scientific reports

OPEN

Accurate and efficient AI-assisted paradigm for adding granularity to ERA5 precipitation reanalysis

Mattia Cavaiola[1]✉, Peter Enos Tuju[2] & Andrea Mazzino[2,3]✉

Scientific inquiry has long relied on deterministic algorithms for systematic problem-solving and predictability. However, the rise of artificial intelligence (AI) has revolutionized data analysis, allowing us to uncover complex patterns in large datasets. In this study, we combine these two approaches by using AI to improve the reconstruction of past precipitation events, which is crucial for understanding climate change. Our objective is to leverage AI to map large-scale atmospheric proxies from the ERA5 climate reanalysis and multi-satellite historical precipitation data from the NASA-IMERG GPM constellation to observed precipitation, enhancing the accuracy and the resolution of climate reanalysis. Accurate climate reanalyses are essential, as they provide the most realistic representations of past atmospheric conditions, serving as benchmarks against which climate models are validated. Our AI-enhanced method offers a more accurate and computationally efficient solution compared to deterministic high-resolution precipitation downscaling methods. Additionally, it shows the capability to generalize predictions to new, previously unobserved locations, making it applicable across various regions. By integrating AI with traditional reanalysis techniques, we open up new opportunities for climate science and geosciences, with the potential to improve the accuracy and reliability of climate data, contributing to a better understanding of climate dynamics.

Keywords Past climate reconstruction, AI and computing in synergy, Hindcasting

Impact models use climate projections to assess climate change impacts. Their validation necessitates high-quality historical meteorological datasets. These datasets play a dual role: they serve as inputs to drive the impact models and as benchmarks to assess their performance during historical periods. Reanalysis datasets have emerged as tools of excellence for these objectives due to their ability to offer a physically consistent global reconstruction of past weather conditions, devoid of spatial or temporal gaps[1]. To achieve this result reanalyses inherently embody the most likely depiction of the atmosphere and ocean conditions, shaped by assimilating available observations and model forecasts from a preceding time-step[1]. Besides climate science, reanalyses are a critical endeavor with far-reaching implications ranging from hydrology[2] and renewable energy-related applications[3], to water resource management[4] and social risks, biodiversity, and ecosystem health[5]. Global reanalysis has made remarkable strides, particularly in the context of global-scale phenomena assessments[6]. Among the contemporary global reanalysis, ERA5, crafted by the European Centre for Medium-Range Weather Forecasts (ECMWF), is the cutting-edge product[7] offering worldwide climate description coverage at a spatial resolution of approximately 0.25^o (equivalent to the unprecedented resolution of about 30 km), on an hourly basis, spanning from 1950 to the present day. ERA5 is the result of a decade of advances in model physics and data assimilation methods capitalized after the release of the predecessor ERA-Interim database[8]. An overview of the main characteristics and general performance of ERA5 and a comparison with ERA-Interim is provided in[7]. Despite the strengths of global reanalysis datasets, ERA5, like other similar high-quality products[9], grapples with a fundamental difficulty: capturing the complex dynamics of local-scale meteorological phenomena[1]. Their precise characterization remains a challenge, a fact that would discourage their direct use for local applications[10], especially those involving rainfall as input[11–14]. At the heart of this challenge lies the necessity for high-resolution downscaling - a computationally expensive technique employed to capture local effects and related phenomena. While downscaling holds the promise of enhancing the spatial resolution of reanalyses[15–21], making them more amenable to local-scale predictions, it introduces a formidable trade-off: the substantial computational burden associated with this process. Table 1 reports some information on relevant examples of European initiatives

[1]CNR-National Research Council of Italy, Institute of Marine Sciences, Via S.Teresa S/N, 19032 Pozzuolo di Lerici, La Spezia, Italy. [2]Department of Civil, DICCA, Chemical and Environmental Engineering, Via Montallegro 1, 16145 Genova, Italy. [3]INFN-Sezione di Genova, Via Dodecaneso 33, 16146 Genova, Italy. ✉email: mattia.cavaiola@sp.ismar.cnr.it; andrea.mazzino@unige.it

Reanalysis (including their deterministic downscaling)	Start date-end date	Spatial resolution	Covered region	Reference paper
ERA5	1940–Present	30 km	Global coverage	Hersbach et al.[7]
UERRA (ERA-Interim downscaling)	1961–2019	11 km (atmosphere), 5.5 km (near-surface)	CORDEX EUR-11 domain	Copernicus CCS[15]
CERRA (ERA5 downscaling)	Early 1980s–Present	5.5 km	Pan-European coverage	Schimanke[16]
COSMO-REA6 (ERA-Interim downscaling)	1995–2015	6 km	CORDEX EUR-11 domain	Bollmeyer et al.[17]
MÉRA (ERA-Interim downscaling)	1981–2019	2.5 km	Ireland, UK, Northern France	Whelan et al.[18]
MERIDA (ERA5 downscaling)	1990–2020	7 km (initial), 4 km (postprocessed)	Italy	Bonanno et al.[19]
BOLAM/MOLOCH (ERA5 downscaling)	1979–2019	2.5 km	Italy	Capecchi et al.[20]
VHR-REA_IT (ERA5 downscaling)	1989–2020	2.2 km	Italy	Raffa et al.[21]

Table 1. Examples of European initiatives bridging the gap between global reanalyses datasets and high-resolution regional-scale reanalysis datasets. Only datasets covering a time span of at least 20 years are reported. Only fully deterministic strategies (i.e. not assisted by AI) have been considered.

on high-resolution regional-scale reanalysis downscaling. The current zenith in downscaling, exemplified by the 2-kilometer resolution reanalysis downscaling, VHR-REA_IT, over Italy by[21], serves as a testament to the challenges of achieving such granularity on a planetary scale. The computational demands and resource-intensive nature of high-resolution downscaling pose practical constraints that limit its applicability on a global scale.

In this landscape, artificial intelligence (AI) emerges as a beacon of hope. Fully data-driven techniques, which have demonstrated exceptional promise in the nowcasting/forecasting realms[22,23], present a tantalizing alternative to the classical deterministic downscaling alluded to above. However, their application in hindcasting faces a fundamental conundrum-dependency on existing reanalyses, such as ERA5, to generate their predicted dynamics. This reliance on the same reanalysis used for training compromises their use to improve upon reanalysis data. However, the remarkable success of AI in predictive tasks is unquestionable, making it an appealing candidate for addressing the gap in local-scale weather reconstruction. Our solution involves harnessing the strengths of existing global reanalyses renowned for their ability to predict large-scale climate patterns, and AI techniques celebrated for their capability to discern intricate relationships between large-scale weather features and localized point observations[10,24]. In essence, our proposed strategy, to be classified as an AI-enhanced strategy, seeks to bridge the chasm between global-scale phenomena captured by state-of-the-art global reanalyses, typified by ERA5, and local-scale phenomena, which still remain inadequately characterized. Our strategy demonstrates the ability to overcome common limitations seen in recent AI applications on reanalysis datasets. Specifically, it does not require costly downscaling from ERA5 using deterministic models[25]. Instead, the downscaling is performed directly from ERA5, bypassing the need for intermediate, computationally intensive steps. The low computational cost of our approach allows for the consideration of much longer training and testing periods compared to more complex and resource-demanding AI networks[26]. This extended period of data usage ensures a more robust assessment of the performance of the developed strategy, as it can be tested over a wider range of temporal conditions and scenarios, providing greater confidence in its applicability and accuracy. The proposed approach strikes an effective balance between computational efficiency and high performance, making it a competitive alternative to traditional, more computationally expensive downscaling methods.

In a recent article featured in Nature Climate Change[27], authors put forth relevant recommendations: they advocate for a balanced approach, emphasizing the importance of harnessing advances in computing and AI to enhance climate modeling accuracy and reliability. In the present paper, we have embraced these recommendations. Our approach aims to unlock the potential of merging the best of two worlds: the skills of the deterministic approach at the basis of all available reanalyses in capturing global-scale weather/climate patterns, and the predictive prowess of AI in extrapolating these relationships to the local scale. Our primary conclusion posits that there is no imperative necessity to attain exceptionally high-resolution reanalyses via dynamical downscaling strategies.

The logic circuit of our AI-enhanced strategy

In our AI-enhanced strategy, precipitation prediction involves two distinct steps sketched in Fig. 1. The first step is a binary classification task that predicts the occurrence (event labeled by '1') or absence of precipitation (event labeled by '0') obtained from a predicted probability of occurrence upon a threshold selection. This step is executed using a neural network classifier that has been appropriately trained and validated through k-fold cross-validation (see next sections for details). The performance assessment phase of the classifier was conducted using the Precision and Recall indices, along with their derived metrics, as defined in the Methods section. It is worth noting that the Precision index is also known as the 'success ratio' (SR), while the Recall index is commonly referred to as 'sensitivity' or 'probability of detection' (POD).

The second step of our strategy follows the first one, involving an additional neural network, now working as a regressor, trained specifically on True Positives (TP) identified by the classifier in the training set, and outputting the accumulated precipitation on different accumulation periods. The regressor will be tested on the observed wet conditions. We have in this way a test set common to all models/reanalysis, which is not the case if, e.g., 'True Positives' were chosen for testing of the prediction models.. The decision threshold in probability to identify events '0' and '1' is selected downstream the classification output to maximize the so-called F1-score (defined in the Methods section as the harmonic mean of Precision and Recall indices) in the training set.

