

# From Hours to Daily Extremes: Evaluating ERA5 and ERA5-Land Biases Across Morocco’s Diverse Climates

Badreddine Alaoui<sup>1\*</sup>, Driss Bari<sup>1</sup> and Abdou Khouakhi<sup>2</sup>

<sup>1\*</sup>National Center of Meteorological Research, General Directorate of Meteorology, Boulevard Mohamed Taieb Naciri–Hay Hassani, Casablanca, 8106, Casablanca, Morocco.

<sup>2</sup>Faculty of Engineering and Applied Sciences, Cranfield University, College Rd, Wharley End, Bedford, MK43 0AL, UK.

\*Corresponding author(s). E-mail(s): [alaoui.badreddine.abe@gmail.com](mailto:alaoui.badreddine.abe@gmail.com);

## Abstract

Reanalysis products increasingly serve as ground-truth references and training/verification data for statistical and machine-learning applications in regions with sparse in-situ networks. This study evaluates ERA5 and ERA5-Land against 36 quality controlled synoptic stations over Morocco, across hourly, seasonal, Köppen-climate, and daily-extreme timescales, from 2016 to 2024. Both reanalysis reproduce the mean diurnal and seasonal cycle of near-surface variables with high fidelity (Pearson  $r$  frequently  $> 0.9$  for temperature), and the two reanalysis products are broadly interchangeable for climatological applications, with small, systematic differences per variable (ERA5 ahead for temperature, humidity and wind bias; ERA5-Land ahead for wind-speed RMSE in complex terrain). Categorical verification of daily extremes reveals a threshold-dependent, rather than uniformly poor, skill: moderate warm/cold/wind thresholds are detected well by both products, but skill collapses specifically at the most extreme, rarest thresholds (POD falling from 0.85 to 0.4 for the hottest T2M\_MAX class; near-zero for the strongest WS\_MAX class), a pattern consistent with the known degeneracy of traditional contingency-table scores for rare events rather than with a specific flaw of either reanalysis. We additionally document a structural ERA5-Land limitation: its land-only grid leaves no valid nearest-neighbour match at some coastal/island-adjacent stations (e.g. Dakhla), a gap that is easily mistaken for missing observations if not checked explicitly. The results argue for anomaly- or percentile-based extreme definitions, and for explicit reporting of land-mask

coverage, when ERA5/ERA5-Land are used for impact studies or as an AI-model verification baseline over Morocco’s complex, station-sparse terrain.

**Keywords:** ERA5 Reanalysis, ERA5-Land Reanalysis, Validation, Morocco, Extreme Weather, Climates

## 1 Introduction

The fifth-generation ECMWF reanalysis, ERA5 [1], and its land-surface companion, ERA5-Land [2], are now the default reference datasets for a wide range of downstream products: bias-correction of numerical weather prediction, hydrological and agrometeorological modelling, climate-service indicators, weather remote sensing validation, and the training and out-of-sample evaluation of data-driven (AI) weather models. This last use case raises the stakes of reanalysis quality assurance considerably, since biases and unresolved local processes in the reanalysis are silently inherited by any model trained on it. On global scale several research studies have been achieved in reanalysis validation or intercomparison. Mainly driven by climate monitoring such studies suggest generally three streams of validation: (1) a climate scale evaluation of the reanalysis quality and accuracy on daily basis, (2) a validation of reanalysis against satellite data and finally (3) a validation against extreme weather events. In recent years the ERA5 reanalysis played a key role in the development of weather fully data driven models. As they found the initial unique complete reliable input covering profound archive periods extending to 40 years. The majority of global weather foundation models (GraphCast [3], Aurora [4], Pangu [5], FourCastNet [6] and AIFS [7]) are commonly trained using ERA5 Reanalysis differing only in architectures with a set based on graphical neural networks and another based on transformers. These data-driven models have matched state-of-the-art physics-based systems like the ECMWF’s IFS HRES in deterministic medium-range forecasting across multiple lead times [8, 9]. Independent benchmark evaluations, such as WeatherBench 2, affirms that these models consistently achieve higher average skill across various parameters and metrics and provide competitive performances to classical NWP models [10–14]. Despite strong overall performance, existing assessments focus generally on average forecast skill rather than extreme events. This shortcoming known as the out of sample data driven models issue remain big challenge despite isolated case studies exist such as cyclone tracking [15] and surface temperature extremes [11] a global, systematic assessment is missing. Average forecast skill does not automatically guarantee strong performance in the tails of the distribution [16–18]. Discrepancy between data driven models forecasts and extreme events are mainly due to three factors: (a) Sparsity of extreme events in training data mainly ERA5 Reanalysis [16]; (b) Inadequate symmetric loss functions that fail to weight extremes appropriately [17] and (c) Architectural optimizations, such as multi-task and multi-step training (e.g., [3, 5]). Since ERA5 Reanalysis acts as the main ground truth baseline and training input for advanced data-driven weather

models, it is essential to validate its accuracy and reliability against local observations in specific regions to detect any regional representation errors before they affect model predictions.

Reliable extreme weather forecasts are vital for disaster risk management [19, 20] and key socioeconomic sectors like energy and insurance [21]. Consequently, validating extreme-weather performance is critical prior to operational adoption [16]. This imperative directly underlines the necessity of locally validating ERA5 and ERA5-Land, which serve as the primary inputs, training targets, and verification benchmarks for data-driven models. Because gridded reanalysis inherently smooth out hyper-local physical dynamics and peak values, an AI model that appears highly accurate against global ERA5 baseline data may still fail to capture severe, real-world local impacts. To guarantee reliable deployment in risk-critical decision-making, the ERA5 and ERA5-Land benchmarks themselves must be rigorously evaluated against local ground observations and extreme-quantile thresholds, ensuring that downstream AI predictions reflect true localized hazards rather than smoothed reanalysis artifacts. ERA5 and ERA5-Land reanalysis play a vital role in weather model verification and validation by serving as the primary benchmark. Numerous benchmark datasets spanning mesoscale to climate scales have been built on ERA5 data to act as ground truth for operational and fully data-driven forecasting models. This foundational role is crucial, as these reanalysis supply the core baseline needed to evaluate model performance, particularly at high altitudes and in data-sparse regions where observational coverage is sparse.

## 2 Literature Review

Many studies on global reanalysis validation generally categorize their evaluations into three main streams: assessing daily accuracy across climate scales, validating data against satellite observations, and testing reliability during extreme weather events. From global validation perspective, authors in [22] conducted a global assessment of ERA5 temperature and precipitation accuracy on monthly scale over the past 70 years against Climatic Research Unit Time-Series version 4.04) gridded data using multiple statistical scores and detection metrics. Authors in [22] report overall good consistency of ERA5 reanalysis with gaps in accuracy of ERA5 between the two periods (1950–1979, 1979–2019) and generally a better performance of ERA5 temperature compared to precipitation. Results also show that mid- and low-latitude precipitation in both hemispheres follows a positive–negative pattern, while for temperatures an inverse distribution between the hemispheres, dominated by warmer conditions in the north and colder conditions in the south is noticed.

