scientific data

Check for updates

OPEN

DATA DESCRIPTOR

A station-based 0.1-degree daily gridded ensemble precipitation dataset for India

Anagha Peringiyil[1], Manabendra Saharia[1]✉, Sreejith O. P.[2], Andrew W. Wood[3], Mrutyunjay Mohapatra[2], Bharti Sabde[2], Aradhana Kumari[2], Bhushan Phadkar[2], Sabeerali C. T.[2], Rohini P.[2], Hosalikar K. S.[2] & M. Rajeevan[4]

Gridded precipitation products are inherently uncertain and predominantly deterministic, which limits their applicability in data assimilation systems and hydrologic modeling. This limitation is significant in developing countries such as India, where the observation network is sparse and non-uniform, topography is complex, and hydrometeorological extremes are frequent. The current official 0.25° observed precipitation dataset of the Indian Meteorological Department (IMD) is deterministic and based on Shephard's interpolation technique. To address these challenges, we have developed the Indian Precipitation Ensemble Dataset (IPED) leveraging the largest network of precipitation gauge stations across India and using a locally weighted spatial regression approach. IPED is a daily 30-member ensemble precipitation product available at 0.1° and 0.25° resolution (1991–2020), accounting for topographical variation in elevation, slope, and aspect. For all thresholds, including the extreme 99th percentile precipitation during monsoon, the developed ensemble product exhibits higher discrimination and reliability. This is the first observation-based ensemble precipitation product over India and is expected to have widespread hydrometeorological applications.

Background & Summary

Gridded meteorological estimates are critical for comprehending global water and energy cycles and can be used in a wide range of scientific and practical applications such as hydrological modeling and climate studies[1]. There are numerous gridded meteorological datasets available from different data sources, varying in spatiotemporal extent, resolution, meteorological variables, and diverse application objectives[1,2]. Most gridded meteorological datasets are deterministic, where they present a single estimate of a particular variable at a given time and location, without accounting for uncertainties. These uncertainties in meteorological data products result in imprecise parameter estimation of the hydrological models[3]. To estimate the uncertainty in such datasets, it is necessary to quantify errors in surface meteorological fields and develop probabilistic or ensemble meteorological frameworks[4]. Understanding these uncertainties is important, especially in developing countries where the station network is sparse and not distributed uniformly. Among the meteorological variables, gridded precipitation products are extensively developed and utilized in India, but they are all deterministic and employ distance-based interpolation techniques[5–8]. An observation-based ensemble dataset that also accounts for topographic heterogeneity is currently missing in India and is a critical requirement for hydrologic and climate impact studies.

Gauge stations are the most trustworthy sources of data for developing and evaluating gridded meteorological datasets due to their high accuracy and longtime coverage[9]. However, maintaining such resources is challenging, particularly in remote locations and mountainous terrains. Additionally, gauge station-based estimates also contain measurement errors due to precipitation under-catch or over-catch as well as instrument errors[4]. This is especially true for isolated regions with challenging climatic and geographical conditions, where nearly all meteorological datasets show significant uncertainty[10,11]. Though gauge stations collect rainfall data at non-uniform points, researchers need spatially continuous data for precipitation and hydrological studies. Interpolating the non-uniformly spaced rain gauge points to a uniform grid is one approach to solve the problem. Various

[1]Department of Civil Engineering, Indian Institute of Technology Delhi, Hauz Khas, New Delhi, 110016, India. [2]Indian Meteorological Department, Delhi, India. [3]National Center for Atmospheric Research, Boulder, Colorado, USA. [4]Vice Chancellor, Atria University, Bengaluru, India. ✉e-mail: msaharia@iitd.ac.in

techniques are utilized for spatial interpolation, such as Inverse Distance Weighting, Geostatistical methods like kriging, and statistical methods like Generalized Least Square Trend[12]. Despite the increasing availability of radars, satellites, and atmospheric model-based datasets, accurate estimation of uniformly distributed spatial rainfall dataset remains challenging[2,13–15].

In addition to the aforementioned challenges, high spatiotemporal variability in rainfall makes the availability of high-resolution rainfall datasets crucial in India[5]. In recent years, there has been a rise in the number of floods reported in India[16], which shows the need for a better understanding of how rainfall is distributed throughout the country. In addition, high-resolution precipitation estimates are essential for developing flash flood severity metric[17,18], understanding the impact of rainfall spatial variability on flooding[19], and assessing the impact of rainfall erosivity on soil erosion[20]. Apart from that, proper uncertainty quantification is required to improve the reliability of hydrologic predictions, as the uncertainties in precipitation translate into hydrologic model simulation[4,21]. Furthermore, the lack of reliable uncertainty estimates for observed meteorological fields is a hurdle towards greater application of data assimilation in hydrology[21]. Hence, high-resolution precipitation datasets along with uncertainty estimation, are important for multiple hydrological applications[4,22–24].

Considerable amount of research effort has been dedicated to the creation of deterministic precipitation datasets pertaining to the Indian region. These include the 1° spatial resolution dataset from 1951–2003[6], 1° spatial resolution dataset for 1901–2004[5], 0.5° spatial resolution dataset for 1971–2005[25], 0.25° spatial resolution dataset for 1901–2010[8], 1° and 0.25° rainfall datasets for the region including the Island area for the period of 1951–2020[26]. These datasets developed by the Indian Meteorological Department (IMD) were created using similar procedures of multistage quality checks followed by algorithms such as Shepard's Interpolation method[27]. Shepard's interpolation method is based on the principle of Inverse Distance Weighing (IDW), which considers both the barriers and directional effects[27]. However, in areas far from known recording stations, the efficiency of such methods becomes very limited[7]. Furthermore, spatial interpolation is an imperfect process that results in prevalent uncertainties in gridded meteorological datasets[3]. According to Jena *et al.*[28], the interpolation method can introduce significant errors in complex terrains, such as the Himalayas, due to sparse and unevenly distributed station. The sparse density of stations in these regions results in higher uncertainty, particularly for extreme events, which are more susceptible to inaccuracies in interpolation. To obtain accurate climatic data in such regions, improvements in interpolation techniques and station maintenance must be addressed, in addition to the option of establishing dense stations[28]. Additionally, Goteti *et al.*[29] has reported a substantial underestimation of precipitation in the IMD dataset over wet regions of India. Moreover, they do not incorporate the uncertainties due to measurement errors caused by factors such as evaporation or wetting loss and under-catch of precipitation[1,11]. Recently, a probabilistic stochastic lattice model approach was implemented for interpolating daily rainfall data over India for the monsoon months of 1951–70[7]. However, this dataset also lacks estimation of precipitation uncertainty.

Several recent developments have occurred in the field of station-based gridded dataset generation[4,11,30]. One of the significant advancements in the field is probabilistic precipitation estimation algorithms that can generate ensemble meteorological datasets[4,15,23,31,32]. Probabilistic datasets are helpful in determining uncertainty and capturing extremes[4,11,21]. The advantage of these ensemble precipitation dataset is, it comes with an estimation of uncertainty, which will be further useful in the advanced data assimilation and land surface and hydrologic modeling[1,4,21]. Newman *et al.*[4] developed a tool named, the Gridded Ensemble Meteorological Tool (GMET), applies this concept via a locally weighted spatial regression for creating gridded ensembles of precipitation and temperature based on observational station data records. GMET has been further refined and expanded through the sequential application initiatives, creating various datasets ranging from regional to continental levels[1,9,15,22–24,31]. A more recent development is the Geospatial Probabilistic Estimation Package (GPEP) which builds upon GMET, while introducing new methodological and usability advancements[9]. The algorithm has enhanced capabilities, including better configurability and variable definition, expanded choices for integrating machine learning approaches, an alternate method for cross validation, and improved flexibility in input formatting. It employs probabilistic interpolation, incorporating both deterministic regression and ensemble-based perturbations. Furthermore, it integrates static predictors such as elevation and slope into its regression model, allowing it to account for orographic effects, rain shadows, and other topographical influences. By introducing Spatially Correlated Random Fields (SCRFs)[21], GMET quantifies uncertainty, especially in regions where data is sparse. It calculates regression residuals for each grid point, explicitly modeling the uncertainty associated with sparse station data or complex terrain[4].

At present, India lacks an observation-based probabilistic ensemble precipitation product. The proposed approach, estimating uncertainty rather than providing a single value for precipitation would be more beneficial, especially in hydrological applications. Furthermore, previous observation-based gridded datasets in India have been prepared using data from gauge stations without considering the influence of the complex underlying terrain. To the authors' knowledge, no prior work in India has utilized spatial features such as latitude, longitude, elevation, and slope in concurrence with rain gauge stations to create an ensemble gridded dataset using observed station network. This study develops a high-resolution (0.25° and 0.1°) daily ensemble gridded precipitation dataset using a locally weighted spatial regression method, for 30 years from 1991 to 2020 over the Indian region. The primary application of the developed dataset is to force in land surface and hydrologic models and conduct hydrologic data assimilation studies. However, it could also be used to validate atmospheric models and identify extreme events.

