you mean something else in this simple equation.

Response:

Done.

****

Please see Section 2.3 in detail:

To prevent unrealistically large precipitation corrections under extreme wind conditions, a strict lower bound of 0.2 was enforced for all catch ratios.

****

L. 150: Which figures show this independent validation?

Response:

Figure 11.

L. 186: I would be slightly more conservative with presenting the framework as a hybrid machine learning/physics-based approach. The machine learning algorithm is trained on physical data and you create additional variables based on thresholds but I think most people would expect a more advanced construct of combining numerical physical models with AI when we talk about "hybrid machine learning".

Response:

Done.

****

Please see Section 3 in detail:

3 Data processing and methodology

The CRISP dataset was developed using a physics-guided machine-learning framework that

incorporates physical information into the construction of training targets, feature engineering, and post-processing, rather than coupling a numerical physical model with a machine-learning algorithm. As illustrated in Figure 3, the framework consists of three sequential modules: spatiotemporal alignment and feature engineering, the CRISP machine-learning engine, and physics-guided post-processing. Physical information is introduced through the use of gauge observations corrected for wind-induced undercatch, the derivation of physically relevant variables and phase-related features, and the application of physically constrained post-processing procedures.

****

L. 277: Can you clarify here what the independent data for hydrometeor phase is and how it compares to satellite-retrieved hydrometeor phase (from IMERG) as well as the hydrometeor classification in ERA5 (from the microphysics scheme)?

Response:

Done. The primary phase observations used for validation are station-based Present Weather codes from the NMIC network, rather than an entirely separate observational network. We have revised the manuscript accordingly and now distinguish between the NMIC Present Weather observations, the IMERG satellite-retrieved phase product, and the ERA5 hydrometeor classification. The latter two were not treated as independent observational references because IMERG contains satellite-derived information related to the CRISP inputs, while ERA5 hydrometeor phase is a model-based diagnostic generated by its microphysics scheme. **In addition, we added an evaluation using the dataset of Wang et al. (2024)**. Although this dataset also originates from the NMIC network, its phase labels were adjusted using two OTT Parsivel^2 disdrometers and were produced through a different processing procedure. We therefore describe it as a supplementary or quasi-independent observational benchmark rather than as fully independent data. The additional results and evaluation details have been included in the Appendix. We have also acknowledged that the limited availability and temporal coverage of ground-based phase observations constrain robust evaluations across elevation zones and for mixed-phase precipitation.

****

Please see Section 3.3.2 in detail:

To support cryospheric research, a dedicated module was incorporated to explicitly classify the physical state of precipitation into four categories: Dry (0), Snow (1), Mixed (2), and Rain (3). In alpine regions, lower atmospheric pressure and reduced humidity intensify the sublimational cooling of falling hydrometeors, allowing solid precipitation to survive at relatively warmer near-surface temperatures. To account for this, the critical 2-m air temperature threshold for snowfall was dynamically parameterized using the SRTM DEM. The threshold shifts linearly from -2.0℃ at lower elevations to -1.0℃ for regions above 4,000 m, while pure rainfall is defined strictly above 2.0℃. To address ERA5 uncertainties in resolving micro-topographic inversions, this logic is further constrained by the IMERG liquid probability layer: probabilities >90% are forcibly reclassified as rain, and <10% as snow. To evaluate the phase-classification results, we used ground-based precipitation-phase observations that were not used for CRISP training. The primary observational reference was the official Present Weather classification from the National Meteorological Information Centre (NMIC) station network. This reference provides direct station-based information on precipitation phase, including rain, mixed precipitation, and snow.

We further conducted a supplementary evaluation using the dataset developed by Wang et al. (2024). This dataset was also derived from the NMIC station network, but its phase classification was adjusted using measurements from two OTT Parsivel^2 disdrometers. It therefore provides a complementary benchmark based on a different observation-processing procedure. Because the two datasets share the same underlying NMIC station network, the Wang et al. (2024) comparison is considered supplementary rather than fully independent validation.