Correspondingly authors in [23] evaluated ERA5 datasets hourly wind speed 10m against HadISD observations across 245 European stations showing that ERA5 effectively captures regional wind characteristics. Authors also highlighted that ERA5 replicates accurately the annual cycle reproducing monthly wind speed cycle across Europe with slightly higher monthly medians in winter and lower medians in summer compared to observational data. Using a frequency distribution-based score ranging from 0 to 1, most stations record scores between 0.8 and 0.9. This indicates a high

capability in mapping the full wind speed spectrum (from light to strong relative frequencies) across different European locations.

Regarding ERA5 validation against satellite data, authors in [24] evaluated the global horizontal irradiance estimates from ERA5 and COSMO-REA6 reanalysis using ground and satellite-based data. Authors found that ERA5 and COSMO-REA6 reduce the quality gap between reanalysis and satellite data. ERA5 bias is like that of satellite data for inland regions with few clouds. COSMO-REA6 is comparable to ERA5 for North and Central Europe. The absolute error of reanalysis is large due to a deficient cloud prediction. COSMO-REA6 underestimates due to the overestimation of aerosols. In similar context [25] validated ERA5 reanalysis against cloud height (CHB) base from satellite images globally, the authors found that ERA5 underestimated globally CHB with strong bias in low latitudes, large bias on low latitudes depends mainly on separated type of cirrus clouds.

ERA5 Reanalysis helped also in monitoring several global weather and climate extremes, authors in [26] assessed how the improvement in ERA5 reanalysis can affect the accuracy of simulating storm surge events using the Global Tide and Surge Model. The authors find that the improved wind and pressure fields in ERA5 compared to ERA-Interim helps increase accuracy by reducing tropical cyclone wind speed errors by 58%, for the Irma use case authors stress out a matching between ERA5 and tide gauge observations compared to large under-estimation using ERA-Interim. This displays that model resolution markedly enhances global coastal flood hazard assessment in early warning systems.

Over the Mediterranean region, authors in [27] applied ERA5-Land to long-term precipitation variability from 1950 to 2024, deriving high-resolution indices of precipitation frequency, intensity and persistence and testing them for trend with the Mann-Kendall test and Sen’s slope estimator. The dataset resolves regional contrasts, including decreasing heavy precipitation frequencies in the western Mediterranean and increasing extremes in the northeastern basin, highlighting its capability to represent spatially heterogeneous climate signals. Likewise, authors in [28] displays that the high spatial and temporal resolution of ERA5 has proven valuable for atmospheric circulation studies, particularly for cyclone detection and climatology. By exploiting ERA5’s enhanced resolution, a newly developed cyclone detection and tracking method identified approximately 40% more cyclones than previous datasets, especially sub-synoptic and baroclinic systems, while accurately reproducing known cyclogenesis regions and revealing previously underestimated areas such as the Aegean Sea and the Tunisian–Libyan coasts. Finally, authors in [29] provided 29-year Mediterranean wind and wave hindcast driven by ERA5 atmospheric fields showed good agreement with buoy and coastal station observations, with significant wave height correlations ranging from 0.89 to 0.96 and improved representation of coastal wind and wave characteristics compared with the original ERA5 fields.

Over Africa, and Morocco in particular, the observation network is comparatively sparse and heterogeneous in exposure, which makes independent, station-based verification both more difficult and more necessary. Prior evaluations of ERA5 over the African continent have focused primarily on precipitation at seasonal-to-interannual scales [30], while Moroccan-focused studies have generally targeted a single basin or a

handful of rain gauges for hydrological applications [31–33]. Authors in [34] compared MERRA-2 and ERA5 reanalysis datasets for wind energy modeling and assess the technical and economic viability across 39 sites, the study displays that MERRA-2 outperformed ERA5.

Authors in [35] reconstructed 44 years continuous monthly precipitation dataset over the Sebou basin in Morocco, spanning 11 stations (5 m to 1478 m elevation) using Quantile Mapping (QM) bias-corrected ERA5 data. The authors underlined ERA5 severe positive biases at high altitudes. Seasonal Quantile Mapping reduced mean bias from +8.9% down to -0.1% and improved mean Nash-Sutcliffe Efficiency (NSE from -0.32 to +0.68)

Authors in [10, 36] suggests two datasets Weather Bench and Weather Bench 2 both datasets are fundamentally based on ERA5 and ERA5-Land reanalysis. WeatherBench provides a standardized framework for evaluating data-driven weather forecasting methods. While physics-based numerical models excel at many tasks, key performance gaps remain. In contrast, artificial intelligence and deep learning leverage observational datasets directly, offering the potential to surpass traditional forecast accuracy. Standardized benchmarks such as WeatherBench and its second-generation iteration, WeatherBench 2 establishes unified evaluation metrics, open-source datasets, and code, with initial tasks focused on global pressure and temperature predictions across 3- to 5-day lead times. By setting these standards, the project aims to bridge AI and traditional numerical modeling, fostering interdisciplinary collaboration to accelerate progress in global weather prediction.

On the other hand, authors in [37] proposed a benchmark on Snow Fields of ERA5-Land in the Northern Regions of North America. The authors, evaluates the accuracy of ERA5-Land snow cover, snow depth, and Snow Water Equivalent (SWE) across monthly, seasonal, and annual scales, within the ecological regions of Canada and Alaska, regions that are characterized by prolonged seasonal snow cover. By comparing ERA5-Land’s snow data against MODIS satellite imagery and Canadian Meteorological Centre analyses, the authors found that ERA5-Land severely overestimates historical averages and year-to-year variability while distorting multi-year trends. Raw values for both variables should not be used directly. However, ERA5-Land stays more reliable for snow cover and density but still requires local bias adjustments before use.

Together, these studies confirm that ERA5 and ERA5-Land provide a robust and physically consistent representation of Global, Mediterranean and Moroccan atmospheric and hydrological conditions in climatological scale, making them reliable benchmarking and training datasets for climate trend analyses, extreme event characterization and cyclone climatology, however a local fine-scale verification at the station level on an hourly basis is always a necessity to assess regional performance.

To our knowledge, a station-dense, multi-variable evaluation of ERA5 datasets across Morocco’s full diurnal cycle, diverse Köppen climate classes [38], and daily extremes remains absent from the literature. This work addresses this benchmark gap, providing the rigorous baseline validation of the performance of both ERA5 and ERA5-Land in capturing hourly observations by climates space and extreme weather events over Morocco for 36 ground stations from 2016 to 2024.

This work conducts a systematic, multi-variable, station-dense verification spanning the full diurnal cycle, the Köppen climate classes present in Morocco, along with daily temperature/wind/precipitation extremes. This is precisely required before ERA5/ERA5-Land can be trusted as a baseline for out-of-sample AI model verification over Morocco.

This work targets the following objectives:

- Hourly validation of ERA5 and ERA5-Land over different climate zones, several stations during 8 years against synoptic observations.
- Daily verification of the ERA5 and ERA5-Land in detecting extreme weather events of temperature, wind speed and precipitation.

This article is organized as follows: Following the introduction and literature review section that presents several works related to the global, local validation of reanalysis, their usage as inputs and benchmarks for data-driven models. The third section is dedicated to the studied domain, the used dataset, and the effect of nearest neighbor's grid point selection for our use case. The fourth part is devoted to methodology where the verification framework is presented along with the metrics used for continuous and binary validation. Results section covers two verification process: continuous verification using bias, RMSE, relative RMSE and correlation and a probabilistic verification process where occurrences of extreme wind, temperature and precipitation events are conducted using an operational set of thresholds. Finally, the conclusion and discussion where further perspectives and limitations of this work are highlighted.