Methods

Study area. The study area for developing the ensemble precipitation product encompasses the whole Indian subcontinent (Fig. 1). The research locations cover the entire Indian region, with a geographical scope ranging from latitudes 6°N to 35.5°N and longitudes 68°E to 100°E. The region exhibits one of the most diverse precipitation climatology of the world. According to the IMD rainfall statistics 2022, northwestern region receives

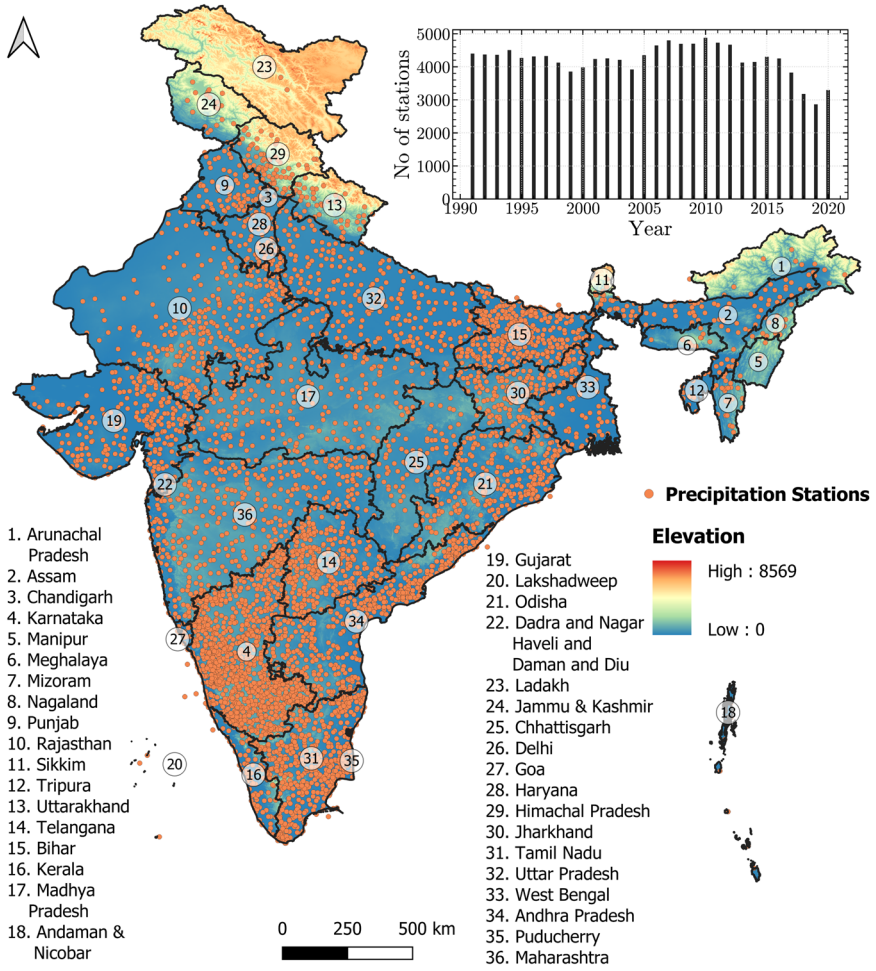

Fig. 1 A map representing the spatial domain of the Indian mainland, including the location of IMD precipitation gauge stations. The thick black line outlines the state administrative boundaries, color contours illustrate the elevation of the terrain, and orange dots represent station locations. The bar plot diagram shows the yearly fluctuation in the number of stations used to generate IPED across India.

an average of 827.6 mm of annual precipitation, and the central India receives 1304.5 mm of rainfall annually. Further, the southern peninsula receives 1394.4 mm and the northeastern and eastern region receives 1815.6 mm[33]. In recent years, the country has experienced several cloudburst events such as in Uttarakhand (2013) and Himachal Pradesh (2015)[28]. The absence of sophisticated ensemble precipitation products in the region coupled with its complex topography is a motivation for this study.

Input data. *Daily rain gauge network.* This study utilizes daily rain gauge station data obtained from the Indian Meteorological Department (IMD). A daily 24-hour accumulated rainfall from gauge stations was collected for a period of 30 years, with an average of 3966 stations per day. This collection consists of varying numbers of stations for each day from 1991 to 2020, that were quality controlled for unique gauges with less missing values. The variability observed in the number of stations employed for the analysis is illustrated on a yearly basis in a bar plot (Fig. 1). The data density exhibited fluctuations across the years, with a peak of over 3750 stations between 1991 and 2016 and the number of stations decreased to less than 3000 from 2017 to 2020. Also, the spatial distribution map shows a higher station density in the southern peninsular region compared to northeast India, northwest India and parts of central India (Fig. 1). Additionally, the number distribution of rainfall stations with different elevations are also shown in the supplementary figures (Fig. S3).

Geospatial attributes. The study used SRTM Digital Elevation Model (DEM) model data which is downloaded from the Google Earth Engine (GEE) with a resolution of 90 m. Custom Python scripts were used to extract elevation, slope (north-south and east-west), and aspect from GEE for gauge station and grid locations. A 2D gradient filter is applied to smooth the slope, reducing high-frequency noise caused by the full-resolution topography. This process also enhances precipitation patterns along mountain ranges, highlighting the broader regions of higher precipitation on the windward side and lower precipitation on the leeward side[4].

IMD gridded rainfall dataset. In this study, the official and widely used 0.25° IMD daily rainfall dataset was utilized as a baseline for comparison with the newly developed rainfall dataset[8]. The Shepard interpolation method was used to develop this product[27]. The interpolated values in the Shepard approach are calculated from the weighted total of the measurements. To determine the weighted average, only rainfall measurements from a couple of nearby stations were used with a minimum of 1 station and a maximum of 4 stations within a search radius of the 1.5° radial distance of grid the point[8]. Interpolation is only valid inside the influence radius. A missing code is assigned to the grid point value when the search distance is equal to or higher than the radius of impact, and no station location is discovered within this range.

Preprocessing and data imputation. The accuracy and reliability of the meteorological station records in this study were ensured through various quality control (QC) procedures and systematic verification. The precipitation data from gauge stations was sourced from the Government agency, the India Meteorological Department (IMD), which archives observations from up to 6,955 rain gauge stations. Each station's location information, including latitude, longitude, and elevation was fetched and was verified to ensure accurate spatial representation. Comprehensive quality checks were conducted to identify and correct errors in the station data. Initially, duplicate entries for the same station and day were identified and removed to eliminate redundancy. Some stations had the same name but were located at different latitudes and longitudes. To address this, we retained only the latitude, longitude pair with the highest average precipitation throughout the study period. Then screening of stations involves the exclusion of stations with less than a decade of reliable data spanning from 1991 to 2020. These stations are expected to possess complete and uninterrupted data records. Research has shown that the precision and pattern of gridded estimations are improved by the implementation of a gap filling strategy[34]. The station data must be present throughout the time and sequentially complete in order to reduce significant fluctuations in computed lapse rates as stations are added, removed, or report missing data for a certain occurrence of the event. We implemented the MICE (Multiple Imputation by Chained Equations)[35] algorithm to make the input gauge station data serially complete. The Iterative Imputer employs a technique of running multiple regressions on randomly selected samples of the data and subsequently consolidating the results to impute the missing values. Once the gap filling is done, the available station data are used to determine the spatial attributes such as elevation, slope and aspect.

Generation of ensembles. Figure 2 depicts the overall methodology employed in the study. We used GPEP tool for developing ensemble precipitation dataset. The GPEP tool was integrated with gauge station data for all non-water pixels to create a gridded daily ensemble of precipitation. The GPEP requires three input files: A station metadata, which is a list of stations along with their latitude, longitude, elevation, and slope orientation towards the north and east; a station data time series file, which is a netCDF file containing daily quality checked and filled gauge station measurement of precipitation; and a netCDF file which specifies the target dataset grid, containing the x and y axis resolution, latitude, longitude, elevation, and north-south and east-west gradients (slopes). GPEP has two procedural steps: spatial regression and ensemble generation, which are run in sequential order. GPEP is coded to accommodate both precipitation and temperature inputs and subsequently produce datasets for both variables while considering their temporal autocorrelation, spatial correlation and variable cross-correlation. Since the density of temperature monitoring stations is much lower[36] than of the precipitation monitoring stations in India, we generate only gridded precipitation outputs. GPEP can estimate deterministic, and ensemble distribution of geophysical variables. This includes the generation of meteorological datasets to facilitate retrospective and real-time modeling on a variety of scales. GPEP is improved upon GMET with expanded functionalities and is written in Python, which is a more flexible and user-friendly programming language[4,9,23].

In this study the representation of mean (μ) in GPEP is accomplished through the utilisation of deterministic gridded estimates derived from locally weighted linear regression (LWLR) with static terrain-related predictors including aspect, latitude, longitude, elevation, and topographic slope. The uncertainty of gridded regression estimates can be calculated using the prediction error or the standard error of the regression. GPEP estimates the probability of precipitation (PoP) for the intermittent variable precipitation using locally weighted logistic regression to enable probabilistic estimation; that is, the binary probability of the event (0 or 1) is regressed against the static predictors, which are also utilised in a precipitation amount regression. The GPEP spatial regression estimation used in this study includes the following four steps (which are also detailed in[9] and previous GMET publications[4,23]):

1. At each grid cell, distance-dependent weights are computed for all stations within a defined radius (Eq. 1).

$$W_{i_{sta},i_{sta}} = \left[1 - \left(\frac{d_{i_{sta}}}{MAXD}\right)^3\right]^3 \quad (1)$$

   Where $W_{i_{sta},i_{sta}}$ is the diagonal matrix of weight. $d_{i_{sta}}$ is the distance to the respective station from the correspond ing grid point and MAXD refers to the specified maximum distance. In our case we have set MAXD to 80 km, and number of stations to 25.
2. The PoP at every point on the grid is estimated via multivariate, locally weighted logistic regression (Eq. 2). This method used the precipitation occurrence data, P_{occur} (yes/no) from the nearest stations, applied locally varying weights (W), and incorporated spatial attributes, such as latitude, longitude, and elevation.