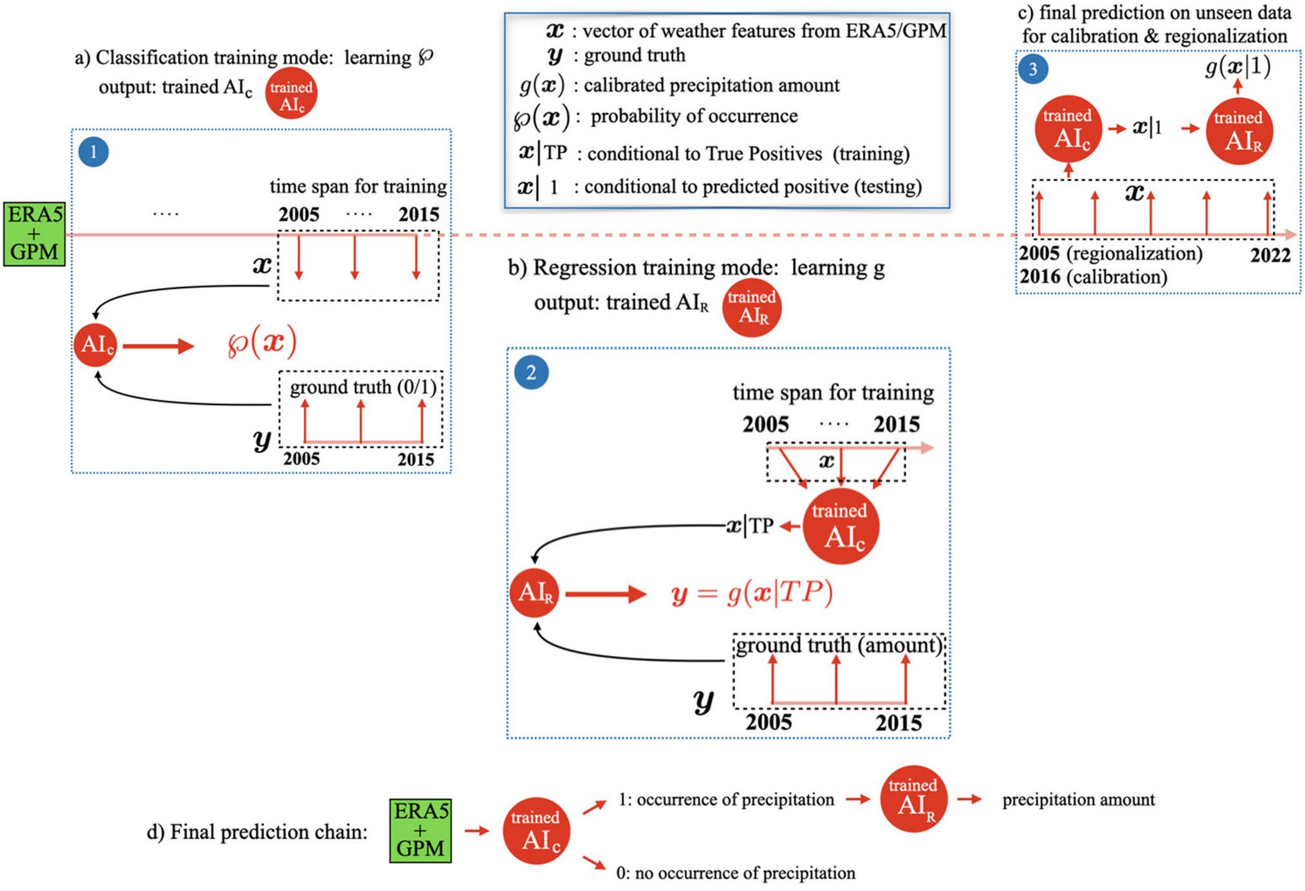
a) Classification training mode: learning ℘
output: trained AI_c
trained AI_c
1
ERA5 + GPM
time span for training
2005 2015
x
AI_c
℘(x)
ground truth (0/1)
y
2005
2015
x : vector of weather features from ERA5/GPM
y : ground truth
g(x) : calibrated precipitation amount
℘(x) : probability of occurrence
x|TP : conditional to True Positives (training)
x| 1 : conditional to predicted positive (testing)
b) Regression training mode: learning g
output: trained AI_R
trained AI_R
2
time span for training
2005 2015
x
trained AI_c
x|TP
AI_R
y = g(x|TP)
ground truth (amount)
y
2005
2015
c) final prediction on unseen data for calibration & regionalization
3
g(x|1)
trained AI_c
x|1
trained AI_R
x
2005 (regionalization)
2016 (calibration)
2022
d) Final prediction chain:
ERA5 + GPM
trained AI_c
1: occurrence of precipitation
trained AI_R
precipitation amount
0: no occurrence of precipitation

Fig. 1. Logic circuit of our AI-enhanced precipitation prediction strategy for different accumulation intervals. The green box represents the features provided by ERA5 reanalysis and GPM satellite data in the period 2005-2022 serving for training, validation, and testing. Panel 1 illustrates the training stage of the classifier, panel 2 represents the training stage of the regressor that operates downstream of the classifier, specifically trained on the true positive events. Both the classifier and regressor, once trained, work in synergy during the testing period (panel 3) from 2016 to 2022 (for calibration) and from 2005 to 2022 (for regionalization). To convert predicted probabilities into class labels (0 for absence and 1 for occurrence of precipitation), an optimal threshold value has been determined. Rather than assuming an equal division with a fixed threshold of 0.5, we optimized the threshold using the F1 score on the training set. This approach ensures a balanced and effective conversion of probabilities into class labels.

These two neural networks are fed with features from both ERA5 and the Satellite Precipitation Estimate (SPE) from the Integrated Multi-satellite Retrievals for the Global Precipitation Measurement Mission[28] (GPM in short, with a spatial resolution of approximately 10 km). Table 2 reports the complete list of used features. Together, the two networks serve as the framework for transitioning from large-scale convection-triggering precipitation proxies to actual observed precipitation.

Three networks are trained for three distinct precipitation accumulation intervals: 1 hour, 12 hours, and 24 hours.

For all reanalysis products, we have chosen not to apply any bias correction. Several reasons underlie this decision. Firstly, due to the absence of a standardized method for correcting precipitation bias, any corrective approach would lead to results highly dependent on the specific correction method employed. Conversely, by refraining from correction, the results obtained unambiguously reflect the performance of the three considered products against our AI network.

Another compelling reason for not bias-correcting is a practical one: in many applications, end-users employ products as they are, without implementing any bias correction. This raises the question of whether an AI network can effectively map features from models that may potentially be affected by bias onto observed data. Our results clearly demonstrate the feasibility of this, greatly simplifying the use of these reanalysis tools. They can be considered as they are without sacrificing a high level of accuracy in the results obtained.

Certainly, there remain open questions regarding how our strategy could further improve by incorporating features after bias removal. We are confident that the results we present will inspire researchers to explore this aspect in future research endeavors.

List of ERA5 features used to train the AI networks
CAPE index (convective available potential energy)
cp - convective precipitation
$u100$ - u component wind velocity at 100m
$v100$ - v component wind velocity at 100m
$u10$ - u component wind velocity at 10m
$v10$ - v component wind velocity at 10m
$d2m$ - dew point at 2m
$T2m$ - air temperature at 2m
blh - boundary layer height
$ucdv$ - cloud base height
z - geopotential height
kx - K-index
lsp - large scale precipitation
msl - mean sea-level pressure
$sshf$ - surface sensible heat flux
tcc - total cloud cover
$vimd$ - vertically integrated moisture divergence
D - divergence at 300, 500, 700, 750, and 850 hPa
PV - potential vorticity at 300, 500, 700, 750, and 850 hPa
R - relative humidity at 300, 500, 700, 750, and 850 hPa
U - u component wind velocity at 300, 500, 700, 750, and 850 hPa
V - v component wind velocity at 300, 500, 700, 750, and 850 hPa
T - air temperature at 300, 500, 700, 750, and 850 hPa
W - vertical velocity at 300, 500, 700, 750, and 850 hPa
Q - specific humidity at 300, 500, 700, 750, and 850 hPa

List of GPM features used to train the AI networks*
Lca - GPM-Late run Multi-satellite precipitation estimate with gauge calibration
Lun - GPM-Late run Multi-satellite precipitation estimate
Lir - GPM-Late run Infrared (IR) only precipitation estimate
Eca - GPM-Early run Multi-satellite precipitation estimate with gauge calibration
Eun - GPM-Early run Multi-satellite precipitation estimate
Eir - GPM-Early run Infrared (IR) only precipitation estimate

Other features used to train the AI networks
$\cos(2\pi H/24)$** - H being the hour of the day
$\sin(2\pi H/24)$**- H being the hour of the day
$\cos(2\pi m/12)$ - m being the month of the year
$\sin(2\pi m/12)$ - m being the month of the year
h - height above the sea level

Table 2. The features used in training our AI-based networks, as extracted from ERA5 and GPM datasets, are reported. ERA5 and GPM features are extracted hourly and utilized without modification when predicting 1-h accumulated precipitation. For 12-h and 24-h accumulated precipitation predictions, these features are averaged over 12-h and 24-h time windows, respectively. For the GPM features, we considered the four values associated with the four closest grid points surrounding the stations. For the ERA5-extracted features, we limited the extraction to the nearest grid point to the stations.

Target areas and observed data for training and testing

We have utilized official open data provided by two Regional Agencies for the Protection of the Environment (ARPA): Arpa Liguria and Arpa Lombardia. These two public Italian bodies manage data within the Liguria and Lombardia regions in north-west Italy (refer to Fig. 2 for an illustration of the covered area). The choice of these two regions is motivated by our desire to tackle the challenge presented by regions with highly diverse climatology. Liguria, which borders the sea, experiences a Mediterranean climate, although it is not uniform due to the rugged terrain, which is mostly mountainous. On the other hand, Lombardia, lacking coastal access, features a continental climate in the Po Valley, transitioning to a predominantly alpine climate in the northern region bordering Switzerland.

With such distinct characteristics, these two regions pose a significant challenge for both purely deterministic downscaling models and our AI-enhanced strategy in accurately predicting precipitation. Their climatic diversity, coupled with the fact that these stations are not assimilated into ERA5, makes them an ideal testing ground for our approach. In Liguria, we considered a total of 71 stations (red circles in Fig. 2, region L1 in short), while the number of stations in Lombardia amounted to 268 (blue circles in Fig. 2, region L2 in short). The time span for regions L1 and L2 covers years from 2005 to 2022 on an hourly basis. All selected stations guarantee a coverage of hourly accumulated precipitation of at least 95 % for the entire period of 18 years.

The benchmark products we considered for the precipitation field include the two reanalyses ERA5[7] (with a spatial resolution of approximately 30 km) and VHR[21] (with a spatial resolution of approximately 2 km), as well as the GPM. The features used to train and test our AI network have been extracted solely from ERA5 and GPM (see Table 2 for the complete list of extracted features). In addition, to further compare the performance of our AI strategy, we employed baseline supervised machine learning algorithms such as Logistic Regression[29] (LR) for the classification task and multilinear ridge regression[30] (MLR) for the regression stage (see the Methods section for details on the implementation of LR and MLR).

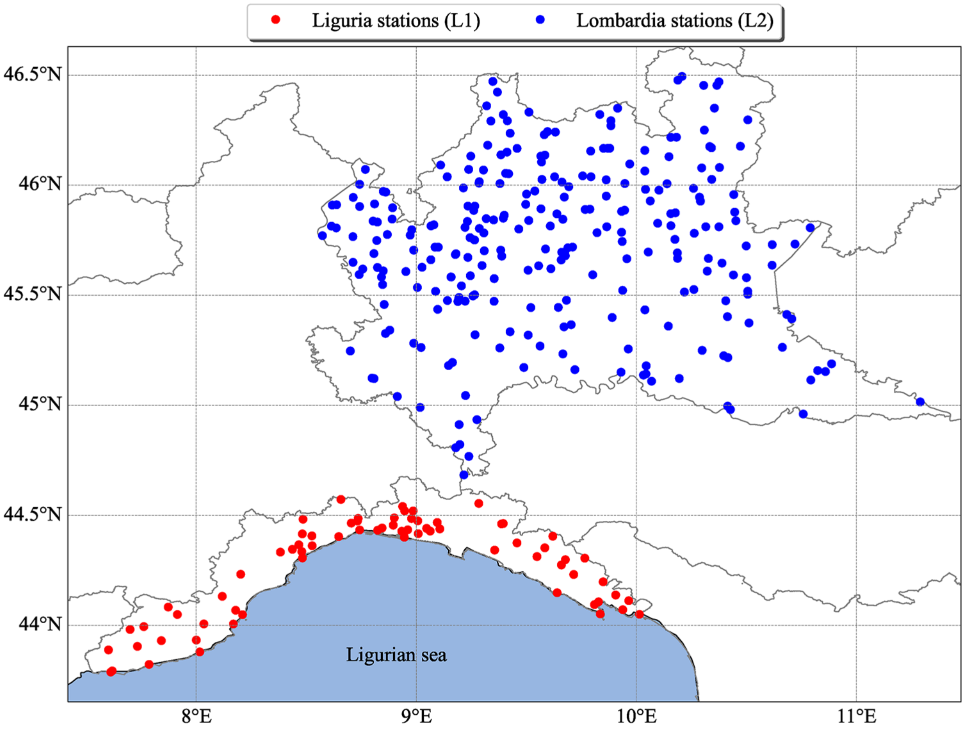

Fig. 2. The two official networks of rain gauge stations used in the present study are as follows: red bullets represent region L1 (Liguria, 71 rain gauges), and blue bullets represent region L2 (Lombardia, 268 rain gauges). The rectangular panel measures approximately 400 km in longitude and 300 km in latitude and is situated in the northwestern part of Italy.