### 3 Dataset and Study Domain

For this study the datasets used are ERA5 Reanalysis [1] and ERA5 Land Reanalysis [2] provided by the European Center of Medium Range Weather Forecasts.

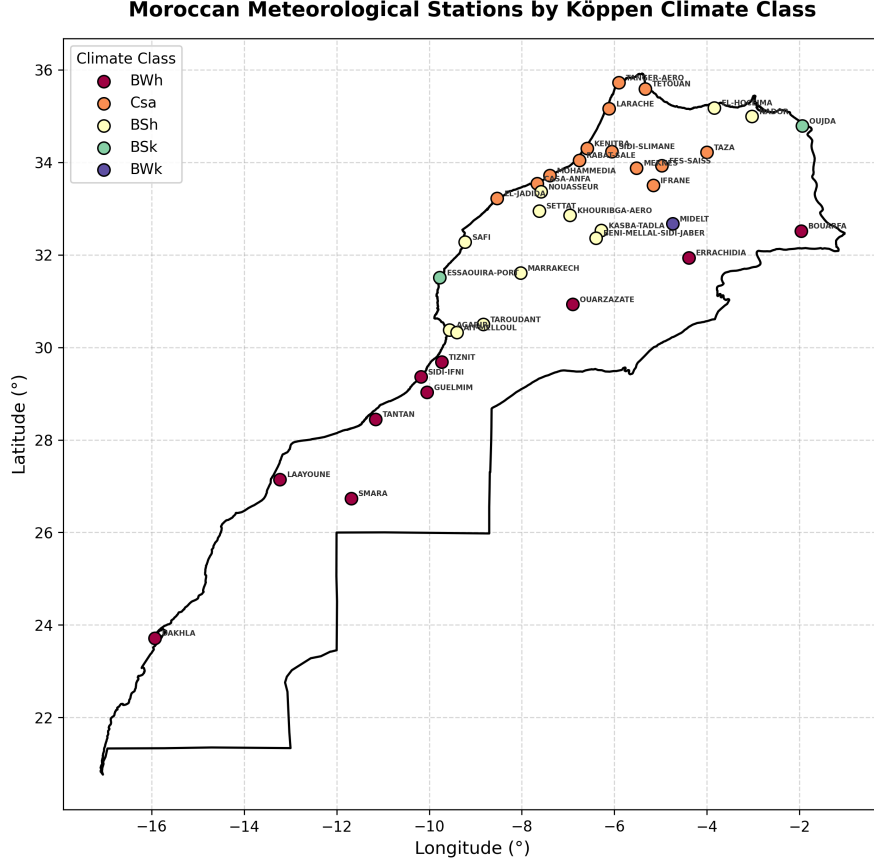


**Fig. 1** Moroccan synoptic observation stations by Köppen climate class spatial distribution

### 3.1 ERA5 Reanalysis

ERA5 is the fifth-generation ECMWF global atmospheric reanalysis delivered for the Copernicus Climate Change Service, operating on the Integrated Forecasting System (IFS Cycle 41r2) to produce a comprehensive multi-decadal climate record. Technically, these reanalysis offer a major upgrade compared to its predecessor by refining the global horizontal grid resolution from approximately 80 km down to 31 km (TCO639) and extending the vertical domain across 137 hybrid sigma-pressure levels from the surface up to 0.01 hPa. Driven by twice-daily 12-hour 4D-Var data assimilation cycles (initialized at 06:00 and 18:00 UTC), ERA5 uniquely delivers hourly outputs for atmospheric, land-surface, and ocean-wave parameters [1]. ERA5 has established itself as the indispensable foundation for a vast span of Earth-system applications, serving as the definitive baseline that bridges traditional Numerical Weather Prediction (NWP), multi-decadal climate monitoring, advanced satellite data assimilation, and the modern revolution of AI-driven weather forecasting models [3, 5].

### 3.2 ERA5-Land Reanalysis

Along with ERA5, ERA5-Land is a state-of-the-art global land reanalysis dataset designed to provide a highly consistent, multi-decadal record of the Earth’s water and energy cycles near the surface. Technically, it is produced by running offline global numerical integrations of the ECMWF land surface model at a high native spatial resolution of 9 km ( $0.1^\circ \times 0.1^\circ$ )—a major refinement compared to the parent ERA5 atmospheric grid (31 km). Driven exclusively by ERA5 atmospheric forcing fields (such as air temperature, humidity, and surface pressure) corrected via lapse-rate adjustments for altitude differences, ERA5-Land outputs hourly variables that capture enhanced land-surface heterogeneity. It has become an indispensable baseline for hydrological studies, flood and drought forecasting, initializing regional NWP and climate models, and analyzing land-atmosphere feedback [2]. Because ERA5-Land is generated over a land-only mask, its nearest grid cell to a coastal or island-adjacent station can be undefined or hydrologically unrepresentative.

### 3.3 Observations

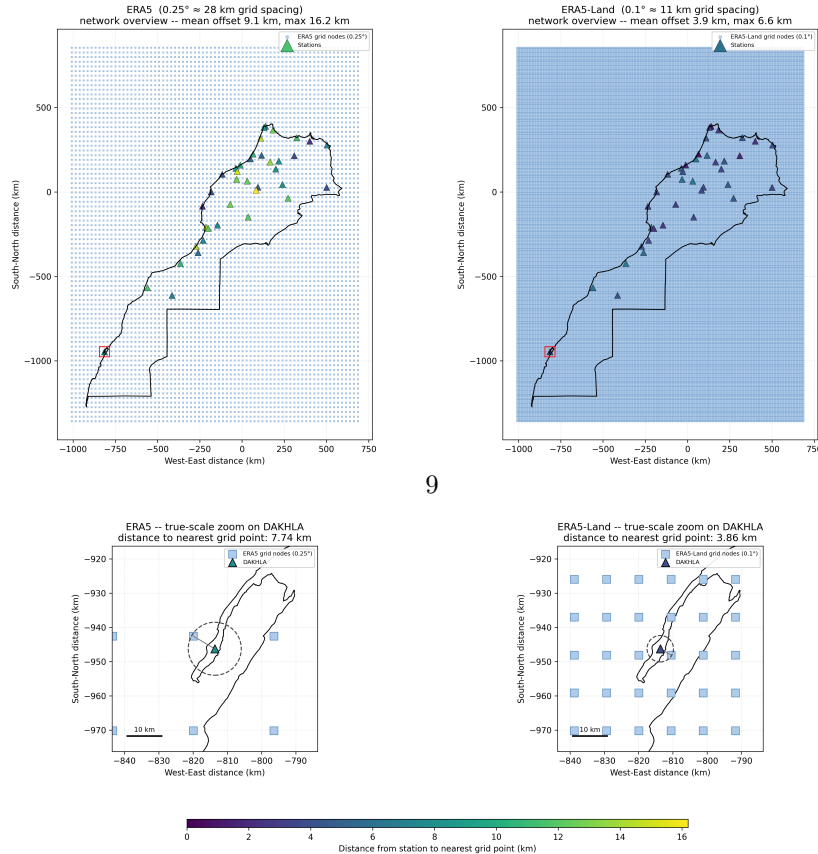
Hourly surface observations from 36 synoptic stations operated by DGM are used as ground truth, covering the full domain of the Moroccan Kingdom. Observations stations span over various climate zones and topographies (See Figure 1 and 1). The climate classification for the validation of reanalysis is driven by the Koppen classification where five classes are present in Morocco semi-arid (BSh) and cold semi-arid (BSk) in the interior; hot desert (BWh) and cold desert (BWk) in the Saharan and high-altitude zones; and hot-summer Mediterranean (Csa) along the coast and northern mountains. Standard quality-control procedures (range, internal-consistency and temporal-continuity checks, following WMO guidance in [39] are applied prior to matching.