For both observational references, precipitation-phase classifications were matched in space and time between the gridded CRISP product and the available station observations. The categorical performance of each phase was evaluated using the probability of detection (POD), false alarm ratio (FAR), and critical success index (CSI). For the multiclass phase classification, each phase was evaluated using a one-versus-rest contingency table. In addition, the volumetric fraction of each phase was calculated as the precipitation amount assigned to that phase divided by the total precipitation amount across all matched samples.

****

L. 287: Which long-term regional climatological volume is taken as the reference value to be restored? Given the discussed uncertainties in the different datasets, do the authors think there is

the possibility that the estimates in long-term regional climatological precipitation volume also suffer from biases? Or is the assumption that undercatch is the main uncertainty which is not visible in the regional climatological mean because the missing precipitation at one point will be included at another point? Please clarify here, the motivation and underlying rationale why temporal and spatial variability in precipitation is assumed to be biased but the climatological mean represents the truth.

Response:

Thank you for raising this important point. **In the revised manuscript, we clarify that the reference distribution used for the climatological calibration is derived from the available NMIC station observations after the wind-induced undercatch correction described in Section 2.3.** The corrected station observations are assigned to the 0.1° grid, and observations from multiple stations within the same grid cell are averaged to form grid-level training targets. The resulting distribution is therefore an observation-constrained and physically adjusted reference, rather than an error-free estimate of the true regional precipitation volume.

We also acknowledge that this long-term regional reference may still contain biases. Possible sources include residual errors in the empirical undercatch correction, gauge representativeness errors, sparse and uneven station coverage, missing observations, uncertainty in the auxiliary wind and temperature data, and the assumption that the correction relationships are transferable across different gauge types, environments, and precipitation regimes. In particular, the climatological calibration does not remove all spatial sampling bias associated with the sparse station network. Nor do we assume that precipitation missed by one gauge is necessarily captured by another gauge: wind-induced undercatch can be spatially coherent, especially during snowfall and strong-wind events, and therefore cannot be eliminated simply by spatial averaging.

The rationale for applying climatological calibration is consequently limited and pragmatic. The machine-learning model is optimized primarily for daily event-level prediction and may produce a small systematic bias in the long-term marginal distribution because of variance reduction, class imbalance, and the uneven distribution of precipitation intensities. CDF mapping adjusts this distribution toward the selected observation-constrained reference while retaining the event-level spatial and temporal ordering supplied by the model. **It does not establish that the reference climatology is unbiased, nor does it imply that temporal and spatial variability is biased while**

the climatological mean is exact. Rather, CRISP treats the long-term reference as a calibration constraint with its own uncertainty. We have revised the text to state this limitation explicitly and have moderated the corresponding claims throughout the manuscript.

****

Please see Section 3.3.3 in detail:

The resulting transfer function maps the empirical quantiles of the preliminary estimates toward those of this corrected-gauge reference. The climatology-adjusted layer (Prcp_Clim) therefore provides improved consistency with the selected long-term observational baseline while retaining the spatial and temporal structure of the model estimates. It should not be interpreted as recovering the error-free regional precipitation volume, because the reference itself may be affected by residual undercatch-correction errors, gauge representativeness errors, and sparse and uneven station coverage. In particular, spatial averaging does not necessarily compensate for precipitation missed coherently by multiple gauges during snowfall or strong-wind conditions.

Because the corrected-gauge reference is derived from a sparse and uneven station network, the climatological calibration is subject to sampling uncertainty. The resulting Prcp_Clim product should therefore be interpreted as a reference-based distributional adjustment rather than as an absolute correction of the true regional precipitation volume.

****

L. 510: perform this negative bias → exhibit this negative bias

Response:

Done.