Calibration and regionalization via k-fold cross-validation

In the geographic region L1 (Fig. 2), the set of 71 stations was divided into 5 folds created by randomly selecting different stations without repetition. Each fold contained approximately 20% of the total number of stations. A further fold has been constructed (fold 6) containing all stations of region L2. We conducted three types of validation: i) validation to assess the predictive capabilities of the neural network on the same stations as in the training set, albeit at different time intervals not present in the training data. This validation step involves a sort of calibration of the ERA5 and GPM products at individual rain gauge stations. ii) Verification of the network's ability to provide predictions for locations different from those used in the training set, even if belonging to the same geographic area L1 considered in the training stage. This validation step involves an extrapolation (defined here as regionalization) toward unseen locations. iii) Verification of the network's ability to provide predictions for locations not belonging to the same region (L1) of the training set. This validation step entails extrapolating to an entirely new region, L2, through the application of the transfer learning strategy (see the next section) exclusively for the regression task. During this process, the classifier remains as initially trained in region L1, without undergoing retraining or modification for L2.

The k-fold cross-validation strategy facilitated the assessment of both i), ii), and iii) within a single loop. In this loop, each of the 5 folds was sequentially designated as the test set for assessing ii), and paired with fold 6 to assess iii). The remaining 4 folds were used for training and for testing property i).

From each training set composed of 4 folds, a subset of samples was randomly selected, constituting 20% of the total training dataset. This validation set, never used for testing, serves for both the network hyperparameter tuning and the selection of the probability decision threshold by maximizing the F1 score defined in the Methods section. Selecting the threshold is required to identify labels '1' and '0' corresponding to the occurrence/absence of precipitation.

The training set encompassed data from the years 2005 to 2015. For the evaluation of i), the years from 2016 to 2022 were reserved for testing. To assess ii) and iii), the test set spanned the years from 2005 to 2022 thank to the fact that the testing stations were different from those used for training.

The sketch of the whole strategy is reported in Fig. 1. Note that, although our study focuses on land precipitation without considering marine areas, we chose to use ERA5 as our reference reanalysis dataset rather than its higher spatial resolution version (approximately 10 km), known as ERA5-Land[31] which provides information only for land, starting from 1950. The reason for this choice is that we want to keep the possibility open for future extensions of our present study to marine areas. Extending to marine areas would allow us to investigate whether different convection-triggering mechanisms, active over land and over sea, may lead to varying levels of accuracy for the different strategies analyzed.

Transfer learning from region L1 to L2 for the regressor

Transfer Learning (TL) is a powerful method[32] in machine learning that involves transferring knowledge from one domain (the source domain, here L1) to another (the target domain, here L2). This technique is particularly useful when the target domain has insufficient data to train a robust model from scratch. By leveraging the patterns and insights learned from a related, well-understood source domain, TL allows for improved learning efficiency and performance in the target domain with minimal data requirement.

In the present study, we have utilized TL to enhance the regression tasks within region L2, building upon a classifier initially trained on region L1 data. For classification in L2, we have employed the L1-trained classifier directly, without applying TL, to ascertain its generalization capability across different climatic and geographical zones. Specifically, we have randomly selected 10% of the stations in region L2 to serve dual purposes: firstly, for performing inference with the L1 classifier over the period 2005-2015 to identify True Positives (TP) in region L2; secondly, this same subset was utilized for TL during the regression phase on the identified TP to refine our model's understanding and prediction accuracy concerning L2's climate behaviors. The training of the regressor has been conducted over 30 epochs. The inference phase of the regressor has then been carried out on the remaining 90$\%$ of the L2 stations for the period 2005-2022, ensuring that our model could accurately generalize to the broader region L2.

The robustness of our findings has been enhanced by employing the same k-fold cross-validation procedure exploited in region L1, and the results presented have been averaged across the outputs of 5 models generated through this process. This approach based on TL points to demonstrate the efficacy of TL in leveraging limited data points for significant predictive advancements in climatological studies.

As a final test of the TL approach's robustness, we created an identical twin model. The only difference is that in this twin model, the weights are initialized randomly, rather than being pre-trained on a source task. This way, the twin model does not benefit from prior knowledge, allowing us to directly measure the impact of the TL approach.

Results

As our AI-enhanced strategy is designed to predict precipitation in two steps - classification followed by regression - evaluating the effectiveness of our approach as a whole involves a separate assessment of the skills of the classifier and the regressor using the performance indices outlined in the Methods section.

Assessing the classification stage: region L1

Figure 3 displays the trends of Precision-Recall (P-R) curves along with their respective values of the area under the P-R curve (AUC) index (see Methods section) for the average of the 5 L1-folds (represented by blue curves) from the cross-validation. For the sake of comparison, we have reported in red the analogous curves from the application of the simpler logistic regression (LR).

The P-R curves have been determined across three accumulation precipitation intervals (1 hour, 12 hours, and 24 hours) and for both 'calibration' and 'regionalization'. The dotted horizontal line represents prevalence-based prediction (see Methods section) and serves as a useful benchmark.

AUC evaluates the classifier's ability to discriminate between precipitation occurrence and absence of precipitation. It measures how well the classifier can rank instances in terms of their predicted probabilities. A high AUC indicates that the classifier is effective at distinguishing between positive and negative instances. Our AI-enhanced prediction model exhibits this highly desirable characteristic, especially when compared to the prevalence-based random model, whose AUC is always largely below 50$\%$. Although the results from the simple logistic regression (LR) are not as strong as those from our deep learning network, they are still quite reasonable, likely due to the model's ability to capture key relationships in the data with fewer parameters and less risk of overfitting.

In general, a classifier can excel in its discriminatory power being not reliable in predicting the associated probability of precipitation occurrence. To assess the reliability of a probabilistic classifier we resort to the reliability diagrams (see Methods section) reported in Fig. 4. The x-axis represents the mean predicted probability within each bin and the y-axis represents the observed frequency of events. Ideally, the points on the reliability diagram should fall along or close to a diagonal line (the 'perfectly calibrated' line), which indicates that the predicted probabilities are accurate estimates of the true event probabilities. Our predicted probability (blue lines) clearly exhibits this highly desirable property for both calibration (left column of Fig. 4) and regionalization (right column), and for all considered precipitation accumulation intervals. Also important to emphasize is the robustness of this conclusion (not shown), which is valid for all considered L1-folds of the cross-validation. Although the calibration level of the logistic regression is lower than that of the more complex network we have developed, its reliability remains competitive, especially considering the simplicity of the LR approach.

The reliability diagrams are accompanied by the sharpness diagrams as insets, with reference to the sole fold 3 for the sake of readability. Because our AI-enhanced strategy predicts probabilities that are concentrated and clustered around specific probability levels, especially for the 24-h accumulation period, our model exhibits sharpness. In other words, when our model assigns probabilities to the two mutually exclusive events of precipitation occurrence/absence, it does so with a high degree of confidence. A similar conclusion applies to the simpler LR strategy, although it yields lower sharpness.

We can now proceed to analyze the classification skills of our strategy against the precipitation estimates provided from GPM, ERA5, its 2-km dynamical downscaled reanalysis VHR, and the benchmark based on the LR algorithm. The results are summarized in Table 3, where Precision, Recall, F1 score, and Accuracy are shown for all the analyzed cases and models. To calculate these indices, it was necessary to set a decision threshold on the probability to identify the occurrence of precipitation. This threshold was chosen based on the requirement that

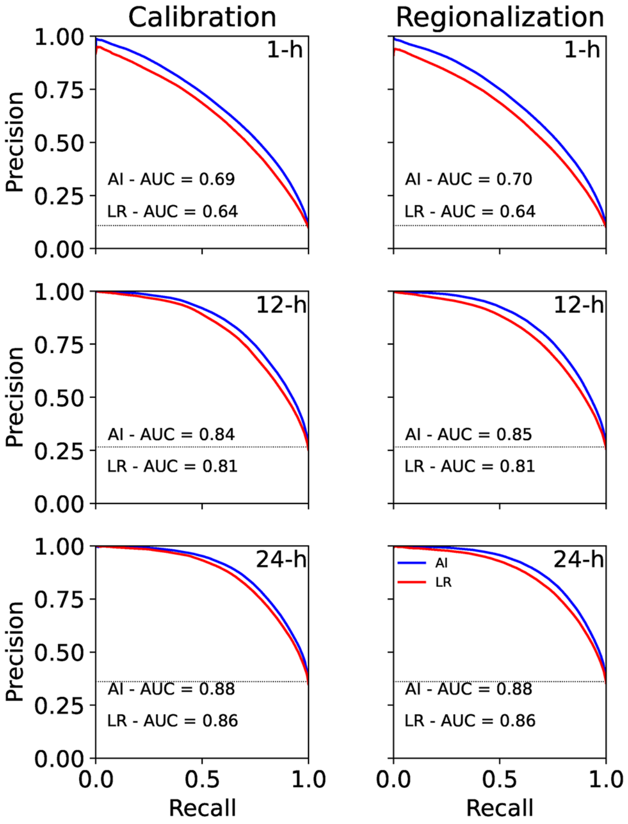

Fig. 3. Precision-Recall (P-R) curves for the average of the 5 L1-folds. The blue lines refer to the results from our AI strategy; the red lines refer to the results from the LR algorithm. Different rows refer to different accumulation precipitation intervals; different columns refer to calibration and regionalization. Each panel also displays the AUC index of the P-R curve along with the corresponding value derived from the prevalence-based random model (dotted line), which serves as a benchmark.

it maximizes the F1 score on the validation dataset described in the subsection "Calibration and regionalization via k-fold cross-validation".

It is worth noting that we cannot compare the AUC score of our strategy with those of the other considered deterministic products as the latter, being purely deterministic, provide single values for Recall and Precision. The adaptability represents a distinct advantage of the approaches based on AI, setting it apart from purely deterministic models that lack this level of flexibility. For instance, when dealing with extreme precipitation events, it is more advantageous to choose a threshold that maximizes Recall rather than Precision, a fact possible with our strategy by fine-tune the decision threshold while this is not for deterministic approaches.

The results of Table 3 clearly show that our strategy outperforms the other approaches with the LR algorithm still delivering competitive results in this case as well. A strength of our strategy emerges both in terms of achieving a higher Accuracy and a superior F1 score. The high Accuracy obtained by virtually all models for the 1-hour accumulated precipitation is primarily a result of the significant dataset imbalance (smaller than 10%), and therefore, it does not accurately reflect the true measure of model performance. It is important not to be misled by the higher Recall exhibited by ERA5, as it is accompanied by a significantly lower Precision value. To emphasize this point further, we intentionally set the decision threshold for probabilities to yield the same Precision value as ERA5 in each of the 5 folds, and similarly as the Precision values of GPM, VHR, and LR. The resulting Recall values are presented in Table 4 and denoted as Recall_1 (at the same Precision value as GPM, table's raw 2); Recall_2 (at the same Precision value as ERA5, table's raw 3); Recall_3 (at the same Precision value as VHR, table's raw 4); Recall_4 (at the same Precision value as LR, table's raw 5). Notably, these values are now higher than those of GPM, ERA5, VHR, and LR confirming the skills of our strategy which for both 12 and 24-hour accumulated precipitations reaches Recall values close to 90%.