### 3.4 Nearest grid points representativeness

Figure 2 displays the spatial density of the minimal distance between each reanalysis grid point and the exact location of the ground station for ERA5 and ERA5-Land. Most of the stations nearest reanalysis grid points are in a less than 10km radius for both ERA5 and ERA5-Land, except for some interior stations (BENI-MELLAL and NOUACEUR) where this distance can climb up to 16km, which goes in line with WMO synoptic spatial representativeness. Figure 2 also highlights a structural ERA5-Land limitation: its land-only grid leaves no valid nearest-neighbor match at some coastal/island-adjacent stations (e.g. Dakhla see Zoom in Figure 2), a gap that is easily mistaken for missing observations if not checked explicitly, this displays a structural limitation of ERA5-Land compared to ERA5 despite the high-resolution advantage. It is to mention that elevation corrections for stations beyond 750m have been carried out for the studied variables.

**Table 1** Station Metadata and Data Availability starting from 01-01-2016

| Wmoid | Station        | Alt    | Records | End_Date   | T2m (%) | TP (%) | WS10m (%) | RH2m (%) |
|-------|----------------|--------|---------|------------|---------|--------|-----------|----------|
| 60033 | LAAYOUNE       | 66.7   | 871476  | 2024-10-31 | 99.74   | 6.72   | 99.66     | 99.69    |
| 60060 | SIDI-IFNI      | 49.5   | 817476  | 2024-09-30 | 95.91   | 31.12  | 91.92     | 94.37    |
| 60096 | DAKHLA         | 11.0   | 788604  | 2024-10-31 | 76.69   | 0.15   | 62.19     | 76.68    |
| 60101 | TANGER-AERO    | 16.0   | 871476  | 2023-12-31 | 99.87   | 5.14   | 98.51     | 99.86    |
| 60105 | LARACHE        | 70.0   | 854148  | 2023-12-31 | 72.68   | 3.71   | 90.07     | 58.17    |
| 60107 | EL-HOCEIMA     | 12.0   | 862236  | 2023-12-31 | 100.0   | 2.73   | 99.92     | 99.98    |
| 60115 | OUJDA          | 464.8  | 879564  | 2024-09-30 | 100.0   | 2.92   | 99.99     | 100.0    |
| 60120 | KENITRA        | 1.9    | 862236  | 2023-12-31 | 100.0   | 4.61   | 99.85     | 99.98    |
| 60127 | TAZA           | 509.2  | 644508  | 2021-12-31 | 99.87   | 4.96   | 54.15     | 97.13    |
| 60135 | RABAT-SALE     | 74.7   | 871476  | 2023-12-31 | 100.0   | 4.28   | 99.84     | 100.0    |
| 60136 | SIDI-SLIMANE   | 51.6   | 862236  | 2023-12-31 | 100.0   | 3.39   | 94.63     | 99.99    |
| 60141 | FES-SAISS      | 571.0  | 953196  | 2024-09-30 | 100.0   | 4.01   | 99.96     | 100.0    |
| 60146 | MOHAMMEDIA     | 4.3    | 843744  | 2024-09-30 | 66.62   | 2.68   | 99.15     | 66.55    |
| 60150 | MEKNES         | 556.0  | 889380  | 2024-09-30 | 100.0   | 4.6    | 99.99     | 100.0    |
| 60155 | CASA-ANFA      | 84.0   | 935004  | 2024-09-30 | 79.37   | 3.29   | 65.17     | 79.07    |
| 60156 | NOUASSEUR      | 200.0  | 919128  | 2024-09-30 | 99.78   | 2.95   | 99.62     | 99.77    |
| 60160 | IFRANE         | 1663.7 | 943956  | 2024-09-30 | 100.0   | 6.08   | 99.97     | 100.0    |
| 60165 | EL-JADIDA      | 27.0   | 897744  | 2024-09-30 | 79.59   | 4.16   | 99.34     | 73.88    |
| 60177 | SETTAT         | 407.5  | 907860  | 2024-09-30 | 61.05   | 6.12   | 60.45     | 60.75    |
| 60178 | KHOUIBGA-AERO  | 772.4  | 548064  | 2024-09-30 | 67.83   | 1.98   | 67.73     | 67.82    |
| 60185 | SAFI           | 43.0   | 925176  | 2024-09-30 | 71.65   | 2.5    | 97.7      | 71.65    |
| 60195 | MIDELT         | 1507.4 | 661836  | 2024-09-30 | 73.62   | 2.07   | 99.45     | 73.54    |
| 60200 | BOUARFA        | 1127.0 | 953196  | 2024-09-30 | 99.97   | 1.48   | 99.78     | 99.96    |
| 60210 | ERRACHIDIA     | 1048.4 | 580116  | 2024-09-30 | 99.94   | 1.62   | 96.17     | 99.94    |
| 60220 | ESSAOUIRA-PORT | 7.1    | 862524  | 2024-09-30 | 78.72   | 0.81   | 78.53     | 51.84    |
| 60230 | MARRAKECH      | 463.5  | 858768  | 2024-09-30 | 99.58   | 1.8    | 99.17     | 99.58    |
| 60250 | AGADIR         | 17.0   | 826140  | 2024-09-30 | 70.4    | 0.52   | 65.21     | 65.29    |
| 60252 | AIT-MELLOUL    | 74.0   | 835668  | 2024-09-30 | 100.0   | 1.09   | 99.64     | 100.0    |
| 60253 | TAROUDANT      | 263.0  | 816624  | 2024-09-30 | 99.99   | 37.2   | 99.99     | 99.99    |
| 60265 | OUARZAZATE     | 1131.0 | 543732  | 2024-09-30 | 99.99   | 1.12   | 58.91     | 99.97    |
| 60270 | TIZNIT         | 264.6  | 825840  | 2024-09-30 | 96.32   | 29.86  | 94.72     | 94.93    |
| 60280 | GUELMIM        | 332.0  | 844332  | 2024-09-30 | 79.94   | 0.96   | 79.07     | 79.9     |
| 60285 | TANTAN         | 192.0  | 835092  | 2024-09-30 | 73.81   | 0.69   | 71.59     | 64.73    |
| 60318 | TETOUAN        | 5.0    | 871476  | 2023-12-31 | 99.77   | 5.24   | 99.58     | 99.77    |
| 60335 | SMARA          | 170.0  | 853572  | 2024-10-31 | 57.17   | 1.5    | 23.04     | 57.14    |
| 60340 | NADOR          | 178.0  | 590508  | 2024-09-30 | 100.0   | 2.27   | 76.29     | 100.0    |

**Station distance to nearest reanalysis grid point, true km scale (39 stations)****Fig. 2** Distance of nearest grid point from observation station's location for ERA5 and ERA5-Land reanalysis.