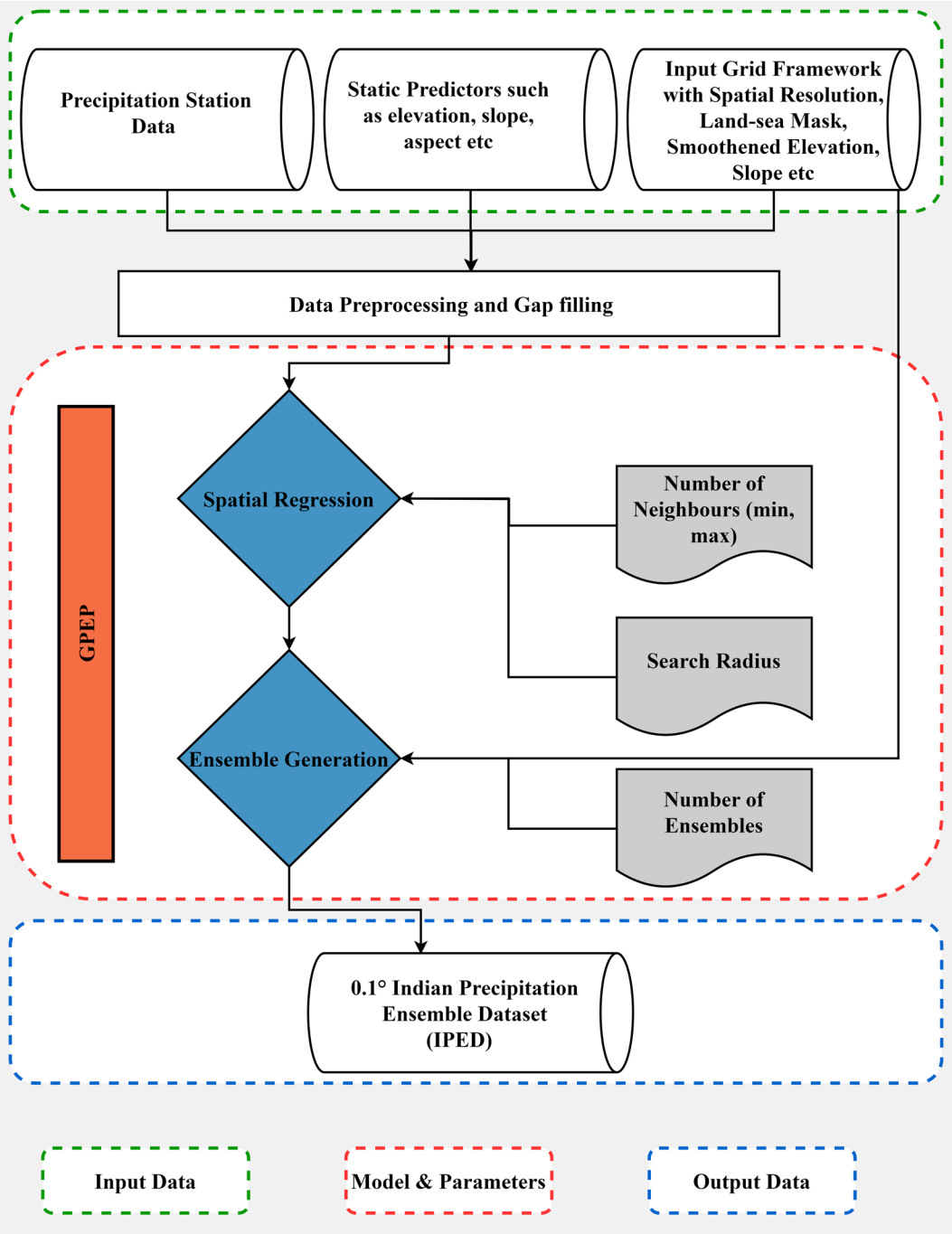

Fig. 2 The methodology adopted for generating the IPED dataset. The input data to the model is represented by a dotted green line. The red dotted part represents the step where the model and its relevant parameter are determined. The grey box displays the critical parameters that impact the model's performance, which are identified through sensitivity analysis. The development of the ensemble precipitation involves a two-stage process, represented by the blue diamonds.

$$PoP_{i_{grid}} = \frac{1}{1 + \exp(-S_{i_{grid}}\beta)} \tag{2}$$

Where $S_{i_{grid}}$ = row vector of spatial attributes such as latitude, longitude, elevation of the grid cell. β is the regression coefficient vector and calculated using the equation Eq. 3, where β is initialized as column vector of ones.

Number of ensemble members	Radius of search (Km)			Number of stations		
30	150	100	80	30	25	20
50						

Table 1. Different parameters and their possible values for sensitivity analysis.

$$\beta_{new} = \beta_{old} + (X^T W V X)^{-1} X^T W (P_{occur} - \pi) \tag{3}$$

Where $\pi = [\pi_1, \pi_2, \ldots. \pi_{nsta}]^T$ is the estimated probability of precipitation at each station. W is the weight matrix and V is the variance matrix. X is the spatial attributes of 25 nearest stations.

3. Locally weighted multivariate linear regression is employed to compute the precipitation values at every grid point. Before the application of any regression steps, all input variables are transformed to a normal distribution using Box-cox technique[9]. The transformed non-zero precipitation is P'_{sta}. Then PCP at a corresponding grid point can be calculated using (Eq. 4):

$$PCP_{i_{grid}} = S_{i_{grid}} \beta^{\alpha} \tag{4}$$

Where

$$\beta^{\alpha} = (X^T W X)^{-1} X^T W P'_{sta} \tag{5}$$

W, X, and V is the weight matrix, spatial attributes of 25 nearest stations and variance matrix, respectively (Eq. 5).

4. Uncertainty for precipitation amounts at individual grid points is estimated from the regression residuals. To estimate the error (Error) in the predicted precipitation (PCP) at the target grid cell, i_{grd} begin by using the method described earlier to calculate the PCP at each of the n_{sta} stations within the modeling domain (Eq. 6).

$$Error_{i_{grid}} = \frac{\sum_{i_{sta}=1}^{n_{sta}} W_{i_{sta},i_{sta}} (PCP_{i_{grid}} - P'_{sta_{i_{sta}}})}{\sum_{i_{sta}=1}^{n_{sta}} W_{i_{sta},i_{sta}}} \tag{6}$$

Following this, the result of the spatial regression is used to produce ensemble members of the surface gridded precipitation. Spatially correlated random fields (SCRFs) are drawn from a normal distribution but are conditioned by the variable spatial correlation length. For each ensemble member, a separate SCRF is generated for each day and used to resample the error distributions of the initial spatial regression step. When the cumulative probability (CP) of a random-normal deviate for that particular grid cell and ensemble member ($C_{\text{igrid,iens}}$) is less than PoP (step2), then the precipitation for that ensemble member is set to zero. If the cumulative probability is greater than POP, precipitation is expected, which requires calculating the quantity of precipitation. The precipitation amount was computed by scaling the CP of SCRF and determine the corresponding standard normal deviate of the scaled CP (RN). Then scale RN, by the distribution of precipitation amounts defined by PCP and E to find the transformed precipitation amount (Eq. 7).

$$Precp'_{ens,i_{grd}} = PCP_{i_{grd}} + RN_{i_{grd}} (E_{i_{grd}}) \tag{7}$$

Where $Precp'_{ens,i_{grd}}$ is the transformed precipitation amount at normal space, $PCP_{i_{grd}}$ is the regression estimated transformed precipitation at the ith grid, $E_{i_{grd}}$ is the ith grid point error or uncertainty estimate. $RN_{i_{grd}}$ is the grid point normal deviate for the precipitation SCRF. As the final step, the actual precipitation amount ($Precp_{ens,i_{grd}}$) for the current ensemble member and grid point is transformed back[9] (Eq. 8):

$$Precp_{ens,i_{grd}} = (Precp'_{ens,i_{grd}})^T \tag{8}$$

Sensitivity analysis. The algorithm requires specifying several parameters, including the number of stations, the maximum distance for searching the neighboring stations, and the number of ensemble members to be generated. A comprehensive sensitivity analysis was conducted to determine the optimal values for the parameters, including the search diameter, the number of stations to be designated as neighbors, and the total number of ensembles to be generated for the analysis. Optimizing the selection of parameters for the model involves conducting multiple runs with varying parameter combinations and selecting the configuration that exhibits the highest correlation with the ground data while also taking into account the algorithm's spatial and temporal complexity. Table 1 describes the possible combinations for the model to run in different iterations with different parameters.