Assessing the classification stage: region L2

All indices presented in Tables 3 and 4 for the L1 region have been calculated for the L2 region as well via Transfer Learning applied to the sole regressor This means that, at the classification level, we use the network entirely trained in region L1 for predictions in region L2. The corresponding indices for the L2 region are summarized in Tables 5 and 6.

From these two tables, it is evident that the added value of our strategy persists when compared to all the benchmarks considered in region L2, including the benchmark based on the LR algorithm. The results of both

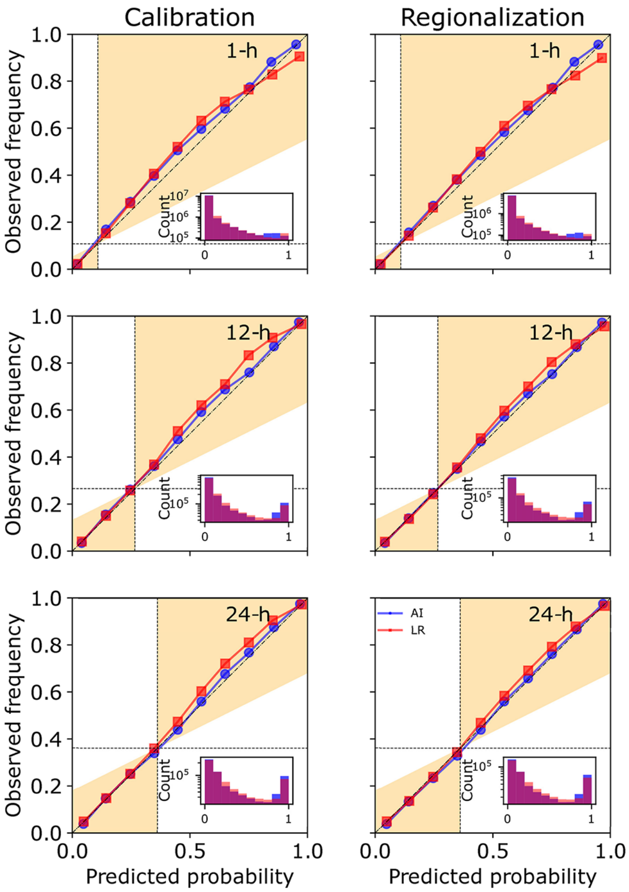

Fig. 4. The reliability diagrams are displayed for the average of the 5 L1-folds of the cross validation for assessing both 'calibration' (left column) and 'regionalization' (right column). Different rows depict different precipitation accumulation intervals from 1 hour (1st row) to 24 hours (3rd row) . Blue lines refer to our AI strategy while red lines refer to the logistic regression, LR. The shaded regions highlight areas where forecasts demonstrate skill. In addition, the horizontal and vertical lines show the prevalence-based probabilities of the event for forecasts and observations. Furthermore, the two smaller diagrams included as insets depict the sharpness diagrams (in reference to the fold 3), illustrating the relative frequency of precipitation occurrence predictions (abbreviated as 'Count') at different probability levels (x-axis of the inset). Colors are coded as in the reliability diagrams.

tables thus reflect the generalization ability of our network used for the classification step, which entirely learns in the L1 region and also produces skillful predictions in the L2 region.

Assessing the regression stage: region L1

Having demonstrated the robustness and generalization capability of the classifier component of our strategy, along with its reliability and significantly higher accuracy compared to both a simple prevalence-based random model, the LR algorithm (see Methods section) and the considered reanalyses, including the 2-kilometer high-resolution product VHR, we now proceed to analyze the performance of the regression component. For this analysis, we have computed robust statistical indices commonly used to assess the skills of a regressor. Namely, the normalized root mean square error (NRMSE) and the Pearson correlation coefficient, both defined in the Methods section. All of these indices have been calculated on the dataset of the observed wet conditions (i.e. the observed '1' events). Figure 5 reports for the calibration task the skill score of NRMSE (1st panel row) and correlation coefficient (2nd panel row) for all considered folds and different accumulation intervals (panel columns). The skill score is a measure used to evaluate how much better (or worse) a prediction is compared to a reference or baseline prediction. The definition of this index is reported in the Methods section. Different colors are used to identify the selected reference model. In all considered cases, our AI strategy outperforms all benchmarks, with the skill score larger than zero, significantly for the correlation coefficient who reaches values close to 60%, and remaining nearly constant across different folds. The superiority of our AI strategy over the simpler multilinear regression (MLR) is evident especially in relation to the correlation coefficient. The reason for this superiority may be due to the ability of the AI model to capture complex, non-linear relationships between variables, which the MLR model cannot. While MLR assumes linearity and limited interactions, the AI model can adapt to more intricate patterns in the data, leading to better predictive skills.

Metric	Task	1 h					12 h					24 h				
		AI	GPM	ERA5	VHR	LR	AI	GPM	ERA5	VHR	LR	AI	GPM	ERA5	VHR	LR
Precision	Calibration	**0.64**	0.43	0.31	0.52	0.61	**0.78**	0.58	0.46	0.68	0.75	**0.80**	0.64	0.55	0.74	0.78
	Regionalization	**0.64**	0.42	0.30	0.52	0.60	**0.77**	0.56	0.47	0.68	0.73	**0.80**	0.62	0.55	0.74	0.77
Recall	Calibration	0.62	0.50	**0.81**	0.41	0.59	0.72	0.64	**0.86**	0.56	0.70	0.76	0.71	**0.89**	0.62	0.76
	Regionalization	0.64	0.53	**0.81**	0.42	0.61	0.74	0.67	**0.87**	0.56	0.72	0.78	0.74	**0.89**	0.62	0.76
F1	Calibration	**0.63**	0.46	0.44	0.46	0.60	**0.75**	0.61	0.60	0.61	0.72	**0.79**	0.67	0.68	0.67	0.78
	Regionalization	**0.64**	0.47	0.44	0.46	0.60	**0.75**	0.61	0.61	0.61	0.72	**0.80**	0.67	0.68	0.67	0.77
Accuracy	Calibration	**0.93**	0.88	0.80	0.90	0.92	**0.88**	0.79	0.71	0.81	0.86	**0.85**	0.76	0.70	0.78	0.84
	Regionalization	**0.93**	0.90	0.79	0.90	0.92	**0.88**	0.78	0.71	0.81	0.86	**0.85**	0.74	0.70	0.78	0.84

Table 3. Evaluation of the classification skills of our strategy against GPM, ERA5, its 2-km dynamical downscaled reanalysis VHR, and LR for the L1 region. L1-fold-averaged Precision, Recall, F1 score, and Accuracy are displayed for all the analyzed cases and models. Best results are highlighted in bold

Metric	Task	1 h					12 h					24 h				
		AI	GPM	ERA5	VHR	LR	AI	GPM	ERA5	VHR	LR	AI	GPM	ERA5	VHR	LR
Recall$_1$	Calibration	**0.83**	0.50	–	–	–	**0.82**	0.64	–	–	–	**0.89**	0.71	–	–	–
	Regionalization	**0.84**	0.53	–	–	–	**0.90**	0.67	–	–	–	**0.91**	0.74	–	–	–
Recall$_2$	Calibration	**0.92**	–	0.81	–	–	**0.94**	–	0.86	–	–	**0.94**	–	0.89	–	–
	Regionalization	**0.92**	–	0.81	–	–	**0.94**	–	0.87	–	–	**0.95**	–	0.89	–	–
Recall$_3$	Calibration	**0.75**	–	–	0.41	–	**0.80**	–	–	0.56	–	**0.82**	–	–	0.62	–
	Regionalization	**0.76**	–	–	0.42	–	**0.82**	–	–	0.56	–	**0.84**	–	–	0.62	–
Recall$_4$	Calibration	**0.65**	–	–	–	0.59	**0.75**	–	–	–	0.70	**0.78**	–	–	–	0.76
	Regionalization	**0.68**	–	–	–	0.61	**0.78**	–	–	–	0.72	**0.80**	–	–	–	0.76

Table 4. For the L1 region, the Recall index was computed using our AI-enhanced strategy while fixing the Precision at the values corresponding to GPM (Recall$_1$, row 2), ERA5 (Recall$_2$, row 3), VHR (Recall$_3$, row 4) , and LR (Recall$_5$, row 5). Best results are highlighted in bold

Metric	1 h					12 h					24 h				
	AI	GPM	ERA5	VHR	LR	AI	GPM	ERA5	VHR	LR	AI	GPM	ERA5	VHR	LR
Precision	**0.62**	0.42	0.31	0.45	0.54	**0.71**	0.54	0.46	0.62	0.67	**0.74**	0.59	0.55	0.68	0.72
Recall	0.63	0.49	**0.82**	0.54	0.60	0.77	0.70	**0.88**	0.74	0.73	0.81	0.78	**0.90**	0.80	0.75
F1	**0.63**	0.45	0.45	0.49	0.57	**0.74**	0.61	0.61	0.67	0.69	**0.77**	0.67	0.68	0.73	0.74
Accuracy	**0.92**	0.87	0.78	0.87	0.90	**0.86**	0.76	0.70	0.80	0.83	**0.83**	0.72	0.69	0.78	0.80

Table 5. Evaluation of the classification skills of our strategy against GPM, ERA5, its 2-km dynamical downscaled reanalysis VHR , and LR for the L2 region. L2-fold-averaged Precision, Recall, F1 score, and Accuracy are displayed for all the analyzed cases and models. Best results are highlighted in bold

Metric	1 h					12 h					24 h				
	AI	GPM	ERA5	VHR	LR	AI	GPM	ERA5	VHR	LR	AI	GPM	ERA5	VHR	LR
Recall_1	**0.83**	0.49	–	–	–	**0.89**	0.70	–	–	–	**0.90**	0.78	–	–	–
Recall_2	**0.91**	–	0.82	–	–	**0.92**	–	0.90	–	–	**0.93**	–	0.90	–	–
Recall_3	**0.81**	–	–	0.54	–	**0.84**	–	–	0.74	–	**0.85**	–	–	0.80	–
Recall_4	**0.72**	–	–	–	0.60	**0.80**	–	–	–	0.73	**0.82**	–	–	–	0.75

Table 6. For the L2 region, the Recall index was computed using our AI-enhanced strategy while fixing the Precision at the values corresponding to GPM (Recall_1, raw 2), ERA5 (Recall_2, raw 3), VHR (Recall_3, raw 4) and, LR (Recall_5, row 5). Best results are highlighted in bold