## 4 Methodology

### 4.1 Verification workflow

The verification workflow is a two parts process (See Figure 3): continuous hourly verification in space, time, seasons and climate classes and daily categorical and continuous verification for extreme events based on operational local fixed thresholds. For the continuous hourly verification three parameters are assessed the temperature at 2m (T2m), wind speed at 10m (WS10m), relative humidity at 2m (RH2m). The choice of those parameters is justified by two main reasons: firstly, by their availability on profound archives of observations and in both reanalysis datasets, secondly those surface parameters characterize the main proprieties of the atmospheric conditions near the surface regarding heat, moisture and dynamics, they are also required for many applications and have impact in many strategic fields in Morocco such as agriculture, hydrology, aviation and marine traffic. The quality of those parameters in both datasets was evaluated daily and hourly using four key metrics: bias, RMSE, Pearson correlation and relative RMSE (see section 4.2 for more details). Regarding the daily categorical verification, 4 parameters are verified as the temperature at 2m daily minimum and maximum (T2m\_Min, T2m\_Max), wind speed at 10m daily maximum (WS10m\_Max) and finally the total daily precipitation (TP). Categorical extreme-event verification employs standard  $2 \times 2$  contingency-table metrics [40], including the probability of detection (POD), success ratio ( $SR = 1 - FAR$ ), frequency bias, and critical success index (CSI), which are summarized using Roebber performance diagrams [41] alongside relative operating characteristic (ROC) curves and the area under the curve (AUC) for daily precipitation exceeding 10mm. Because these scores are known to degenerate toward trivial values as the verified event becomes rarer [42], results are evaluated across the full range of sampled thresholds rather than at a single nominal value, with the threshold-dependence itself treated as a primary finding. Daily extremes are evaluated using station-specific climatological thresholds across a spectrum of percentiles and absolute values.

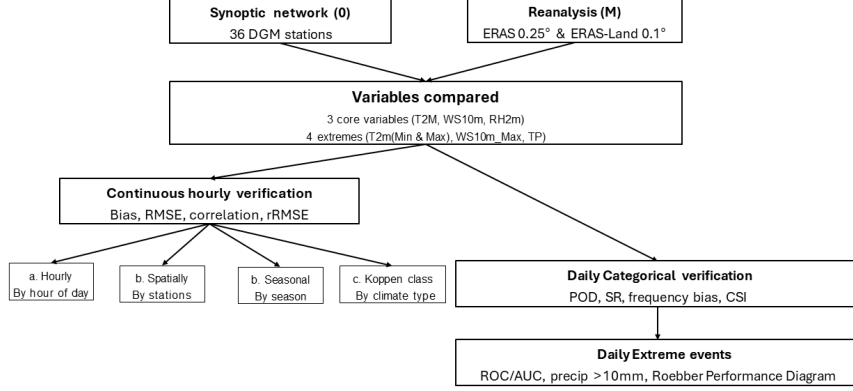


**Fig. 3** The validation workflow

## 4.2 Continuous verification

Continuous verification metrics evaluate the numerical error and linear association between continuous meteorological fields  $M_i$  and corresponding observations  $O_i$ .

### 4.2.1 Bias (Mean Error)

Bias measures the average tendency of the reanalysis model to overestimate or underestimate observed values. A bias of zero indicates an unbiased system, whereas positive (negative) values denote systematic overestimation (underestimation).

$$Bias = \frac{1}{N} \sum_1^N (M_i - O_i) \quad (1)$$

Where  $N$  is the total number of evaluation pairs.

### 4.2.2 Root Mean Square Error (RMSE)

RMSE calculates the sample standard deviation of the differences between model forecasts and observations. It gives disproportionately high weight to large errors due to the squaring operation, making it sensitive to outliers and extreme values.

$$RMSE = \sqrt{\frac{1}{N} \sum_1^N (M_i - O_i)^2} \quad (2)$$

#### 4.2.3 Relative Root Mean Square Error (Relative RMSE)

Relative RMSE normalises RMSE by the mean or standard deviation of observations (or model values), facilitating dimensionless comparisons across variables with different units or magnitudes.

$$rRMSE(\%) = (RMSE/\bar{O}) * 100\% \quad (3)$$

Where  $\bar{O}$  is the observation mean

#### 4.2.4 Pearson Correlation Coefficient (R)

Pearson correlation measures the strength and direction of the linear relationship between reanalysis estimates and observations, ranging from -1 to +1. It is insensitive to additive and multiplicative systematic biases.

$$R = \frac{\sum_1^N (M_i - \bar{M}) * (O_i - \bar{O})}{\sqrt{\sum_1^N (M_i - \bar{M})^2} * \sqrt{\sum_1^N (O_i - \bar{O})^2}} \quad (4)$$

### 4.3 Categorical (Binary) Verification & Contingency Tables

Categorical verification assesses how well a model predicts the occurrence or non-occurrence of discrete events (e.g., precipitation exceeding a specific threshold, frost, gale-force winds). Continuous or probabilistic fields are converted into binary outcomes using a defined threshold.

#### 4.3.1 The $2 \times 2$ Contingency Table

The  $2 \times 2$  contingency table is corner stone of categorical verification, which cross-classifies forecasts and observations into four distinct categories based on whether the event was forecasted and whether it occurred. From this verification tools one can extract a diversity of probabilistic verification scores (POD, FAR, CSI...)

**Table 2** Contingency Matrix

| Contingency Matrix   | Observed Yes (O = 1) | Observed No (O = 0)   |
|----------------------|----------------------|-----------------------|
| Forecast Yes (M = 1) | Hits (a)             | False Alarms (b)      |
| Forecast No (M = 0)  | Misses (c)           | Correct Negatives (d) |

#### 4.3.2 Probability of Detection (POD)

POD (Hit Rate) measures the fraction of observed events that were correctly forecasted by the model. Ranges from 0 to 1.

$$POD = \frac{a}{a + c} \quad (5)$$

#### 4.3.3 False Alarm Ratio (FAR)

FAR measures the fraction of model-forecasted events that did not actually occur in observations. Optimal value is 0.

$$FAR = \frac{b}{a + b} \quad (6)$$

#### 4.3.4 Success Ratio (SR)

SR represents the proportion of correct positive forecasts among all positive forecasts. It is the complement of the False Alarm Ratio.

$$SR = 1 - FAR = \frac{a}{a + b} \quad (7)$$

#### 4.3.5 Critical Success Index (CSI) / Threat Score

CSI measures the overall fraction of observed and/or forecasted events correctly predicted, penalizing both false alarms and misses.

$$CSI = \frac{a}{a + b + c} \quad (8)$$

#### 4.3.6 Frequency Bias

Frequency bias compares the total number of events forecasted to the total number observed, indicating systematic over-forecasting ( $>1$ ) or under-forecasting ( $<1$ ).

$$Frequency\ Bias = \frac{a + b}{a + c} \quad (9)$$

#### 4.3.7 Relative Operating Characteristic (ROC) Curve

The ROC curve plots the Probability of Detection (POD) on the vertical axis against the Probability of False Detection (POFD) across various decision thresholds. POFD is defined as:

$$POFD = \frac{b}{b + d} \quad (10)$$

#### 4.3.8 Area Under the ROC Curve (AUC)

AUC summarizes overall discriminative skill across all possible decision thresholds into a single scalar value from 0.5 (no skill) to 1.0 (perfect discrimination).