Ensemble verification metrics. To assess the performance of the generated dataset, we utilize ensemble verification metrics.

Relative Operating Characteristics (ROC) Curve. Discrimination is the ability of an ensemble to distinguish between an event and a non-event. Summarizing forecast discrimination in probabilistic prediction verification commonly involves the utilization of the area that is under the relative operating characteristic (ROC) curve (AUC)[37]. The ROC curve plots the hit rate (true positive rate) against the false alarm rate for different thresholds (e.g. exceeding 10 mm of rain a day). The area under ROC curve (AUC) is a measure of the discrimination ability of the ensemble, with a value of 1 (close to upper left corner) indicating perfect discrimination and 0.5 (close to or below diagonal) indicating no discrimination.

Reliability. Reliability diagrams depict the calibration function of the ensemble, which is a line graph showing the conditional probability of an event (precipitation > threshold) based on the estimated probability[4]. The reliability diagram illustrates the conditional probability of an event happening across various probability categories. More precisely, the probabilistic estimates are divided into m categories, with each category representing probabilities between 0.0 and 0.1, between 0.1 and 0.2, and so on up to between 0.9 and 1.0. For each category, the average estimated probability and the average of the observed binary data are computed[21]. The calibration function will adhere to the 1–1 line, ensuring that a flawlessly reliable ensemble has the same estimated probability for every observed probability for every event threshold. The calibration function will be located above the 1–1 line in the case of a dry bias (underestimation of events), while the calibration function will be located below the 1–1 line in the case of a moist bias (overestimation of events).

Evaluation metrics. To further validate the datasets, we compared the 0.25° IPED against 0.25° IMD dataset and the 0.1° IPED against the global 0.1° EM-Earth dataset using the following metrics:

Mean Absolute Error (MAE). It is determined by adding together the magnitude differences between the model estimates and observations and dividing the result by the overall number of occurrences. It includes dividing the 'total error' by n after summing the magnitudes (absolute values) of the errors. MAE shows the actual magnitude of error between the ensemble mean and deterministic precipitation.

$$\sum_{i=1}^{n} \frac{|y_i - x_i|}{n} \tag{9}$$

The Eq. 9 is the equation for calculating MAE, in which y_i represents the anticipated value and the x_i represents the actual value, while n represents the total number of points.

Root Mean Square Error (RMSE). It is the total of the square root of the mean differences between the ensemble mean of IPED and deterministic datasets.

Correlation. Correlation is a statistical measure indicating the relationship between two variables. The Pearson correlation coefficient is a statistical method utilized to quantify the degree of linear association between two given sets of data.

Data Records

The IPED dataset consists of a 0.1° observation –based daily precipitation dataset at resolution publicly available in a Zenodo repository[38] (https://zenodo.org/doi/10.5281/zenodo.8199138). It contains two directories for the daily mean and standard deviation of the precipitation ensembles. The IPED dataset has been uploaded as a yearly netCDF file for ease of downloading and usage. These netCDF files contain latitude, longitude, and time as coordinates. The ensemble mean precipitation variable is named pcp and the variable for the standard deviation is named pcp_std. The size of the file per year is approximately 124 MB, and the total data size is 7.28GB.

Technical Validation

Figure 3(a–c) shows the annual mean precipitation of the 0.25° IMD, 0.25° IPED, and 0.1° IPED[38], respectively for 1991–2020. There is general concurrence among the spatial patterns in the 30 ensemble members produced. Nevertheless, every member of the ensemble exhibits a distinct possibility. Further, Fig. 3(d,e) depicts the standard deviation of the 0.25° and 0.1° ensemble respectively, which represents the uncertainty in the ensemble estimates. The uncertainty in precipitation is greater in the Northeastern and Western coastal region of India. The northeastern region, with fewer stations compared to the rest of India, is primarily in the Brahmaputra River basin, which experiences extreme floods every year. This underscores the need for strengthening the station network in this region.

The use of Spatially Correlated Random Fields (SCRF) to generate ensemble values involves enhancing grid point occurrences at unobserved sites with improved values, such as increased precipitation that surpasses the observations, due to topographic corrections. This distinction highlights the significant importance of GPEP compared to other interpolation-based methods, where grid point occurrences are limited to the highest recorded values along with topographic adjustments. This characteristic is expected to enhance the representation of precipitation occurrences within the ensemble. Both the 0.25° and 0.1° IPED estimates illustrate the significant precipitation variability observed across the northeast region. In contrast, the deterministic approach to precipitation modeling shows a comparatively lower magnitude of precipitation in the northwestern region (Fig. 3a).

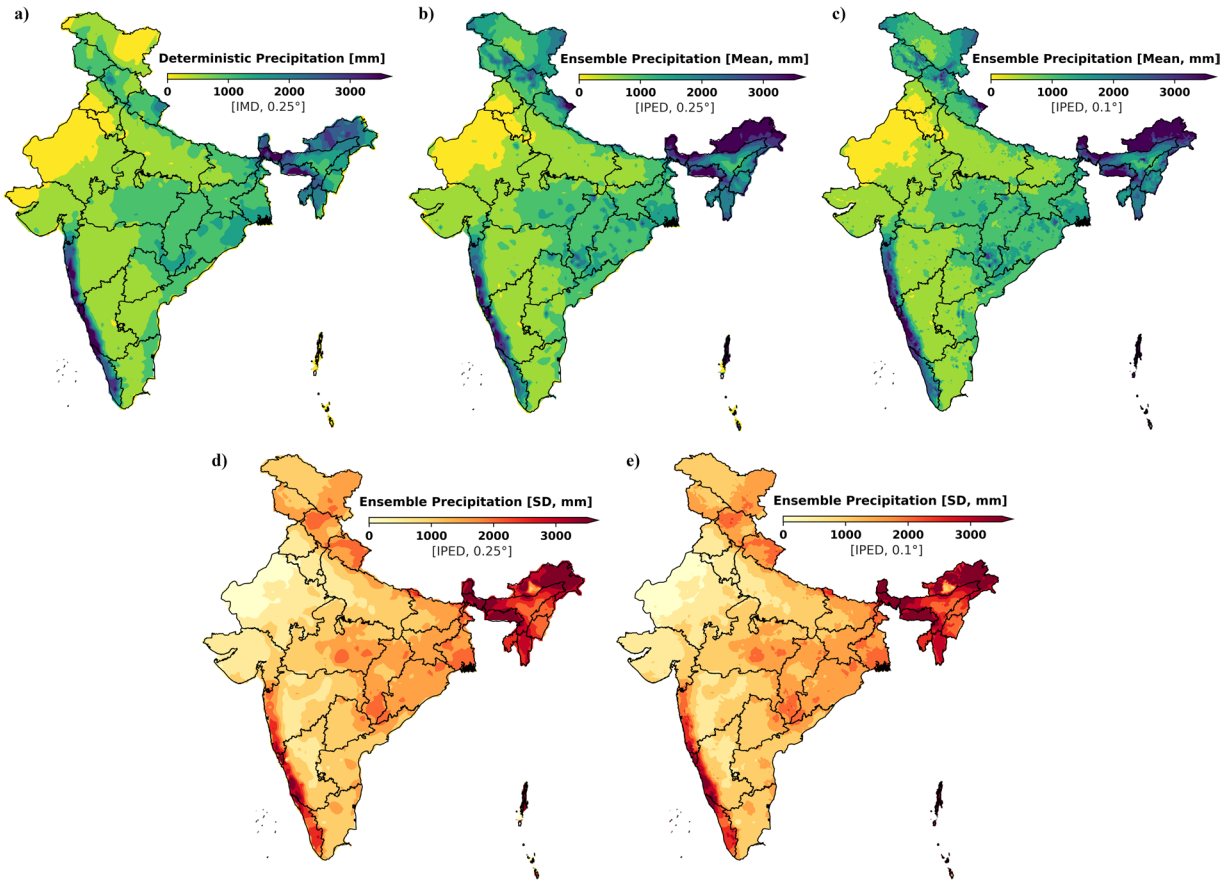

Fig. 3 (**a**) 30-year mean annual precipitation from 0.25° IMD and (**b**) 30-year mean annual precipitation form 0.1° IPED 30 ensembles (**c**) 30 years mean annual precipitation from 0.25° IPED 30 ensembles (**d**) annual mean standard deviation of the ensemble 0.25° IPED and (**e**) annual mean standard deviation of ensemble 0.1° IPED for the duration of 1991 to 2020.

Metrics (Median)	IMD 0.25°	IPED 0.25°	IPED 0.1°
Correlation	0.58	0.66	**0.68**
MAE	2.64	2.69	**2.64**
RMSE	8.51	7.14	**7.05**

Table 2. Error metrics of station precipitation vs their corresponding pixel values for all three precipitation products, 0.25° IMD, 0.25° IPED and 0.1° IPED, from 1991–2020. The overall median of the error metrics is represented as the national value. The bold value highlights the best scores.

We have computed the error statistics, including Mean Absolute Error (MAE), Correlation, and Root Mean Square Error (RMSE), for the three datasets at national scale. Table 2 shows the summary of these error statistics over the domain and demonstrates the superior performance of the 0.1° IPED dataset exhibiting highest correlation, lowest MAE, and the lowest RMSE values among the three.