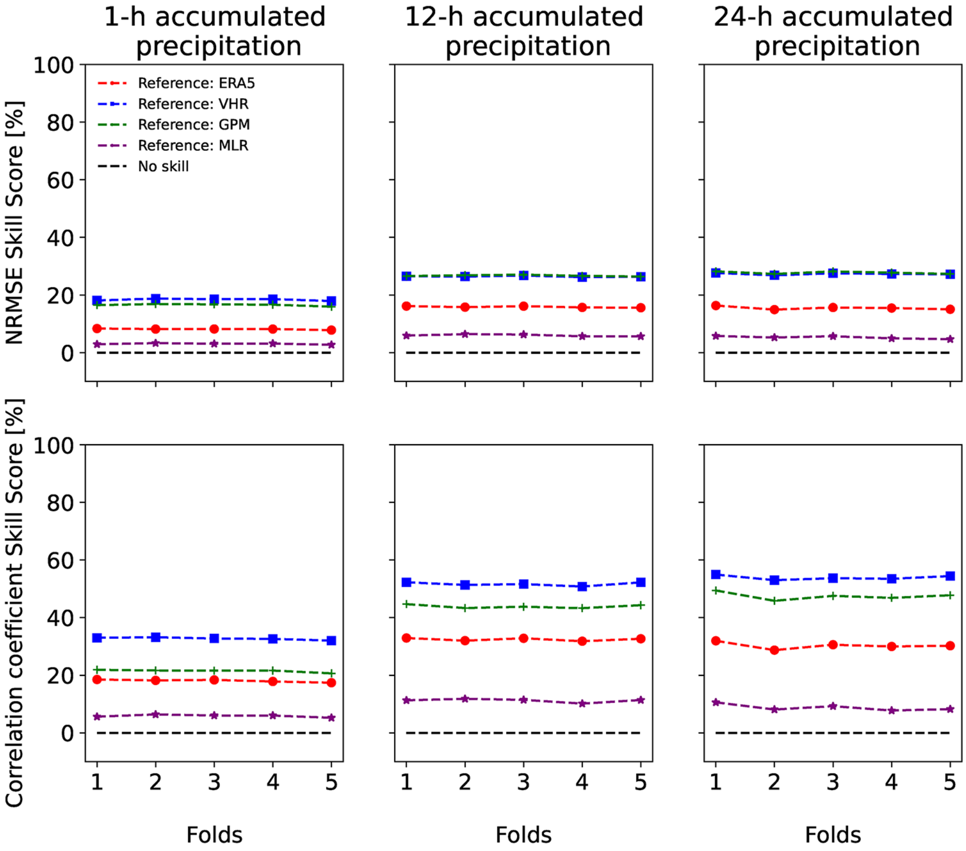

Fig. 5. For the calibration task in region L1, the skill score of NRMSE (1st panel raw) and correlation coefficient (2nd panel raw) are shown for all considered folds and different accumulation intervals (panel columns). Different line colors identify the different selected reference models.

To further assess the network's capability to predict in locations not part of the training set (regionalization), for the region L1, we have presented in Fig. 6 the skill score station by station, for NRMSE (top row of the panel) and correlation coefficient (bottom row of the panel). The resulting skill scores being largely positive for all the considered rain-gauge stations of region L1, this figure confirms our AI-network's ability to generalize, corroborating the findings discussed earlier when analyzing Fig. 5. Although the MLR algorithm consistently delivers lower performance compared to the more complex strategy we have developed, its value should not be underestimated due to its relative simplicity of implementation.

In Table 7 a detailed quantitative analysis reveals a seemingly contradictory observation: the skill scores calculated using GPM, ERA5, VHR, and MLR as a reference are notably higher for 'regionalization' than for 'calibration'. This conclusion holds true for the three accumulation intervals we have considered.

This implies that the network demonstrates greater predictive skills when it is trained on data from neighboring stations rather than solely on historical data from the specific station in question. However, this contradiction is only apparent. During the regionalization process, when the network makes predictions for a particular station at a specific time, it incorporates information from the corresponding time frame of neighboring stations, acquired in the training phase. This information, despite being from adjacent stations, can contribute valuable insights relevant to the inference on the target station. The answer to this paradox lies in the spatial correlations between neighboring stations. Such an observation underscores the potential advantages of employing convolutional architectures in future upgrades of the proposed AI network. While the current network does not employ convolutional layers, these architectures could prove adept at extracting spatial features, thereby advancing beyond the approach of point-wise learning. This promising avenue opens up the prospect of further exploration into convolutional architectures in future work, aiming to harness their full potential for enhancing spatial feature recognition in AI-driven climate reconstruction.

Assessing the regression stage: region L2

The stations considered in Fig. 6 belong to the same geographic area (L1) used for the network training. To rigorously assess and showcase the generalization capability of the employed AI strategies, we have employed transfer learning (TL) to evaluate the adaptability of our approach on stations within region L2. It is critical to clarify that while the classifiers were developed without exposure to L2 stations during their training phase, ensuring a stringent assessment of their generalization potential across new and climatically diverse environments, the strategy for the regressors was more nuanced. Specifically, for the regressors (both based

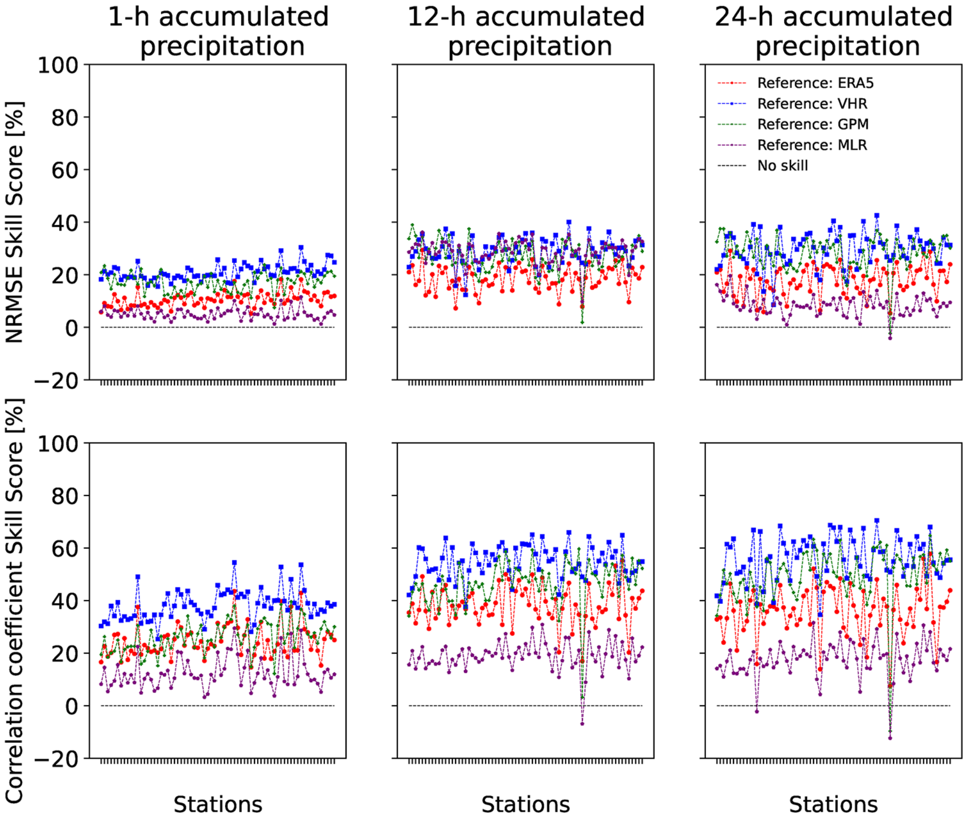

Fig. 6. For the regionalization task in L1, skill scores are computed station by station for those not used in the AI training stage, considering NRMSE (top row of the panel) and correlation coefficient (bottom row of the panel). ERA5, VHR, GPM , and MLR are used as reference predictions, indicated by different colors. The analysis refers to years from 2005 to 2022 (2020 for VHR). The stations are arranged from left to right, starting with those at lower elevations and progressing to those at higher elevations.

		Calibration		Regionalization	
		SS NRMSE [%]	SS Corr [%]	SS NRMSE [%]	SS Corr [%]
1-h accum.	GPM	17	21	17	26
	ERA5	8	18	11	25
	VHR	18	31	21	37
	MLR	3	6	5	12
12-h accum.	GPM	27	44	27	47
	ERA5	16	32	18	38
	VHR	25	48	28	53
	MLR	6	11	29	19
24-h accum.	GPM	28	47	28	50
	ERA5	15	30	18	36
	VHR	25	49	29	54
	MLR	5	9	8	18

Table 7. For region L1, and across the three precipitation accumulation intervals, we have computed the averaged skill scores (across the 5 L1-folds and the L1 stations) for the Normalized Root Mean Square Error (SS NRMSE) and the correlation coefficient (SS Corr) associated to our AI-based predictions. Skill scores have been obtained taking GPM, ERA5, VHR , and MLR as reference predictions (different rows in the table). Averages have been obtained from the skill scores we have presented in Fig. 5 (for calibration) and in Fig. 6 (for regionalization).

on our AI strategy and on a simple MLR), we adopted transfer learning by incorporating a targeted 10 $\%$ of randomly selected L2 stations into the training. This dual-faceted approach, leveraging both unexposed and partially exposed training scenarios, allows us to highlight the efficacy of transfer learning in bolstering model's capability to make accurate predictions in previously unseen territories, thereby affirming the robustness and adaptability of our analytical framework.

The results are shown in Fig. 7 for the five combinations of training sets discussed in the sub-section "Calibration and regionalization via k-fold cross-validation". This figure is analogous to Fig. 6, with the key distinction that it represents an even more stringent extrapolation test. It is remarkable that, in spite of that, the resulting skill scores are overwhelmingly positive for the vast majority of rain-gauge stations. This observation emphasizes that, despite the inherent limitations imposed by using only 10 $\%$ of the stations in the transfer learning process of the regressor alone, our neural network demonstrates a noteworthy ability to adapt to previously unseen scenarios. This ability is also superior to that of the simpler MLR strategy, especially for the correlation coefficient, highlighting the advantages of employing a more complex nonlinear model in handling such tasks.

We conclude this section by quantifying the contribution of Transfer Learning (TL) to the results presented in Fig. 7. This contribution is detailed in Table 8, where the skill score is reported both for the configuration with Transfer Learning (under the column header 'AI with TL') and without Transfer Learning (under the column header 'AI without TL'). From the analysis of the values in the table, the added value of TL is clearly evident.