#### 4.3.9 Roebber Performance Diagrams

Roebber performance diagrams provide an advanced geometric framework specifically engineered to visualize dichotomous forecast performance without the visual bias associated with rare events. By plotting Success Ratio ( $SR = 1 - FAR$ ) on the x-axis against Probability of Detection (POD) on the y-axis, Roebber performance diagrams simultaneously display two critical diagnostic relationships:

- Critical Success Index (CSI): Represented as non-linear curved contours across the diagram, allowing direct reading of threat score skill levels.
- Frequency Bias: Represented as radial straight lines originating from the origin (0,0), enabling immediate identification of under forecasting or over forecasting tendencies.

Unlike traditional ROC space—where rare events can create an illusion of high performance due to the dominance of correct negatives (d)—Roebber performance diagrams omit correct negatives entirely. This makes them exceptionally powerful for evaluating high-impact extreme events (such as heavy daily precipitation thresholds  $> 10$  mm) where traditional metrics degenerate.

## 5 Results and Discussion

### 5.1 Hourly Validation

#### 5.1.1 Hourly Bias and RMSE of ERA5 and ERA5-Land

Figure 4 displays hourly 24 hours mean bias and RMSE of surface weather parameters (T2m, RH2m, WS10m) for single level ERA5 and ERA5 Land reanalysis for 36 stations over Morocco. Regarding 2-meter temperature (T2m), the ERA5 single-levels reanalysis performs best, exhibiting a bias near  $0^{\circ}\text{C}$  (with a maximum magnitude of  $0.5^{\circ}\text{C}$ ) at daytime hours. In contrast, ERA5-Land displays a mean bias of  $0.5^{\circ}\text{C}$ . In terms of RMSE, both configurations show a consistent value of approximately  $2^{\circ}\text{C}$ . For 2-meter relative humidity (RH2m), ERA5 single-levels show a near-zero mean bias for most hours, except for isolated peaks during daytime hours where bias magnitudes reach up to  $-4\%$ . Conversely, ERA5-Land RH2m bias ranges between  $-6\%$  and  $2\%$ . The corresponding RMSE values vary from  $10\%$  for ERA5 single-levels to  $12.5\%$  for ERA5-Land. For 10-meter wind speed (WS10m), ERA5 exhibits a maximum mean bias of  $-0.5$  m/s, outperforming ERA5-Land's  $-1.0$  m/s. The highest WS10m biases for both datasets are concentrated during the evening and night. However, both configurations yield highly comparable RMSE values of up to  $2$  m/s.

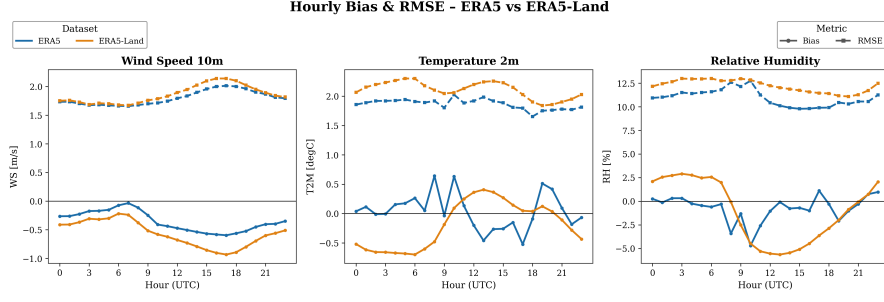


**Fig. 4** Hourly (24h) Mean Bias (solid dotted lines) and RMSE (dashed dotted lines) curves for 10m wind speed (WS10m), 2m temperature and humidity (T2m, RH2m) for ERA5 (blue) and ERA5-Land (orange)

### 5.1.2 Spatial verification of ERA5 and ERA5-Land

Figure 5 display the spatial distribution of ERA5 and ERA5-Land bias and RMSE spatial distribution for RH2m and T2m across the 36 study stations. For relative humidity (RH2m), both ERA5 and ERA5-Land exhibit extremely small biases and RMSE values, with nearly all stations falling within the  $\pm 10\%$  range. Both reanalysis products exhibit negative bias along central atlantic coast showing that reanalysis underestimate near surface moisture. This confirms that both reanalysis struggles capturing the sharp, microscale transition of the coastline boundary and daytime sea breezes. The models tend to underestimate the shallow, highly humid marine layer and its sharp boundary along the coastline. This localized dry bias is likely driven by the models' inability to fully resolve the cold coastal ocean upwelling, coupled with an over-penetration of warm, dry inland airmasses or a slight daytime warm bias in land-surface temperature within coastal grid cells [43]. The higher resolution of ERA5-Land does not confer any tangible advantage for humidity. Moving inland towards the Atlas foothills and eastward into continental/semi-arid zones, the bias shifts to a near-neutral or slight positive bias. The reanalysis captures the dry, continental air mass characteristics reasonably well, though it tends to slightly overestimate humidity in some interior pockets. The highest RMSE values (exceeding 15% to 20%) align perfectly with the regions of high negative bias along the Atlantic coast. The elevated error here points to strong diurnal or seasonal phase mismatches in humidity, which can be driven by the complex local wind regimes (sea/land breezes) and low-level marine stratus clouds that the reanalyses struggle to time correctly [44]. Interior stations show a more uniform, lower RMSE pattern ranging between 7.5% and 12.5%. In these regions, atmospheric moisture could be largely driven by larger-scale synoptic features rather than complex land-sea interfaces, making it easier for the reanalysis systems to simulate accurately [45, 46]. Although ERA5-Land boasts a higher spatial resolution (9 km vs. 31 km), it does not fundamentally correct the coastal humidity biases or lower the RMSE at these stations. This suggests that the errors are not purely a function of horizontal topography resolution. Instead, they are likely rooted in shared boundary layer physics parameterized in the driving model, soil-moisture-atmosphere coupling schemes, or the representation of the

cold Canary Current’s marine boundary layer influences. For temperature, the central Atlantic coast both reanalysis exhibit a noticeable positive temperature bias (rising up to  $2^{\circ}\text{C}$ ). The models are systematically overestimating surface temperatures here. Moving east of the Atlas chain and into the continental interior, the bias flips to negative (from  $-3^{\circ}\text{C}$  to  $-1^{\circ}\text{C}$ ), meaning the reanalysis underestimate heat in the interior. For temperature RMSE the highest RMSE values ( $2.5^{\circ}\text{C}$  to  $3.5^{\circ}\text{C}$ ) are concentrated directly along complex topography, and eastward interior flatlands. The Atlantic coast shows a simultaneous warm bias (T2m overestimated) and dry bias (RH2m underestimated). This strong negative correlation points to a structural issue in how the reanalysis handle the marine-terrestrial boundary. For wind speed (WS10 m) (not shown) :The maps show a more complex and spatially structured picture. Bias patterns (colour scale from  $-2$  to  $+2$  m/s) indicate that both ERA5 and ERA5-Land tend to under-estimate wind speed over the interior and mountainous regions and over-estimate along coastal areas and the Strait of Gibraltar. This dipole pattern is typical of reanalysis struggling to represent the interplay between large-scale flow and local orography. Importantly, ERA5-Land shows a clear improvement in RMSE over ERA5. While ERA5 exhibits RMSE values reaching up to 5 m/s in the Atlas Mountains and along the Atlantic coast, ERA5-Land reduces the RMSE to typically 4 m/s in the same regions. This reduction could be attributed to ERA5-Land’s higher spatial resolution ( 9 km vs. 31 km), which allows for a better representation of wind with finer land grid and the sharp land-sea contrast. The bias magnitude, however, remains similar between the two models – both display the same sign and approximate amplitude of systematic error, suggesting that the improvement in ERA5-Land lies mainly in better capturing the temporal variability (lower random error). Co-occurring warm (T2m overestimated) and dry (RH2m underestimated) biases along coastal margins—alongside anomalous 10-meter wind speeds—stem from grid-scale blending of land-sea masks that poorly resolves steep marine-terrestrial boundaries [47, 48]. This coarse spatial smoothing amplifies daytime land thermal responsiveness while diminishing adjacent water-vapor availability, elevating local temperatures and depressing relative humidity. Simultaneously, the muted surface roughness representation distorts coastal friction profiles, generating systematic errors in low-level wind speed behavior [49].