To emphasize the locations where the ensemble product has shown the significant improvements in correlation, we calculated the correlation between station precipitation and the corresponding pixel values from the 30-member IPED ensemble mean and the IMD deterministic dataset. Four (H1-H4) hotspots have been identified in Fig. 4, where the 0.25° IMD deterministic dataset has a lower correlation, while the 0.25° and 0.1° IPED ensemble products demonstrate higher correlation. H1 is the northwestern hilly region where cloudbursts and devastating flash floods are common. H2 is the eastern region which covers majority of Bihar state, which is very susceptible to flooding. H3 is the northeastern region where monsoonal heavy rainfall often occurs, and is one of the wettest areas in India. H4 is the southern peninsular region are also prone to extreme rainfall events and related flooding. The 0.1° IPED exhibits a strong correlation coefficient and R^2 value, as well as the lowest RMSE, across all four hotspot regions. With the exception of hotspot H1, all three other hotspots have a low MAE for the same product. Therefore, these four regions (H1, H2, H3, and H4) are highlighted as areas where our dataset shows improved performance.

Table 3 shows the correlation, MAE, RMSE, and R^2 in the four hotspots for all the three products against the station precipitation. Among all the products, the 0.1° IPED performs better than the others.

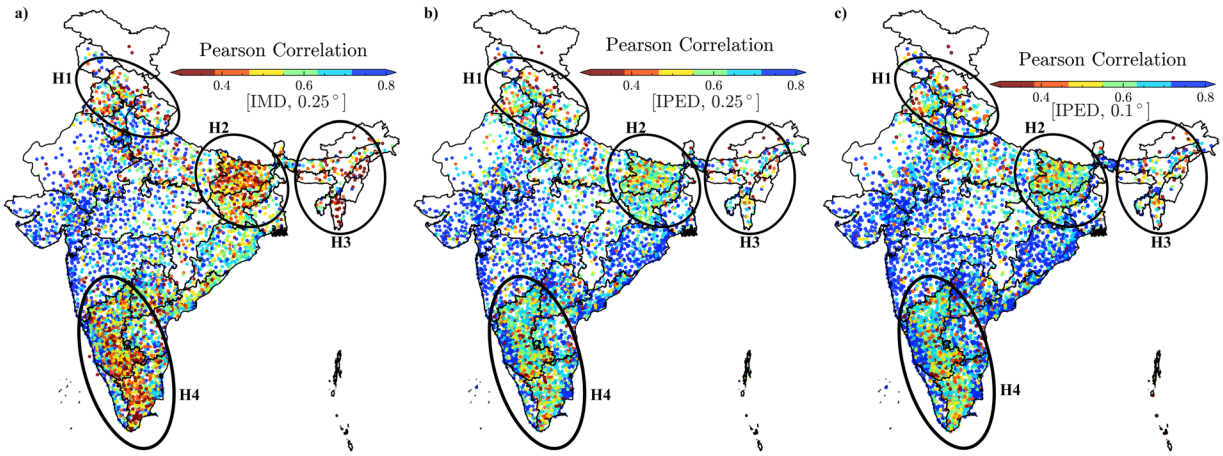

Fig. 4 The correlation of the station precipitation values from 1991 to 2020 for all station vs their corresponding values (**a**) from 0.25° IMD (**b**) from 0.25° IPED (**c**) from 0.1° IPED. The 4-hotspot regions where there is a significant change in the value of correlation were identified and represented as H1, H2, H3, and H4. The oval shape marks the different hotspot regions.

Hotspot	Product	RMSE	MAE	Correlation Coefficient	R^2
H1 (Northwestern)	0.1° IPED	**8.44**	2.38	**0.63**	**0.40**
	0.25° IPED	8.67	2.47	0.61	0.36
	0.25° IMD	8.65	**2.22**	0.63	0.37
H2 (Eastern)	0.1° IPED	9.40	**2.91**	**0.65**	**0.42**
	0.25° IPED	**9.39**	2.93	0.65	0.42
	0.25° IMD	10.99	3.27	0.55	0.21
H3 (Northeastern)	0.1° IPED	**15.08**	**5.72**	**0.65**	**0.41**
	0.25° IPED	15.78	5.93	0.61	0.36
	0.25° IMD	16.87	6.09	0.55	0.27
H4 (Southern peninsula)	0.1° IPED	**8.87**	**2.81**	**0.74**	**0.55**
	0.25° IPED	9.54	2.99	0.69	0.48
	0.25° IMD	9.86	2.93	0.68	0.44

Table 3. Statistical comparison of 0.1° IPED, 0.25° IPED, and 0.25° IMD against station precipitation over the four hotspots [H1-H4]. Numbers in bold represent the highest value for that particular error matrix.

Precipitation product	JJAS Mean precipitation
0.1° IPED	893.86 mm
0.25° IPED	887.83 mm
0.25° IMD	851.02 mm
Station	889.31 mm

Table 4. The national yearly average of JJAS monsoon precipitation of 0.1° IPED, 0.25° IPED, and 0.25° IMD datasets along station precipitation.

Monsoonal rains between June and August contribute to more than 75% of the annual rainfall in India[39] and performance of any precipitation product during this period is critical for agriculture, floods, and droughts. In recent years, the Indian region has endured several catastrophic floods due to extreme rainfall events over a short period, rainfall accumulation over consecutive days, or events like cloudbursts, etc. Several significant incidents include the intense precipitation and consequent inundation in Mumbai (2005), Uttarakhand (2013), Chennai (2015), Kerala (2018), and Assam (2020). Table 4 compares the national yearly averaged seasonal (June, July, August, and September (JJAS)) rainfall over the Indian subcontinent for all the three products. It is observed that the deterministic IMD dataset's seasonal precipitation is the lowest of all the precipitation products. Similarly, the rainfall values of both 0.25° and 0.1° IPED are higher and closer to the station precipitation.

Wet day fraction. The long-term wet day fraction (WDF) is the ratio of number of days that have recorded precipitation values greater than 1 mm to the total number of days. The fraction of wet days can potentially introduce a significant level of uncertainty in the data used to force hydrologic models, particularly when other

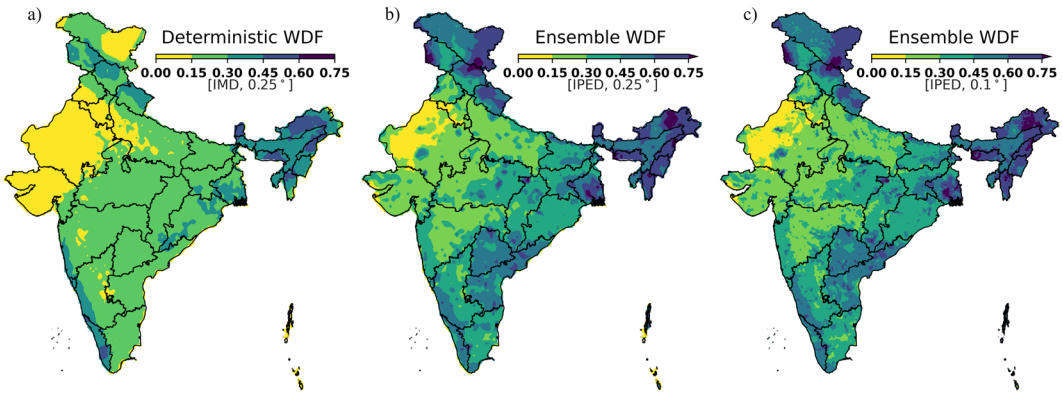

Fig. 5 The long-term wet day fraction (WDF) for (**a**) 0.25° IMD, (**b**) 0.25° IPED and (**c**) 0.1° IPED.

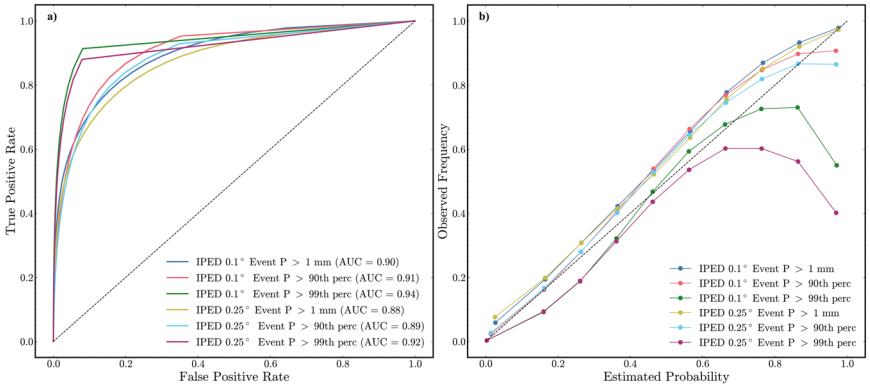

Fig. 6 (**a**) ROC for different event thresholds of precipitation such 1 mm, and 91.2 mm (99 percentile values) for the JJAS seasonal station precipitation from 1991 to 2020. (**b**) Reliability for different event thresholds of precipitation such 1 mm, 22.3 mm (90 percentile values), 91.2 mm (99 percentile values) for the JJAS seasonal station precipitation from 1991 to 2020. The black line indicates the case of perfect reliability.

meteorological parameters are inferred from precipitation data. The mean WDF is computed for both 0.1° and 0.25° IPED by taking the mean of the 30 ensemble members and compared with the WDF derived from the IMD dataset (Fig. 5).