Regionalization of climatological variables

In the previous sections, we have focused on assessing our network's capability to predict accumulated precipitation over various time intervals, referenced for each individual test station. Here, we shift our attention to climatological variables, where precipitation is accumulated yearly, aggregating data from all stations within regions L1 and L2 separately. This results in a form of areal precipitation accumulation for each of the two regions, L1 and L2, year by year, from 2005 to 2022 (2020 for the VHR reanalysis). Results have been obtained from our AI network, as well as through the MLR strategy, providing the 24-h accumulated precipitation, the skills of which have been already discussed in previous sections. Essentially, this represents another test of the generalization skills of our predictive strategy, this time conducted on spatially aggregated variables, with their skills observed over a 18-year time span. Note that the test conducted in the L2 region poses a significantly greater challenge compared to L1 region. This is because only a fraction, specifically 10$\%$, of the stations in

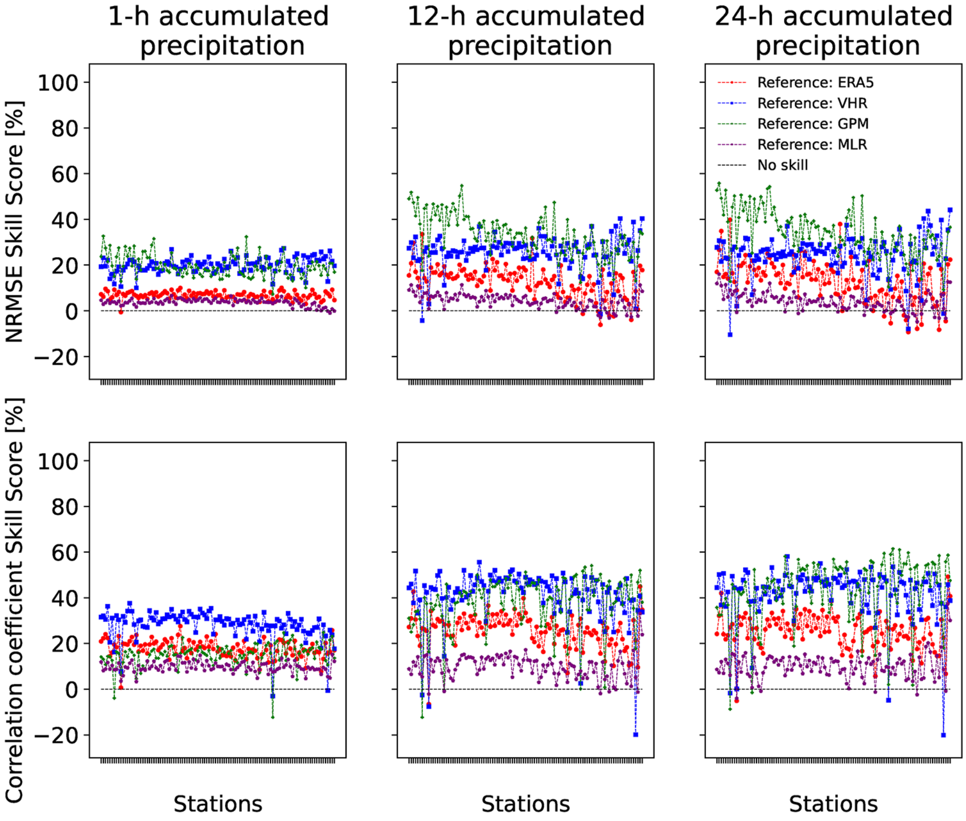

Fig. 7. For the regionalization task in L2, skill scores are computed station by station, considering NRMSE (top row of the panel) and correlation coefficient (bottom row of the panel). ERA5, VHR, GPM , and MLR are used as reference predictions, indicated by different colors. None of the considered rain-gauge stations have been used for training the network. The analysis refers to years from 2005 to 2022 (2020 for VHR). The stations are arranged from left to right, starting with those at lower elevations and progressing to those at higher elevations.

		AI with TL		AI without TL	
		SS NRMSE [%]	SS Corr [%]	SS NRMSE [%]	SS Corr [%]
1-h accum.	GPM	20	15	18	14
	ERA5	7	18	5	16
	VHR	20	28	18	27
	MLR	4	10	2	8
12-h accum.	GPM	33	38	31	34
	ERA5	13	26	10	21
	VHR	26	41	23	37
	MLR	4	10	1	4
24-h accum.	GPM	34	44	31	41
	ERA5	13	25	9	22
	VHR	24	42	21	39
	MLR	4	10	-0.4	6

Table 8. For the regionalization task in L2, and across the three precipitation accumulation intervals, we have computed the averaged skill scores for the Normalized Root Mean Square Error (SS NRMSE) and the correlation coefficient (SS Corr) associated to our AI-based predictions. Skill scores have been obtained taking GPM, ERA5, VHR, and MLR as reference predictions (different rows in the table). Averages have been obtained from the skill scores we have presented in Fig. 7 (values under the column header 'AI with TL). The second block of the table reports the skill scores obtained without transfer learning (values under the column header 'AI without TL).

the L2 region were used for training the network via transfer learning. These stations are both geographically distant and exhibit distinct climatological characteristics compared to those used in the training process within the L1 region. Also note that because the precipitation accumulation has been considered both in time (yearly accumulation) and in space (over the whole regions), the resulting error metric safeguards us against the 'double-penalty effects'[33], potentially penalizing fine-scale reanalysis products.

In Fig. 8, the normalized difference between the accumulated annual precipitations obtained from all reanalysis products [ERA5 (red boxes), VHR (blue boxes), GPM (green boxes), our AI-enhanced strategy (black boxes), MLR (purple boxes)] and the corresponding observed values are presented separately for both L1 (upper panel) and L2 (bottom panel) regions across all years between 2005 and 2022 (up to 2020 for VHR). The normalization was performed by dividing the differences between reanalysis products and observations by the corresponding observed accumulated precipitations.

For the test conducted in the L1 region, our AI-based strategy provides significantly lower mean errors compared to the other reanalyses. This result appears robust over the extended time span considered from 2005 to 2022 (2020 for VHR). Both strategies based on AI surpass the performance of all the benchmarks considered, even in region L2. This finding is interesting given that the AI networks only see a small fraction of stations in the L2 region during the transfer learning phase. Even more interesting is the fact that the AI-based strategies are skillful despite the fact that the variable we considered is large-scale (as previously mentioned, precipitation has been aggregated both in space and over a long time interval), and we have already discussed how traditional reanalyses are typically accurate in predicting such large-scale variables. The resulting Normalized BIAS (NBIAS) and Mean Absolute Percentage Errors (MAPE), both defined in the Methods section, averaged over the 2005-2020 time span, are reported in Table 9 for all considered models and reanalysis and for both regions L1 and L2. These values confirm the considerations made above, once again highlighting the strengths of AI-based strategies compared to traditional reanalyses.

There are future possibilities to further improve the performance of our strategy, as well as the one based on MLR, in relation to the reconstruction of the statistics of climatological variables. Indeed, it is worth remembering that, even in the present case, the AI networks have been trained using non-climatological variables. As a future improvement, the set of training variables could be expanded by explicitly constructing climatological variables to complement those used in the present strategy.

Discussion

Can a computationally efficient AI-based strategy effectively reproduce the variability of local-scale meteorological variables using large-scale information from state-of-the-art spatially-coarse reanalyses? Through an extensive and rigorous analysis spanning from 2005 to 2022, we have provided strong quantitative evidence supporting this conclusion. While our focus has primarily been on the precipitation field, our affirmative answer is expected to extend to a much broader spectrum of meteorological variables.

Our study introduces a strategy termed AI-enhanced, demonstrating its significant superiority both over state-of-the-art dynamic downscaling, which is computationally very demanding, and over the simpler AI-based approaches LR and MLR. This conclusion holds true not only for pointwise precipitation estimates but also for aggregations in space and time relevant for climatological analyses.

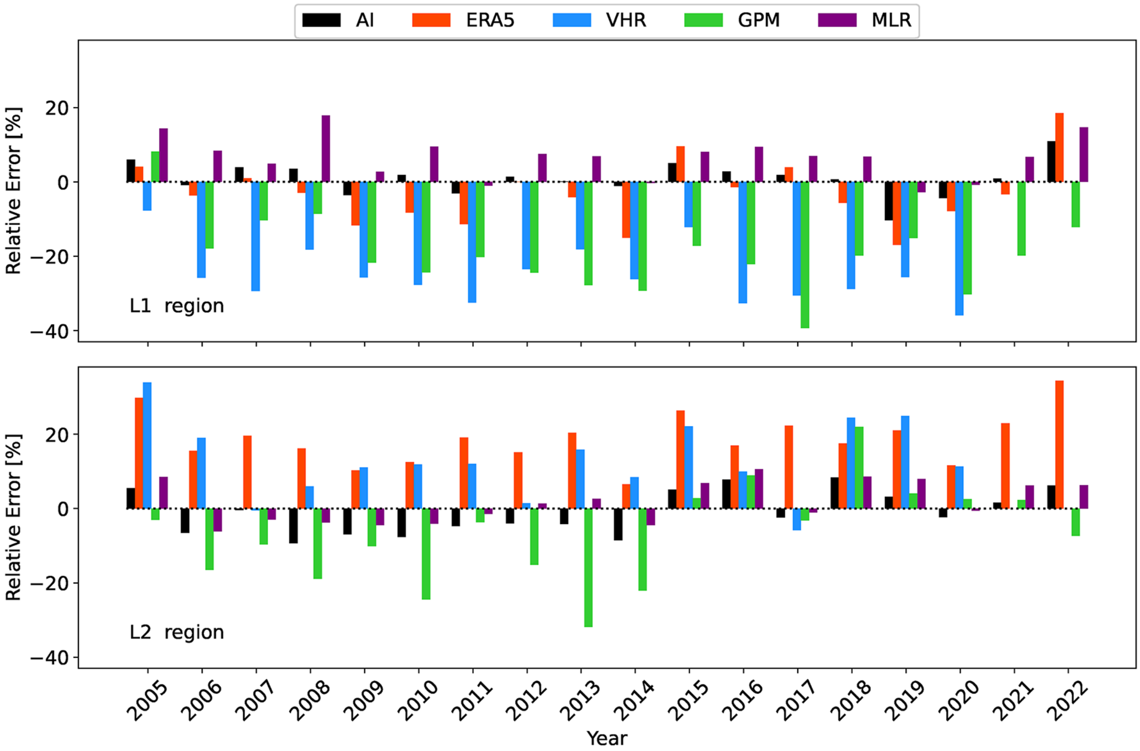

Fig. 8. For ERA5 (red boxes), VHR (blue boxes), GPM (green boxes), MLR (purple boxes) and, our AI-enhanced strategy (black boxes), the normalized error index $(Cum_{rean} - Cum_{obs})/Cum_{obs}$ is reported year by year, from 2005 to 2022 (2020 for the VHR reanalysis), for both regions L1 (upper panel) and L2 (bottom panel). Here, Cum_{rean} denotes the yearly areally-accumulated precipitation from reanalysis ('rean' in short) products, while Cum_{obs} refers to the same quantity measured ('obs' in short) by rain-gauge stations.

	L1		L2	
	NBIAS [%]	MAPE [%]	NBIAS [%]	MAPE [%]
AI	**0.25**	**3.2**	− 1.7	5.4
VHR	− 25.3	25.3	13.7	12.8
ERA5	− 4	6.7	17.5	17.5
GPM	− 20	21	− 7.4	12.4
MLR	6.2	6.7	**1.1**	**4.7**

Table 9. The Normalized BIAS (NBIAS) and Mean Absolute Percentage Errors (MAPE) averaged over the 2005–2020 time span. All considered models and reanalysis are reported for both regions L1 and L2.

These results suggest the potential for rethinking the future role of high-resolution deterministic reanalysis models and high-resolution climate models. While increasing spatial and temporal resolution remains crucial, the findings of this study highlight the growing relevance of AI techniques as complementary tools for reconstructing local-scale variability of surface fields, warranting further exploration alongside traditional approaches.

Our study suggests that although still in its early stages, the integration of AI with deterministic models has already shown remarkable potential, anticipated to reach beyond applications in precipitation. Recent findings related to predicting lightning flashes in the medium-term forecast horizon[24] further affirm our expectations.