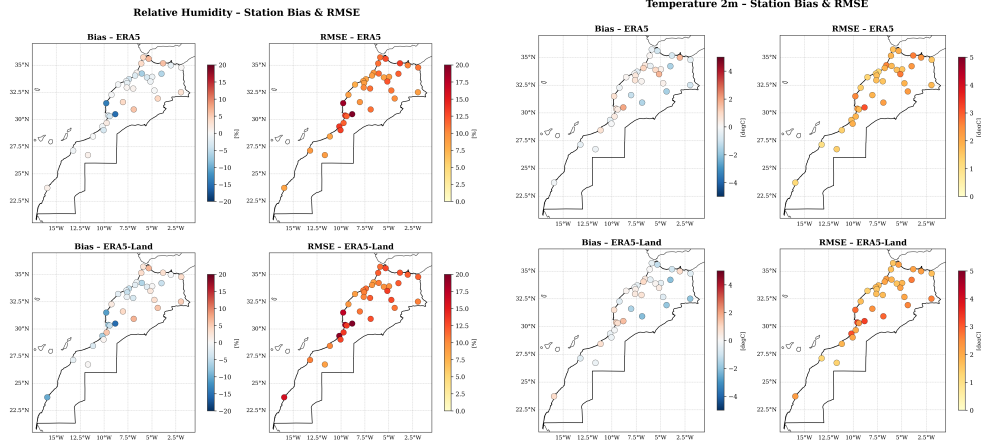


**Fig. 5** Mean bias and RMSE spatial distribution for T2m and RH2m for ERA5 and ERA5-Land over the 36 stations of Morocco

### 5.1.3 Diurnal verification by stations : rRMSE and Pearson Correlation

Figure 6 displays the hourly correlation and relative RMSE between ground observations and ERA5 and ERA5-Land reanalysis for all the studied stations. Regarding T2m correlation scores, most of the hours and stations display high values of correlation for both reanalysis with values exceeding 60%. The lowest values of correlation are displayed by central Atlantic coastline stations where there is a strong influence of sea land frontier micro-climates process. For relative RMSE scores values do not exceed 17.5% for both reanalysis, but high concentration of high relative RMSE values is concentrated within the night and early morning part of the day, with special focus of high scores of rRMSE on mountainous and Atlantic coastline stations. It is worth mentioning that ERA5 reanalysis provides better scores than ERA5-Land. For RH2m (not shown), both reanalysis display high correlation values along all stations, with slightly better performance of ERA5 compared to ERA5-Land. When focusing on relative RMSE worst performances are of 35% observed only for isolated stations, the worst performances are shown during the day part from 12h to 18h, with better performance of ERA5. For WS10m (not shown), both reanalysis display very close moderate correlation values around 60%. Regarding relative RMSE scores, poor performances with up to 100% scores are illustrated during the night phase, moderate relative scores less than 50% are noticed during the day phase (12h-18h). The pronounced contrast in error metrics—where relative RMSE spikes up to 100% during stable nighttime phases but drops below 50% during afternoon convective periods (12–18h)—reflects structural NWP limitations in resolving stable boundary layers [24]. During the night, radiative cooling generates shallow, strongly stable regimes characterized by low absolute wind and turbulence values; minor absolute model mismatches in these decoupled layers are mathematically magnified into massive relative percentage errors [50]. Conversely, daytime solar heating drives vigorous convective

mixing that overrides micro-meteorological noise, allowing global models to track large-scale atmospheric dynamics with significantly higher accuracy and lower relative uncertainties. Approaching 100% rRMSE at some stations could be consequence of rRMSE's denominator, since nocturnal mean wind speeds at these stations are themselves small, so even a modest absolute error translates into a large percentage.