The IPED dataset exhibits higher WDF in the northeastern region, eastern region, and Western Ghats. The IMD dataset indicates a notably low WDF in the arid regions of Gujarat and Rajasthan, while both the IPED dataset reports a WDF up to 0.3 in those areas. Similar results have been reported by Joseph *et al.*[40], indicating that the summer monsoon in northwestern India has increased by 40% from 1979 to 2022. Additionally, Newman *et al.*[4] stated that the ensemble product generates a more genuine precipitation statistic (wet day fraction), which influences the empirical derivation of other fields in land surface and hydrologic modeling. The IMD deterministic data shows less WDF due to the presence of sparse station network. Since IPED utilizes the station data as well as other spatial attributes, it is able to show a greater number of wet days over these areas, which will be helpful in identifying extreme events over these regions. In general, the ensemble datasets exhibit a higher WDF compared to the IMD dataset.

Discrimination and Reliability during Monsoon. The ROC curve for monsoon season (JJAS) is constructed using the station and ensemble daily mean for 1991–2020 (Fig. 6a). The national 99th percentile station precipitation value is 91.2 mm and the national 90th percentile station precipitation value is 22.3 mm. The findings suggest that there is an inverse relationship between the magnitude of the event and the AUC values. For the 0.1° IPED, the smaller threshold of precipitation of 1 mm gives an accuracy rate of 90%, while the higher precipitation thresholds show a higher accuracy rate. For instance, the event and non-event discrimination accuracy for 22.3 mm and 91.22 mm are 91% and 94%, respectively. Similarly, for the 0.25° IPED, the smaller threshold of 1 mm has an accuracy of 88% and the higher threshold of 91.2 mm has 92% accuracy in distinguishing between events and non-events.

Figure 6b illustrates reliability diagrams for three distinct precipitation thresholds: 1 mm, 22.3 mm (90th percentile), and 91.2 mm (99th percentile) for the JJAS period of 1991 to 2020. The reliability diagrams demonstrate that the ensemble exhibits high reliability for all occurrences of precipitation. A minor wet bias emerges at higher probabilities (0.8), and a very slight dry bias appears at a low probability of occurrence as the precipitation event threshold increases. However, probabilistic estimates are very reliable for 22.3 mm and 91.2 mm precipitation

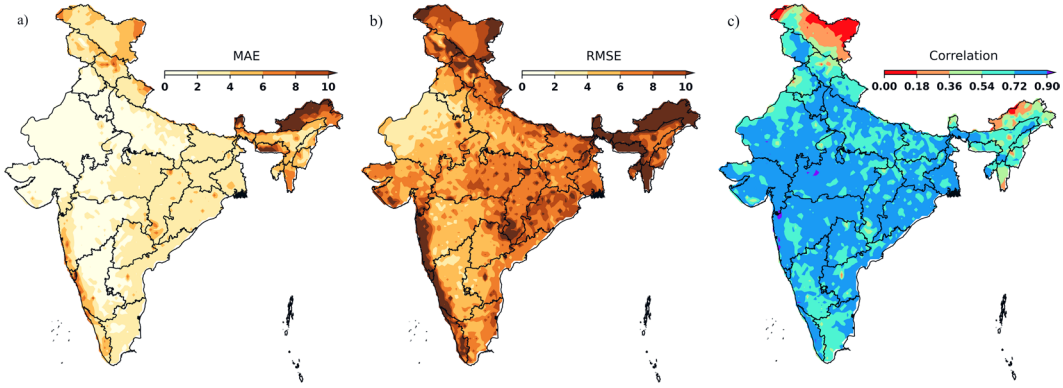

Fig. 7 Comparison of IPED 0.25° with IMD deterministic dataset for the period of 1991 to 2020 with different error metrics such as (**a**) MAE, (**b**) RMSE and (**c**) Pearson Correlation.

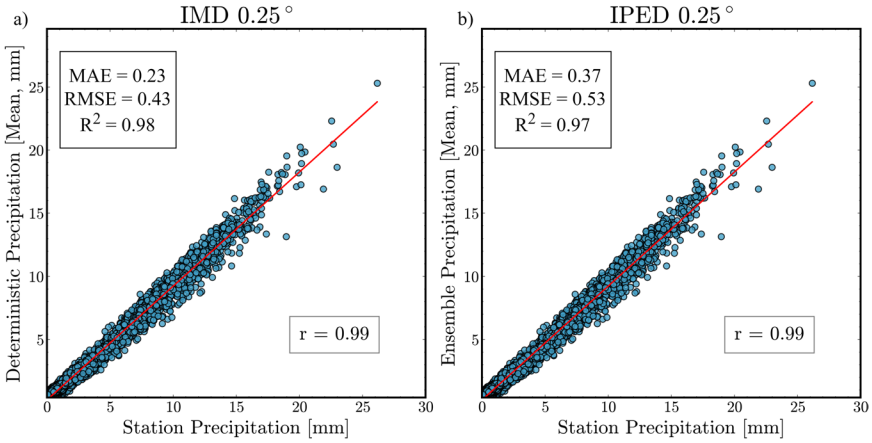

Fig. 8 (**a**) Correlation plot of daily national average precipitation of 0.25° IMD vs station precipitation (**b**) Correlation plot of daily national average precipitation of 0.25° IPED ensembles mean vs station precipitation. The upper left text box shows the error metrics associated with the station data and the corresponding dataset. The text box in the bottom right shows the correlation coefficient of station precipitation and corresponding product.

thresholds. Nevertheless, precipitation events of 91.22 mm (99th percentile) are infrequent, leading to probabilistic estimates nearly in the lowest category. A persistent minor wet bias is observed at the 91.2 mm threshold, in addition to increased sampling uncertainty. This performance closely resembles the findings of Clark *et al.*[21] over the Colorado region. The wet bias is slightly lower for the 0.1° IPED compared to the 0.25° IPED.

Comparison with IMD deterministic. We calculated the mean of the daily difference between the ensemble product (0.25° IPED) and the deterministic product (0.25° IMD) from 1991 to 2020 (Fig. 7). Except for Northeast India, the Mean Absolute Error (MAE) between the two datasets is similar, indicating a high level of agreement across the regions. The Root Mean Square Error (RMSE) emphasizes larger errors by squaring the individual differences, averaging them, and then taking the square root, providing a measure that is sensitive to significant deviations. This means that the RMSE will increase significantly compared to the MAE if there are a few instances where the two datasets differ significantly. This suggests that the datasets generally agree, but there are specific extreme events where they diverge significantly. In addition, a low Pearson correlation is observed between the two datasets over the region such as Ladakh, Arunachal pradesh and parts of Mizoram (Fig. 7). The average daily precipitation patterns among the two datasets exhibit a high degree of similarity across central India.

We evaluated the performance of 0.25° IPED and 0.25° IMD datasets in representing station-based observations. We identified the nearest pixel from both datasets to each gauge station for each day, considering the varying number of stations across days and computed the daily spatial station averages. Similarly, we averaged the corresponding nearest-pixel values from each dataset for the same day. Finally, we calculated the correlation between the station-based daily averages and those from the IPED dataset, as well as between the station data and the IMD dataset, to assess how well each dataset aligns with the observed station data (Fig. 8). The performance of 0.25° IPED vs station precipitation shows a coefficient of correlation of 0.99, a mean absolute error of 0.37, and a root mean square error of 0.53. Similar comparison has been made with 0.25° IMD, which shows a

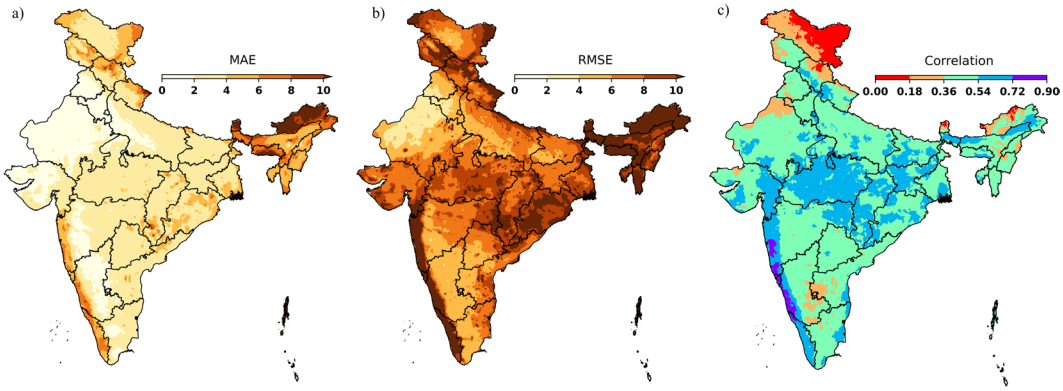

Fig. 9 Comparison of 0.1° IPED with 0.1° probabilistic EM-Earth dataset for the study area period 1991 to 2019 represented by different error metrics (**a**) MAE, (**b**) RMSE and (**c**) Pearson Correlation.