****

Please see Section 4.6 in detail:

Satellite algorithms (IMERG, GSMaP) largely exhibit this negative bias, consistently failing to satisfy the SWE baseline.

****

Fig. 6: Does the black curve labeled "Observations" here refer to all station records?

Response:

Thank you for the question. **The black curve labelled "Observations" in Figure 6 represents the empirical distribution of the quality-controlled NMIC daily gauge observations used as the reference for the climatological calibration.** Before constructing this distribution, the raw gauge measurements were corrected for potential wind-induced undercatch following the procedure described in Section 2.3. The corrected station records were then matched with the corresponding 0.1° CRISP grid cells and dates. Thus, the curve does not represent all raw station records, nor does it represent an independently gridded regional precipitation climatology. We have clarified this point in the figure caption and the main text.

****

Please see Section 4.3 in detail:

The total results of this calibration are validated by the CDF curves (Figure 6c). Due to the known drizzle effect, the cumulative probability for ERA5 increases far too quickly compared to the actual ground truth. Conversely, the CDF of CRISP_Clim closely follows the corrected-gauge reference distribution. This indicates that the calibration reduces the discrepancy in the marginal precipitation distribution relative to the selected observational baseline. This precise distributional alignment indicates that the CRISP framework preserves the daily high accuracy generated by the regressor process while improving consistency with the long-term corrected-gauge distribution.