Thus, the synergy between AI and deterministic models could redefine how we approach local-scale phenomena and should be further explored in various facets of climate science. The successful implementation of AI-enhanced strategies, as demonstrated in our study, relies heavily on the availability of high-quality observational data. These data not only serve as input for AI models but also play a pivotal role in the evaluation, validation, and improvement of AI-based methodologies. This aspect of our study underscores the growing significance of observed weather data and the policies governing their sharing within the realm of climate science. The collaborative sharing of observational data across international boundaries is essential, as climate phenomena transcend geopolitical borders.

The discussion around the adoption of AI-enhanced strategies, the potential reconfiguration of high-resolution modeling goals, and the integration of AI in climate science is poised to shape the future of research and applications in this field. Further studies and a robust interdisciplinary dialogue are essential to fully harness the potential of these emerging tools.

Methods

The deep learning architecture

Figure 9 shows a schematic representation of the neural network used in this work. The architecture consists of four fully connected dense layers stacked and interspersed with dropout layers. We refer to the fully connected layers as $Dense_i$, with i denoting the *ith* dense layer, as illustrated in Fig. 9, progressing from top to bottom. Specifically, the input is followed by dropout with a rate of 0.15, applied to 96 features. Dense_1 comprises 288 nodes, utilizes the *relu* activation function, and has a dropout rate of 0.5. Dense_2 consists of 372 nodes, employs the *relu* activation function, and has a dropout rate of 0.65. Dense_3 consists of 372 nodes, utilizes the *relu* activation function, and has a dropout rate of 0.65. Dense_4 consists of 288 nodes, employing the *relu* activation function.

The output of Dense_4 is added to the output of Dense_1, creating a residual link, followed by the PReLU activation function. Finally, two additional dense layers are stacked, followed by the output node. Dense_5 comprises 200 nodes, utilizing the *swish* activation function, and Dense_6 consists of 200 nodes, employing the *LeakyReLU* activation function. The architecture remains consistent for both the classifier and the regressor, as well as the features used as input, reported in Table 2. The only difference being the activation function used at the output node.

For the classification task, the output exploits the *sigmoid* activation function in order to get a value between 0 and 1 which represents the probability associated with the precipitation occurrence. For the regression task, at the output node, we chose to use the *relu* activation function in order to output a value greater than 0 representing the accumulated precipitation.

The number of neurons for each dense layer, dropout rates, and regularizations were carefully chosen to mitigate overfitting on the validation dataset. Due to resource limitations, fine-tuning the hyperparameters individually was not feasible. Instead, the hyperparameters and architecture settings were manually selected, with the goal of identifying configurations that consistently produced strong generalizations across both the training and validation datasets. More details about the neural network can be seen in the code available at link https://doi.org/10.5281/zenodo.13766864.

The network is trained using the *Adam* optimizer with a learning rate of 0.0001. The loss function used to train the classifier is the *binary cross entropy*, while for the regressor we use the *mean square error*. Additionally, we trained the networks using early stopping with a patience of 15 epochs and learning rate reduction with a patience of 10 epochs, employing a reducing factor of 0.1. The validation loss threshold for learning rate reduction is set at 0.0001.

It is worth noting that for the training of the regressor, the network parameters were initialized using those obtained from the classifier's training process. This approach establishes a natural connection between the classifier and regressor, thereby leading to improved performance of the latter with respect to starting from a random initialization.

Training and inference speeds

On a cluster equipped with 40 Intel(R) Xeon(R) Gold 6230 CPUs at 2.10 GHz, training one fold of the AI framework on the L1 region takes approximately 2 hours for the 1-hour accumulation period, 25 minutes for the 12-hour period, and 15 minutes for the 24-hour period, respectively. For inferring one fold for the 1-hour, 12-hour, and 24-hour accumulation periods on the same cluster, each of the three deep networks requires about 11 minutes, 95 seconds, and 50 seconds, respectively.

Prevalence-based random model

In the context of a classification problem, we define *Prevalence*, denoted as $\wp$, as the proportion of '1' instances (N_1) relative to the total number of events ($N_1 + N_0$). This is mathematically expressed as $\wp = N_1/(N_1 + N_0)$ and can be also interpreted as the probability of observing the instance '1', given a randomly selected sample out of a total of N. A baseline random model can be built assuming that the labels '1' and '0' are predicted randomly with probability q and $(1 - q)$, respectively. According to this model, given N predictions, the

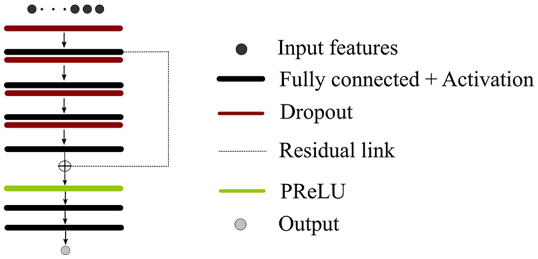

Fig. 9. A schematic representation of the neural network used in this work.

expected number of positive instances (class ‘1’) is qN of which $\wp qN$ are expected to be True Positive (TP) and $(1-\wp)qN$ are expected to be False Positive (FP). The expected number of negative instances (class ‘0’) is $(1-q)N$ of which $(1-\wp)(1-q)N$ are expected to be False Negative (FN). The resulting Precision and Recall thus read: $P = TP/(TP+FP) = \wp$ and $R = TP/(TP+FN) = q$. The random model also provides a baseline prediction for the Area Under the Curve (*AUC*): $AUC = \wp$.

These predictions serve as valuable references for evaluating the performance improvements achieved by more complex prediction strategies.

Logistic and multilinear ridge regression

To implement logistic and multilinear ridge regression (referred to as LR and MLR, respectively), we use the Python library Sklearn[34]. Logistic regression is employed for the classification task, with the training process following the same approach as our AI strategy, as well as utilizing the same input features listed in Table 2. The output of the LR model is a value between 0 and 1, representing the probability of precipitation occurrence. The MLR algorithm, on the other hand, is trained on the TP from the classifier based on the Logistic Regression (LR), and its output is a real number greater than 0, representing the accumulated precipitation over the reference time period. The hyperparameters of the LR and MLR models have mostly been left at their default values from the Sklearn library, except for the maximum number of iterations in LR, which has been set to 1000. For the MLR model, we set the α value to 0.1, which controls the strength of the L2 regularization term.

Assessment of forecast skills

Classifier

In the evaluation of model performance, Precision (also known as success ratio, SR, and related to the false alarm ratio (FAR) by the relation Precision=1-FAR) and Recall (also known as sensitivity or probability of detection, POD) scores are essential, particularly when dealing with unbalanced datasets[35,36]. Precision (P) and Recall (R) are defined as $P = TP/(TP+FP)$ and $R = TP/(TP+FN)$, where TP stands for true positives, FP for false positives, and FN for false negatives. An ideal prediction yields P = R = 1. To determine TP, FP, and FN, a threshold between 0 and 1 is set based on the classifier probability output, distinguishing event ‘1’ (precipitation occurrence) from event ‘0’ (no precipitation). Varying this threshold generates a Precision-Recall (P-R) curve. The AUC (area under the P-R curve) score is an important metric for binary classification models and represents the integral of the P-R curve[37]. A perfect classifier scores an AUC of 1, while a random classifier approximates the prevalence of positive instances in the dataset. An AUC value between 0 and 1 indicates the model’s ability to balance Precision and Recall. Two additional scores have been employed to assess the model’s classification performance: the F1 score, which is the harmonic mean of Precision and Recall and is defined as $F1 = 2P\,R/(P+R)$, and Accuracy, which is defined as $(TP+TN)/(TP+TN+FP+FN)$. These two metrics provide concise and balanced measures of the model’s classification skills[38].

Let us now delve into the evaluation of the reliability of our AI-enhanced strategy. Reliability, in essence, refers to the agreement between the predictions made by a model and the observed frequency of a particular phenomenon[39]. A perfectly calibrated forecast would predict that a set of cases has, for example, a 70% probability of being a precipitation event, and within that set, the actual frequency of precipitation events would indeed be 70%.

To evaluate model reliability, we conducted a series of assessments for each test set within the 5-fold cross-validation process. Specifically, we calculated the probabilities of precipitation occurrences and then divided these probabilities into ten subsets, each representing a distinct range between 0 and 1 in terms of probability. For each of these subsets, we computed the relative frequency of instances corresponding to precipitation occurrences (commonly referred to as the fraction of positive events) and plotted this against the calculated relative frequency of precipitation events, often referred to as the mean predicted probability. The graphical representation of this process is depicted in the resulting reliability diagram.

Regressor

To complete the process of evaluating the skills of our strategy, we assessed the accumulated precipitation values against observed data for the two reanalysis datasets considered and for GPM. The statistical indices considered for this assessment are[40] the Normalized Root Mean Square Error (NRMSE), the Correlation Coefficient and the Normalized BIAS. Namely, denoting by O_i the ith observation (among a total of n) and by F_i the ith prediction (from reanalysis and GPM) we have:

$$\mathrm{NRMSE} = \sqrt{\frac{\sum_{i=1}^{n}(O_i - F_i)^2}{\sum_{i=1}^{n} O_i^2}} \tag{1}$$

$$\text{Correlation Coefficient (Pearson)} = \frac{\sum_{i=1}^{n}(O_i-\bar{O})(F_i-\bar{F})}{\sqrt{\sum_{i=1}^{n}(O_i-\bar{O})^2}\sqrt{\sum_{i=1}^{n}(F_i-\bar{F})^2}} \qquad \bar{O} = \frac{1}{n}\sum_{i=1}^{n} O_i \qquad \bar{F} = \frac{1}{n}\sum_{i=1}^{n} F_i \tag{2}$$

All the statistical indices above have been calculated for AI, ERA5, VHR, GPM, and MLR predictions conditional on the observed wet conditions.

For climatological observables, we have used the Mean Absolute Percentage Error (MAPE) defined as

$$\text{MAPE} = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{F_i - O_i}{O_i}\right| \tag{3}$$

and the Normalized BIAS index

$$\text{NBIAS}\frac{1}{n}\sum_{i=1}^{n}\frac{F_i - O_i}{O_i} \tag{4}$$

To calculate the MAPE and NBIAS indices, no conditioning on observed wet conditions has been used.

Skill score index

To demonstrate the predictive capabilities of our AI strategy in comparison to a reference, we employ the widely recognized skill-score index[40,41]. This index is defined as $SS = (a - a_{ref})/(a_{opt} - a_{ref})$, where a represents an error index used for evaluating prediction quality, a_{ref} denotes the error index associated with a reference prediction, and a_{opt} represents the index value corresponding to optimality. The skill score index is used here to assess both classifier and regressor skills.