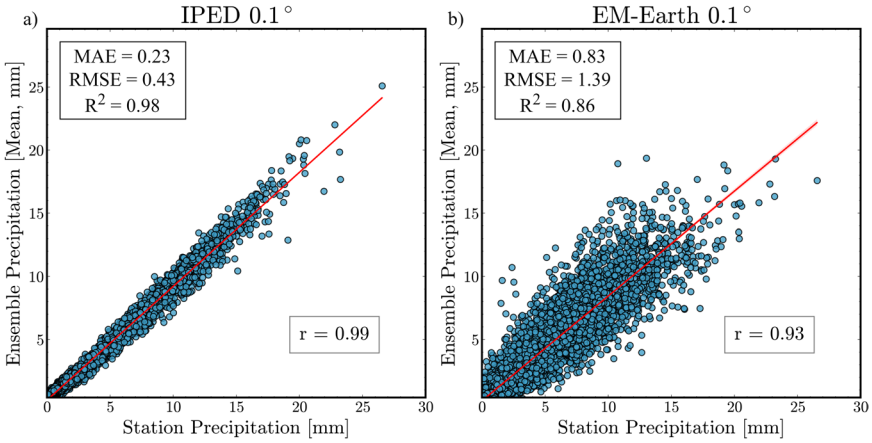

Fig. 10 (**a**) Correlation plot of daily national average precipitation of 0.1° IPED ensembles mean vs station precipitation. (**b**) Correlation plot of daily national average precipitation of 0.1° EM-Earth ensembles mean vs station precipitation. The upper left text box shows the error metrics associated with the station data and the corresponding dataset. The text box in the bottom right shows the correlation coefficient of station precipitation and corresponding product.

coefficient of correlation of 0.99, a mean absolute error of 0.23, and a root mean square error of 0.43. Although the deterministic precipitation exhibits lower MAE and RMSE values when the precipitation is temporally averaged, the ensemble product of 0.25° IPED also shows comparable performances with uncertainty specifications. Hence, we conclude that our 0.25° IPED dataset demonstrates comparable performance to the 0.25° IMD in terms of temporal correlation with the station precipitation.

Comparison with the global probabilistic EM-Earth. To assess the performance of this ensemble dataset, we have compared it against a state-of-the-art global probabilistic dataset called EM-earth, derived from station data at a resolution of 0.1°[1]. EM-Earth offers deterministic estimates on an hourly and daily basis, as well as daily probabilistic estimates with 25 ensemble members. EM-Earth dataset is derived from a station-based Serially Complete Earth (SC-Earth) dataset[1], which eliminates the gaps in time in the original station observations. SC-Earth dataset contains a 6826 number of complete precipitation station dataset over Asia, from 1950 to 2019[34]. The study area duration was taken from 1991 to 2019, since the availability of EM-Earth was till 2019[41]. The results show (Fig. 9) that MAE between them is very similar except for Northeast India and the Western Ghats part, which indicates an overall high-level agreement between the 0.1° IPED and 0.1° EM-Earth dataset over the regions. Similarly, high RMSE indicates that there are specific instances where extreme events diverge significantly between these datasets over regions like Northeast, Eastern part and Western Ghats of India (Fig. 9). In contrast, the Pearson correlation between the 0.1° IPED and EM-Earth shows that the southwest region and few parts of central India show higher correlation.

Figure 10 shows the temporal correlation of ensemble mean of 0.1° IPED and EM-Earth[41] against station precipitation. The correlation between the station-based daily averages and those from the IPED dataset, as well as between the station data and the EM-Earth dataset, was calculated to evaluate the degree to which each dataset is consistent with the observed station data. The same procedure used in the comparison of IPED versus IMD was applied here as well. EM-Earth shows a lower correlation coefficient of 0.93 and a higher MAE/RMSE

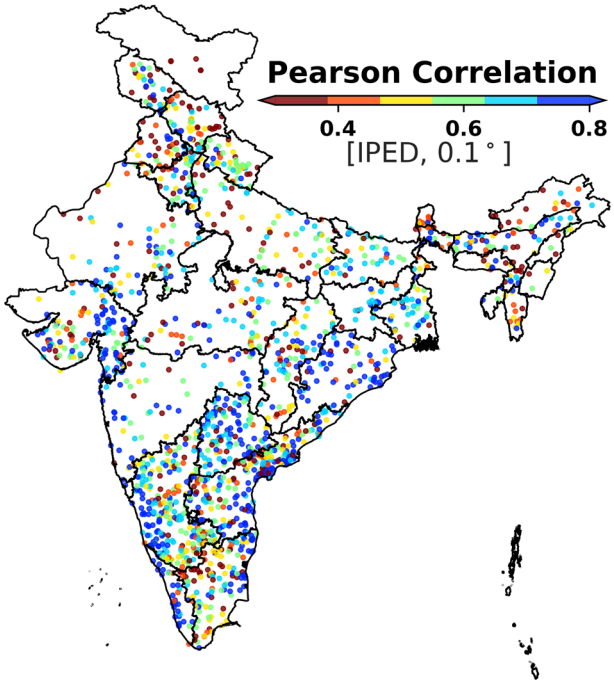

Fig. 11 The correlation of unseen station precipitation with the 0.1° IPED ensemble mean.

compared with the 0.1° IPED. This establishes the vital importance of developing regional datasets of higher fidelity and the superior performance of the 0.1° IPED against the most sophisticated global observation-based gridded dataset currently in existence.

Furthermore, the comparison of 0.1° IPED with AERA-5 Asia (0.1°, hourly)[42,43] from 2000 to 2015 has been performed and shown in supplementary figures (Fig S1 and Fig S2). AERA-5 Asia shows a lower correlation coefficient of 0.97 and a higher MAE/RMSE compared with the 0.1° IPED. This further supports the robustness of our developed IPED dataset[44–46].

Validations using Out-of-Sample stations. All precipitation datasets involve methodological decisions that result in variations in the corresponding final products. To assess how the dataset performs in diverse terrain conditions, we can check its performance in locations where we have station data that has not been used for training purposes. For this, we have taken gauge station data filtered out during the preprocessing stage, which contained less than ten years of daily data. There were 1781 unique stations that have less than ten years of records, and hence, we have not used them for our analysis. Figure 11 illustrates the spatial distribution of those points and their correlation with the 0.1° IPED dataset. The result demonstrated that our methodology enhances the accuracy of precipitation estimates for unknown locations. The stronger correlation reflects the better estimations of precipitation. The regions with the lowest correlation points are Ladakh, Punjab, and Arunachal Pradesh, which have very low station density. Hence, the 0.1° IPED datasets show superior performance over the out-of-sample stations also, which further indicates the potential of 0.1° IPED datasets.

Usage Notes

The IPED dataset contains gridded files with ensemble mean precipitation (PCP) variable and standard deviation, written as daily outputs with a resolution of 0.1° in netCDF format. The developed dataset includes 30 ensemble members spanning 30 years, totaling approximately 145 GB. This comprehensive resource provides valuable material for analysis. However, its large size requires substantial computational power and storage capacity for effective analysis, as well as understanding of ensemble methods. The developed dataset has many potential applications. One of the primary uses of the IPED ensemble members is to assist in estimating uncertainties in the hydrological modeling. Additionally, the mean of the ensemble members can help identify extreme weather events, such as cloudbursts.

Code availability

The dataset was generated using the GPEP tool[9], which is a geospatial probabilistic estimation package written in python, that supports interpolation of specified variables over user-specified regional or global domains using station data. The code is publicly available on https://github.com/NCAR/GPEP.

Received: 10 October 2024; Accepted: 13 January 2025;
Published online: 25 February 2025