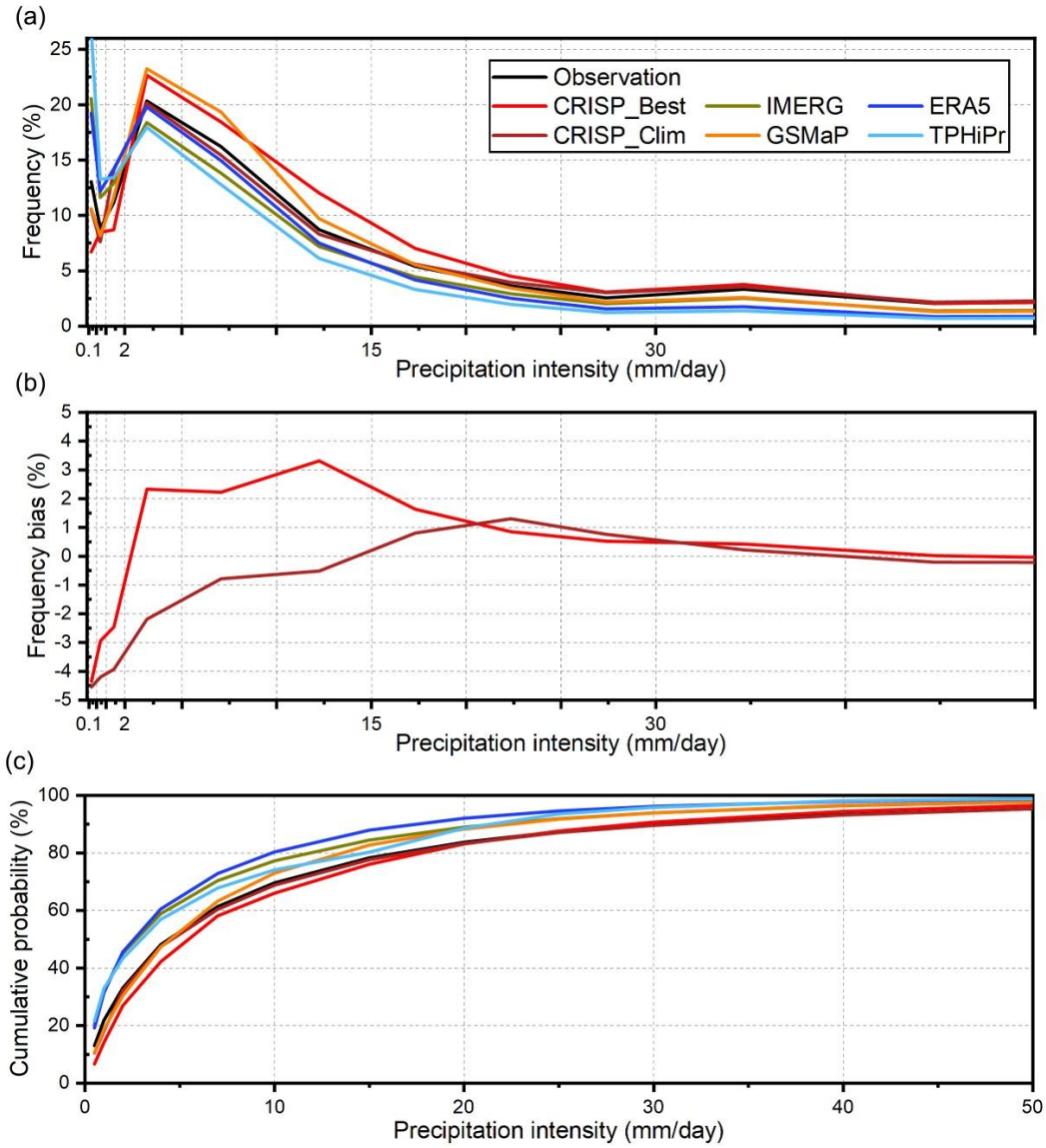

Figure 6. Climatological distribution and calibration performance of daily precipitation intensities. **(a)** Probability density functions (PDF) displaying the occurrence frequency across different precipitation intensity bins. **(b)** Frequency bias of the uncalibrated preliminary estimate (CRISP_Best) and the climatologically calibrated output (CRISP_Clim) relative to the gauge baseline. **(c)** Empirical Cumulative Distribution Functions (CDF) of the evaluated datasets. The black "Observations" curve represents quality-controlled NMIC station-day precipitation records after wind-induced undercatch correction.

****

References

Lyu, Y., Yong, B., Huang, F., Qi, W., Tian, F., Wang, G., & Zhang, J. (2024). Investigating twelve mainstream global precipitation datasets: Which one performs better on the Tibetan Plateau?. *Journal of Hydrology*, 633, 130947. https://doi.org/10.1016/j.jhydrol.2024.130947

Ma, Z., Xu, J., Dong, B., Hu, X., Hu, H., Yan, S., ... & Weng, F. (2025). GMCP: a fully global multisource merging-and-calibration precipitation dataset (1-hourly, 0.1°, global, 2000–the present). *Bulletin of the American Meteorological Society*, *106*(4), E596-E624. https://doi.org/10.1175/BAMS-D-24-0051.1

Ma, Z., Xu, J., Ma, Y., Zhu, S., He, K., Zhang, S., ... & Xu, X. (2022). AERA5-ASIA-Asia: a long-term asian precipitation dataset (0.1°, 1-hourly, 1951–2015, Asia) anchoring the ERA5-land under the total volume control by APHRODITE. *Bulletin of the American Meteorological Society*, *103*(4), E1146-E1171. https://doi.org/10.1175/BAMS-D-20-0328.1

Ma, Z., Xu, J., Zhu, S., Yang, J., Tang, G., Yang, Y., ... & Hong, Y. (2020). AIMERG: a new Asian precipitation dataset (0.1∘/half-hourly, 2000–2015) by calibrating the GPM-era IMERG at a daily scale using APHRODITE. Earth System Science Data, 12(3), 1525-1544. https://doi.org/10.5194/essd-12-1525-2020

Wang, C., Zhao, L., Ma, L., Hu, G., Zhang, L., Zou, D., ... & Zhao, J. (2024). Precipitation adjustment by the OTT Parsivel2 in the central Qinghai–Tibet Plateau. *Earth Surface Processes and Landforms*, *49*(8), 2424-2441. https://doi.org/10.1002/esp.5836