Data availability

For training and testing our AI-enhanced network, we downloaded a subset of the ERA5 dataset, for single-level and pressure level, from 2005 to 2022, totaling 8.6 GB (Grib format, almost equally distributed between regions L1 and L2), from the Climate Data Store (CDS) at the ECMWF Copenicus Climate Change Service (C3S), accessible upon registration at URL: https://cds.climate.copernicus.eu/#!/home. For region L1, data are downloaded from longitudes between 7.0 E and 10.5 E and latitudes between 43.25 N and 45.25 N, while, for region L2, data are downloaded from longitudes between 8.25 E and 11.75 E and latitudes between 44.50 N and 46.75 N. Additionally, for training and testing our AI network, we have obtained precipitation estimates from the GPM constellation, specifically the IMERG-E and IMERG-L gridded products, which provide estimates of precipitation accumulation on a 30-minute timespan at a spatial resolution of approximately 10 km. These data cover the period from 2005 to 2022, totaling approximately 45 GB (almost equally distributed between regions L1 and L2). The two products, IMERG-E and IMERG-L, have been downloaded in two domains, each encompassing regions L1 and L2, defined by the same latitudes and longitudes used for downloading ERA5 data. The GPM data can be accessed from the Goddard Earth Sciences (GES) Data and Information Services Center (DISC) of the NASA Earthdata. Detailed instructions for accessing data are available at https://disc.gsfc.nasa.gov/information/documents?title=Data%20Access For testing purposes, we downloaded the VHR-REA_IT dataset, which is available at the following link: https://dds.cmcc.it/#/dataset/era5-downscaled-over-italy. This dataset allows for the Python API to download the desired data for a specific time period and area. The downloaded data cover the period from 2005 to 2020, on the same domains L1 and L2 as for ERA5 and GPM, totaling approximately 15 GB. The rain-gauge data used for training and testing our AI network have been downloaded from https://ambientepub.regione.liguria.it/SiraQualMeteo/script/PubAccessoDatiMeteo.asp (for the Liguria Region, years from 2005 to 2022, data volume of about 600 Mb) and from https://www.dati.lombardia.it/Ambiente/Mappa-Stazioni-Meteorologiche/8ux9-ue3c for Lombardia region (years from 2005 to 2022, data volume of about 610 Mb). In both cases data are freely available for research purposes. We have prepared the dataset containing ERA5 and GPM variables, and the rain-gauge data for the L1 and L2 regions relative to the 5 folds considered in the paper. The datasets generated and/or analysed during the current study are available in the zenodo repository. https://doi.org/10.5281/zenodo.13766864.

Received: 15 May 2024; Accepted: 23 October 2024
Published online: 30 October 2024

References

1. Cucchi, M. et al. WFDE5: Bias-adjusted ERA5 reanalysis data for impact studies. *Earth Syst. Sci. Data* **12**, 2097–2120. https://doi.org/10.5194/essd-12-2097-2020 (2020).
2. Chan, W. C. H. et al. Added value of seasonal hindcasts for UK hydrological drought outlook. *Nat. Hazards Earth Syst. Sci. Discuss.* **1–21**, 2023. https://doi.org/10.5194/nhess-2023-74 (2023).
3. Ferrari, F., Besio, G., Cassola, F. & Mazzino, A. Optimized wind and wave energy resource assessment and offshore exploitability in the mediterranean sea. *Energy* **190**, 116447. https://doi.org/10.1016/j.energy.2019.116447 (2020).
4. Ozturk, U., Saito, H., Matsushi, Y., Crisologo, I. & Schwanghart, W. Can global rainfall estimates (satellite and reanalysis) aid landslide hindcasting?. *Landslides* **18**, 3119–3133 (2021).
5. Pielke, R. A., Adegoke, J., Hossain, F. & Niyogi, D. Environmental and social risks to biodiversity and ecosystem health-a bottom-up, resource-focused assessment framework. *Earth* **2**, 440–456. https://doi.org/10.3390/earth2030026 (2021).
6. Cardoso, R. M. & Soares, P. M. M. Is there added value in the EURO-CORDEX hindcast temperature simulations? Assessing the added value using climate distributions in Europe. *Int. J. Climatol.* **42**, 4024–4039 (2022).
7. Hersbach, H. et al. The ERA5 global reanalysis. *Q. J. R. Meteorol. Soc.* **146**, 1999–2049 (2020).
8. Dee, D. P. et al. The ERA-interim reanalysis: Configuration and performance of the data assimilation system. *Q. J. R. Meteorol. Soc.* **137**, 553–597 (2011).
9. Saha, S. et al. The NCEP climate forecast system reanalysis. *Bull. Am. Meteor. Soc.* **91**, 1015–1058 (2010).
10. Cavaiola, M., Tuju, P. E., Ferrari, F., Casciaro, G. & Mazzino, A. Ensemble machine learning greatly improves ERA5 skills for wind energy applications. *Energy AI* **13**, 100269 (2023).
11. Bandhauer, M. et al. Evaluation of daily precipitation analyses in E-OBS (v19.0e) and ERA5 by comparison to regional high-resolution datasets in European regions. *Int. J. Climatol.* **42**, 727–747. https://doi.org/10.1002/joc.7269 (2022).

12. Jiang, Y. et al. A downscaling approach for constructing high-resolution precipitation dataset over the Tibetan Plateau from ERA5 reanalysis. *Atmos. Res.* **256**, 105574 (2021).
13. Zhang, W., Villarini, G., Scoccimarro, E. & Napolitano, F. Examining the precipitation associated with medicanes in the high-resolution ERA-5 reanalysis data. *Int. J. Climatol.* **41**, E126–E132. https://doi.org/10.1002/joc.6669 (2021).
14. Ferrari, F. et al. Impact of model resolution and initial/boundary conditions in forecasting flood-causing precipitations. *Atmosphere* **11**, 592 (2020).
15. Copernicus Climate Change Service. Complete UERRA regional reanalysis for Europe from 1961 to 2019, (2019). Accessed on 12-10-2023. https://doi.org/10.24381/cds.dd7c6d66
16. Schimanke, S. et al. CERRA sub-daily regional reanalysis data for Europe on single levels from 1984 to present, (2021). Accessed on 12 Oct 2023. https://doi.org/10.24381/cds.622a565a
17. Bollmeyer, C. et al. Towards a high-resolution regional reanalysis for the European CORDEX domain. *Q. J. R. Meteorol. Soc.* **141**, 1–15 (2015).
18. Whelan, E., Gleeson, E. & Hanley, J. An evaluation of MÉRA, a high-resolution mesoscale regional reanalysis. *J. Appl. Meteorol. Climatol.* **57**, 2179–2196 (2018).
19. Bonanno, R., Lacavalla, M. & Sperati, S. A new high-resolution meteorological reanalysis Italian dataset: MERIDA. *Q. J. R. Meteorol. Soc.* **145**, 1756–1779 (2019).
20. Capecchi, V., Pasi, F., Gozzini, B. & Brandini, C. A convection-permitting and limited-area model hindcast driven by ERA5 data: Precipitation performances in Italy. *Clim. Dyn.* **61**, 1411–1437 (2023).
21. Raffa, M. et al. VHR-REA_IT dataset: Very high resolution dynamical downscaling of ERA5 reanalysis over Italy by COSMO-CLM. *Data* **6**, 88 (2021).
22. Bi, K. et al. Accurate medium-range global weather forecasting with 3d neural networks. *Nature* **619**, 533–538 (2023).
23. Zhang, Y. et al. Skilful nowcasting of extreme precipitation with NowcastNet. *Nature* **619**, 526–532 (2023).
24. Cavaiola, M., Cassola, F., Sacchetti, D., Ferrari, F. & Mazzino, A. Hybrid ai-enhanced lightning flash prediction in the medium-range forecast horizon. *Nat. Commun.* **15**, 1188 (2024).
25. Sun, H. et al. *J. Hydrometeorol.* **23**, 1663–1679. https://doi.org/10.1175/JHM-D-22-0015.1 (2022).
26. Wang, F., Tian, D. & Carroll, M. Customized deep learning for precipitation bias correction and downscaling. *Geosci. Model Dev.* **16**, 535–556. https://doi.org/10.5194/gmd-16-535-2023 (2023).
27. Schneider, T. et al. Harnessing ai and computing to advance climate modelling and prediction. *Nat. Clim. Chang.* **13**, 887–889 (2023).
28. Huffman, G. J. et al. *Integrated Multi-satellite Retrievals for the Global Precipitation Measurement (GPM) Mission (IMERG), 343–353* (Springer International Publishing, 2020).
29. Hosmer, D. W. Jr., Lemeshow, S. & Sturdivant, R. X. *Applied logistic regression* (Wiley, 2013).
30. McDonald, G. C. Ridge regression. *Wiley Interdiscip. Rev. Comput. Stat.* **1**, 93–100 (2009).
31. Muñoz Sabater, J. et al. ERA5-Land: A state-of-the-art global reanalysis dataset for land applications. *Earth Syst. Sci. Data* **13**, 4349–4383. https://doi.org/10.5194/essd-13-4349-2021 (2021).
32. Pan, S. J. & Yang, Q. A survey on transfer learning. *IEEE Trans. Knowl. Data Eng.* **22**, 1345–1359. https://doi.org/10.1109/TKDE.2009.191 (2010).
33. Cassola, F., Ferrari, F. & Mazzino, A. Numerical simulations of Mediterranean heavy precipitation events with the WRF model: A verification exercise using different approaches. *Atmos. Res.* **164–165**, 210–225 (2015).
34. Pedregosa, F. et al. Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).
35. Hoens, T. R. & Chawla, N. V. Imbalanced datasets: from sampling to classifiers. Imbalanced learning: Foundations, algorithms, and applications 43–59 (2013).
36. Sofaer, H. R., Hoeting, J. A. & Jarnevich, C. S. The area under the precision-recall curve as a performance metric for rare binary events. *Methods Ecol. Evol.* **10**, 565–577 (2019).
37. Branco, P., Torgo, L. & Ribeiro, R. P. A survey of predictive modeling on imbalanced domains. *ACM Comput. Surv. (CSUR)* **49**, 1–50 (2016).
38. Goutte, C. & Gaussier, E. A Probabilistic Interpretation of Precision, Recall and F-Score, with Implication for Evaluation. In *Advances in Information Retrieval, 345–359 (Springer* (eds Losada, David E. & Fernández-Luna, Juan M.) (Berlin Heidelberg, 2005).
39. Silva Filho, T. et al. Classifier calibration: a survey on how to assess and improve predicted class probabilities. Mach. Learn. 1–50 (2023).
40. Wilks, D. S. *Statistical Methods in the Atmospheric Sciences* Vol. 100 (Academic Press, 2011).
41. Casciaro, G., Cavaiola, M. & Mazzino, A. Calibrating the CAMS European multi-model air quality forecasts for regional air pollution monitoring. *Atmos. Environ.* **287**, 119259. https://doi.org/10.1016/j.atmosenv.2022.119259 (2022).

Acknowledgements

Fondazione Compagnia di San Paolo and *Fondazione CDP* are warmly acknowledged for the financial support (project AIxtreme, call “Intelligenza Artificiale” 2022). M.C. thanks the project “RAISE - Robotics and AI for Socio-economic Empowerment” and has been supported by European Union - NextGenerationEU. Useful discussions with Francesco Ferrari and Daniele Carnevale are also acknowledged.

Author contributions

A.M. designed the research; A.M. and M.C. proposed the initial concept, which M.C. refined into a working strategy; M.C. developed the final AI-enhanced framework; P.E.T. handled all reanalysis datasets and prepared them to be used by the AI strategy; M.C., P.E.T., and A.M. analyzed the data; M.C. carried out model testing and validation; A.M. and M.C. wrote the paper.

Declarations

Code availability

The AI framework (to be used for both training and inference) is accessible at the following link https://doi.org/10.5281/zenodo.13766864. The codes are based on TensorFlow, a Python-based library for deep learning. Other Python libraries, such as NumPy, SKlearn, and Matplotlib have also been used. The statistical indices have been calculated using Sklearn[34]. All the details, including network architecture, modules, optimizations, and hyperparameters, are also available in the repository.

Additional information

Correspondence and requests for materials should be addressed to M.C. or A.M.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.