References

1. Tang, G., Clark, M. P. & Papalexiou, S. M. EM-Earth: The Ensemble Meteorological Dataset for Planet Earth. *Bulletin of the American Meteorological Society* **103**, E996–E1018 (2022).
2. Sun, Q. *et al*. A Review of Global Precipitation Data Sets: Data Sources, Estimation, and Intercomparisons. *Reviews of Geophysics* **56**, 79–107 (2018).
3. Tang, G. *et al*. The Impact of Meteorological Forcing Uncertainty on Hydrological Modeling: A Global Analysis of Cryosphere Basins. *Water Resources Research* **59**, e2022WR033767 (2023).
4. Newman, A. J. *et al*. Gridded Ensemble Precipitation and Temperature Estimates for the Contiguous United States. *Journal of Hydrometeorology* **16**, 2481–2500 (2015).
5. Rajeevan, M., Bhate, J. & Jaswal, A. K. Analysis of variability and trends of extreme rainfall events over India using 104 years of gridded daily rainfall data. *Geophys. Res. Lett.* **35**, L18707 (2008).
6. Rajeevan, M., Bhate, J. & Kale, J. D. A High Resolution Daily Gridded Rainfall Data for the Indian Region: Analysis of break and active monsoon spells. *Current Science* **91**, 296–306 (2006).
7. Khouider, B. *et al*. A Novel Method for Interpolating Daily Station Rainfall Data Using a Stochastic Lattice Model. *Journal of Hydrometeorology* **21**, 909–933 (2020).
8. Pai, D. S. *et al*. Development of a new high spatial resolution (0.25° × 0.25°) long period (1901-2010) daily gridded rainfall data set over India and its comparison with existing data sets over the region. *MAUSAM* **65**, 19 (2014).
9. Tang, G., Wood, A. W., Newman, A. J., Clark, M. P. & Papalexiou, S. M. GPEP v1.0: the Geospatial Probabilistic Estimation Package to support Earth science applications. *Geosci. Model Dev.* **17**, 1153–1173 (2024).
10. Kochendorfer, J. *et al*. Testing and development of transfer functions for weighing precipitation gauges in WMO-SPICE. *Hydrol. Earth Syst. Sci.* **22**, 1437–1452 (2018).
11. Tang, G. *et al*. EMDNA: an Ensemble Meteorological Dataset for North America. *Earth Syst. Sci. Data* **13**, 3337–3362 (2021).
12. Li, J., Heap, A. D., Potter, A. & Daniell, J. J. Application of machine learning methods to spatial interpolation of environmental variables. *Environmental Modelling & Software* **26**, 1647–1659 (2011).
13. Hu, Q. *et al*. Rainfall Spatial Estimations: A Review from Spatial Interpolation to Multi-Source Data Merging. *Water* **11**, 579 (2019).
14. Kirstetter, P. *et al*. Probabilistic precipitation rate estimates with ground-based radar networks. *Water Resources Research* **51**, 1422–1442 (2015).
15. Newman, A. J. *et al*. Use of Daily Station Observations to Produce High-Resolution Gridded Probabilistic Precipitation and Temperature Time Series for the Hawaiian Islands. *Journal of Hydrometeorology* **20**, 509–529 (2019).
16. Saharia, M. *et al*. India flood inventory: creation of a multi-source national geospatial database to facilitate comprehensive flood research. *Nat Hazards* **108**, 619–633 (2021).
17. Saharia, M. *et al*. Mapping Flash Flood Severity in the United States. *Journal of Hydrometeorology* **18**, 397–411 (2017).
18. Kumar, A., Saharia, M. & Kirstetter, P. Mapping a Novel Metric for Flash Flood Recovery Using Interpretable Machine Learning. *Journal of Hydrometeorology* **25**, 1863–1875 (2024).
19. Saharia, M. *et al*. On the Impact of Rainfall Spatial Variability, Geomorphology, and Climatology on Flash Floods. *Water Resources Research* **57**, e2020WR029124 (2021).
20. Raj, R., Saharia, M., Chakma, S. & Rafieinasab, A. Mapping rainfall erosivity over India using multiple precipitation datasets. *CATENA* **214**, 106256 (2022).
21. Clark, M. P. & Slater, A. G. Probabilistic Quantitative Precipitation Estimation in Complex Terrain. *Journal of Hydrometeorology* **7**, 3–22 (2006).
22. Huang, C., Newman, A. J., Clark, M. P., Wood, A. W. & Zheng, X. Evaluation of snow data assimilation using the ensemble Kalman filter for seasonal streamflow prediction in the western United States. *Hydrol. Earth Syst. Sci.* **21**, 635–650 (2017).
23. Bunn, P. T. W. *et al*. Improving station-based ensemble surface meteorological analyses using numerical weather prediction: A case study of the Oroville Dam crisis precipitation event. *Journal of Hydrometeorology* https://doi.org/10.1175/JHM-D-21-0193.1 (2022).
24. Liu, H., Wood, A. W., Newman, A. J. & Clark, M. P. Ensemble Dressing of Meteorological Fields: Using Spatial Regression to Estimate Uncertainty in Deterministic Gridded Meteorological Datasets. *Journal of Hydrometeorology* **23**, 1525–1543 (2022).
25. Rajeevan *et al*. A high resolution daily gridded rainfall dataset (1975-2005) for mesoscale studies. *Current Science* **96** (2009).
26. Sridhar, L. *et al*. *Development of Daily Gridded Rainfall Data Sets over the Indian Islands at 1°x1° & 0.25°x0.25° Spatial Resolutions for the Period (1951-2020)*. 23 (2021).
27. Shepard, D. A two-dimensional interpolation function for irregularly-spaced data. in *Proceedings of the 1968 23rd ACM national conference on* - 517–524, https://doi.org/10.1145/800186.810616 (ACM Press, Not Known, 1968).
28. Jena, P., Garg, S. & Azad, S. Performance Analysis of IMD High-Resolution Gridded Rainfall (0.25° × 0.25°) and Satellite Estimates for Detecting Cloudburst Events over the Northwest Himalayas. *Journal of Hydrometeorology* **21**, 1549–1569 (2020).
29. Goteti, G. & Famiglietti, J. Extent of gross underestimation of precipitation in India. *Hydrol. Earth Syst. Sci.* **28**, 3435–3455 (2024).
30. Van Den Besselaar, E. J. M., Van Der Schrier, G., Cornes, R. C., Iqbal, A. S. & Klein Tank, A. M. G. SA-OBS: A Daily Gridded Surface Temperature and Precipitation Dataset for Southeast Asia. *J. Climate* **30**, 5151–5165 (2017).
31. Newman, A. J., Clark, M. P., Wood, A. W. & Arnold, J. R. Probabilistic Spatial Meteorological Estimates for Alaska and the Yukon. *J. Geophys. Res. Atmos.* **125** (2020).
32. Caillouet, L., Vidal, J.-P., Sauquet, E., Graff, B. & Soubeyroux, J.-M. SCOPE Climate: a 142-year daily high-resolution ensemble meteorological reconstruction dataset over France (2019).
33. S.C. Bhan, Dr. Ashok Kumar Das, Rahul Saxena & S.K. Manik. Rainfall Statistics of India 2022.pdf. (2023).
34. Tang, G., Clark, M. P. & Papalexiou, S. M. SC-Earth: A Station-Based Serially Complete Earth Dataset from 1950 to 2019. *Journal of Climate* **34**, 6493–6511 (2021).
35. Resche-Rigon, M. & White, I. R. Multiple imputation by chained equations for systematically and sporadically missing multilevel data. *Stat Methods Med Res* **27**, 1634–1649 (2018).
36. Srivastava, A. K., Rajeevan, M. & Kshirsagar, S. R. Development of a high resolution daily gridded temperature data set (1969–2005) for the Indian region. *Atmospheric Science Letters* **10**, 249–254 (2009).
37. Ben Bouallègue, Z. & Richardson, D. S. On the ROC Area of Ensemble Forecasts for Rare Events. *Weather and Forecasting* **37**, 787–796 (2022).
38. Peringiyil, A. & Saharia, M. Indian Precipitation Ensemble Dataset (IPED). *Zenodo* https://doi.org/10.5281/zenodo.8199138 (2024).
39. Gadgil, S. The Indian monsoon: 3. Physics of the monsoon. *Reson* **12**, 4–20 (2007).
40. Joseph, L., Skliris, N., Dey, D., Marsh, R. & Hirschi, J. Increased Summer Monsoon Rainfall Over Northwest India Caused by Hadley Cell Expansion and Indian Ocean Warming. *Geophysical Research Letters* **51**, e2024GL108829 (2024).
41. Tang, G., Clark, M. & Papalexiou, S. EM-Earth: The Ensemble Meteorological Dataset for Planet Earth. Federated Research Data Repository https://doi.org/10.20383/102.0547 (2022).
42. Ma, Z. *et al*. AERA5-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**, E1146–E1171 (2022).
43. Ma, Z. *et al*. AERA5-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 (1999–2015). *Zenodo* https://doi.org/10.5281/zenodo.4264452 (2020).
44. Hunt, K. M. R. & Menon, A. The 2018 Kerala floods: a climate change perspective. *Clim Dyn* **54**, 2433–2446 (2020).

45. Mishra, V. & Shah, H. L. Hydroclimatological Perspective of the Kerala Flood of 2018. *J Geol Soc India* **92**, 645–650 (2018).
46. Ray, K., Bhan, S. C. & Bandopadhyay, B. K. The catastrophe over Jammu and Kashmir in September 2014: a meteorological observational analysis. *Current Science* **109** (2015).

Acknowledgements

This research was conducted in the HydroSense lab (https://hydrosense.iitd.ac.in/) of IIT Delhi, and the authors acknowledge the IIT Delhi High Performance Computing facility for providing computational and storage resources. The authors gratefully acknowledge IMD Pune for allowing access to the datasets. Dr. Manabendra Saharia gratefully acknowledges financial support for this work through grants from Ministry of Earth Sciences/ IITM Pune Monsoon Mission III (RP04574) and Ministry of Earth Sciences DeepINDRA Project (RP04741).

Author contributions

Anagha Peringiyil: Methodology and formal analysis; Data Curation, Writing - Original Draft. Manabendra Saharia: Conceptualization, Methodology, Writing - Review & Editing. Sreejith O. P: Reviewing. Andy Wood: Methodology; Reviewing and editing. Mrutyunjay Mohapatra: Reviewing. Bharti sabde: Datasets. Aradhana Kumari: Datasets. Bhushan Phadkar: Datasets. Sabeerali C. T: Datasets. Rohini P: Datasets. Hosalikar K. S: Datasets. M. Rajeevan: Reviewing.

Competing interests

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41597-025-04474-2.

Correspondence and requests for materials should be addressed to M.S.