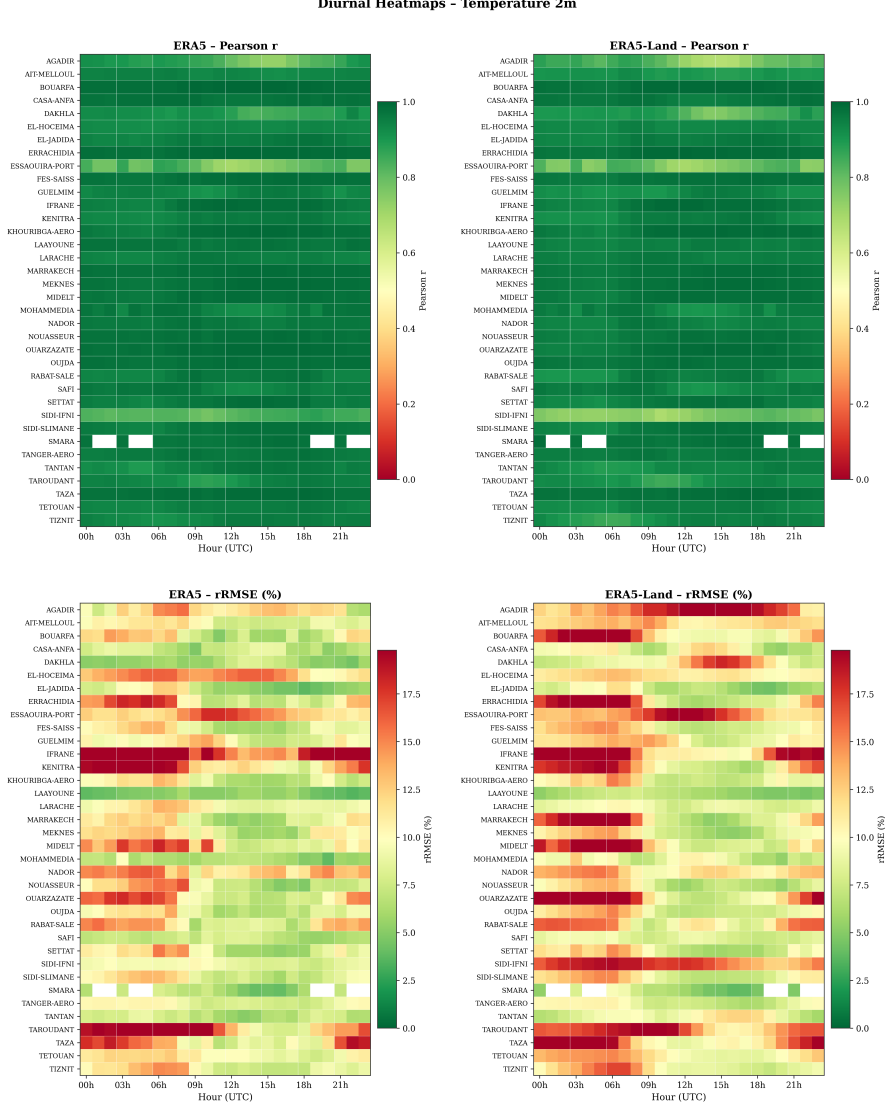


**Fig. 6** Diurnal T2M Pearson Correlation and Relative RMSE of ERA5 and ERA5-Land for all stations

#### 5.1.4 Diurnal Verification by Köppen Class

Figure 7 compares two reanalysis datasets—ERA5 and ERA5-Land—across five Köppen climate classes, evaluated at each hour of the day (UTC) using Pearson-r (correlation) and rRMSE % (relative error) metrics. Regarding relative humidity (RH2m), performance is substantially stronger and more regular, particularly in desert environments where BWh and BWk maintain robust correlations of 0.88–0.95 and 0.80–0.89 in ERA5 respectively, though ERA5-Land drops slightly to 0.84–0.91 in BWh. Crucially, the afternoon breakdown for BSk does not manifest in RH2m where BSk, BSh, and Csa all perform solidly within the 0.64–0.84 range, proving that the midday issue is strictly isolated to dew point representation rather than a generalized afternoon modeling failure. Relative humidity error structures are significantly low with the Mediterranean Csa class yielding the lowest errors overall at 13–18%, followed by BSh at 15–25%, BWh at 18–23%, and the cold semi-arid and desert zones (BSk and BWk) experiencing the highest errors at 21–35%. Across all classes, relative humidity errors peak around solar noon due to daytime convective mixing and decline into the evening, with ERA5-Land consistently displaying higher rRMSE values than ERA5, such as a 35% error peak compared to 32% in BWk. A cross-variable synthesis including 2m air temperature T2m establishes a clear predictability hierarchy where T2m significantly outperforms RH2m. Temperature stands out as the best-predicted variable, with all five classes sitting in an exceptional Pearson-r range of 0.87–0.99 and a low ERA5 rRMSE range of 6–17%. While ERA5 temperature errors drop to a minimum of 6–8% during the well-mixed afternoon, they peak in the early morning hours (3h to 9h) this may be due to the turbulent complexity of the decoupled nocturnal boundary layer, an error that roughly doubles in ERA5-Land where morning spikes reach 20–25% in BWk climates. Furthermore, across all moisture and temperature fields within these arid, semi-arid, and Mediterranean zones, the higher spatial resolution of ERA5-Land consistently introduces additional noise during stable nocturnal boundary layer conditions rather than improving near-surface atmospheric skill. For WS10m (not shown) the BSk and BSh zone experiences a striking middle day collapse, with correlations plummeting to 0.2–0.4 for ERA5 and ERA5-Land, suggesting a systematic failure to capture afternoon wind speed variability in cold and hot semi-arid regions. In contrast, BWh, BWk and Csa climates maintain moderate, stable correlations 0.65–0.8 across the day with no dramatic diurnal breakdown. For relative RMSE both reanalysis display high relative RMSE values going from 0.5 to 0.75% mainly concentrated during the early morning and night periods for BSk and BWk climate classes, showing the limited ability of both reanalysis in forecasting wind speed at 10m in cold arid and semi-arid zones. The best relative RMSE performances for both reanalysis were concentrated in the middle of day period for the BWh climate class.

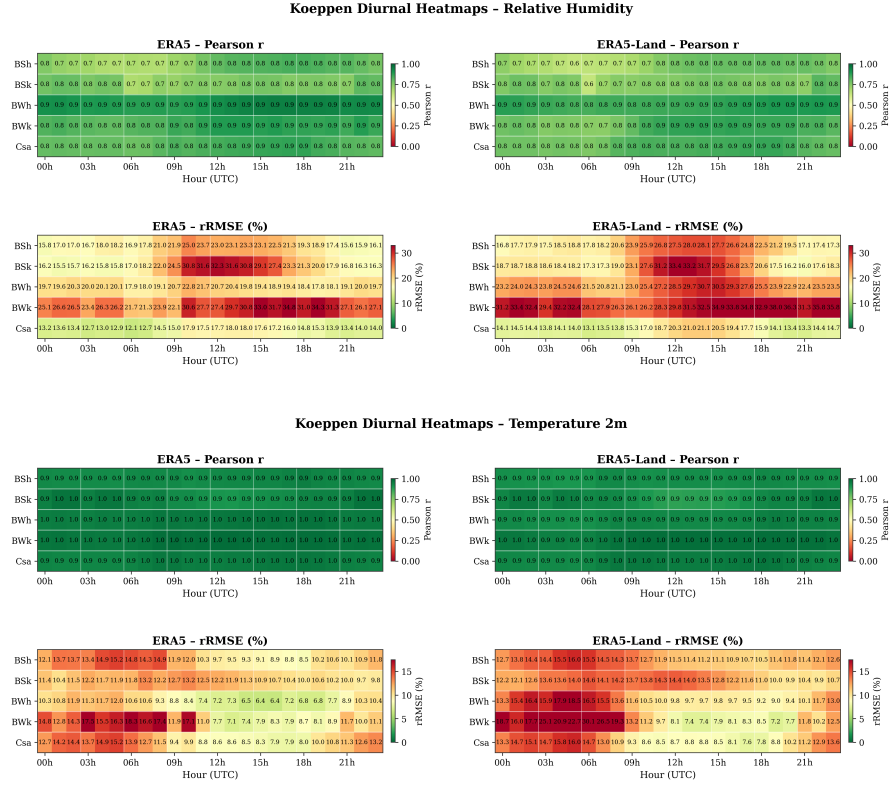


**Fig. 7** Diurnal Pearson Correlation and Relative RMSE of T2m, RH2m across five Köppen climate classes.

### 5.1.5 Seasonal Verification

Figure 8 displays the mean bias and RMSE by seasons for ERA5 (blue) and ERA5-Land (orange), for T2M, WS10m, and RH2m. Regarding WS10M, ERA5 shows the lowest bias distribution in Spring, Autumn and Winter with bias medians going from 0.2 to 0.3 m/s. For summer, ERA5 provide better performance compared to ERA5-Land with lowest bias median. It is to mention that ERA5-Land bias values are less dispersed compared to ERA5 showing longer ranges of bias values. For RMSE instead both reanalysis display close distribution over all seasons for WS10m. For T2M ERA5 provides the lowest bias for spring and winter, while for summer both reanalysis display close bias values, ERA5-Land shows large dispersed bias values for Autumn compared with ERA5 with values spanning from -0.25°C to 0.5°C, while ERA5 shows bias values around 0.5°C for the same season. Nevertheless, for RMSE values ERA5 shows the lowest errors values for all seasons with medians not exceeding 2°C. When considering moisture bias and RMSE distributions along seasons, generally ERA5 display lowest, well distributed values of bias and RMSE for all seasons compared to ERA5-Land.

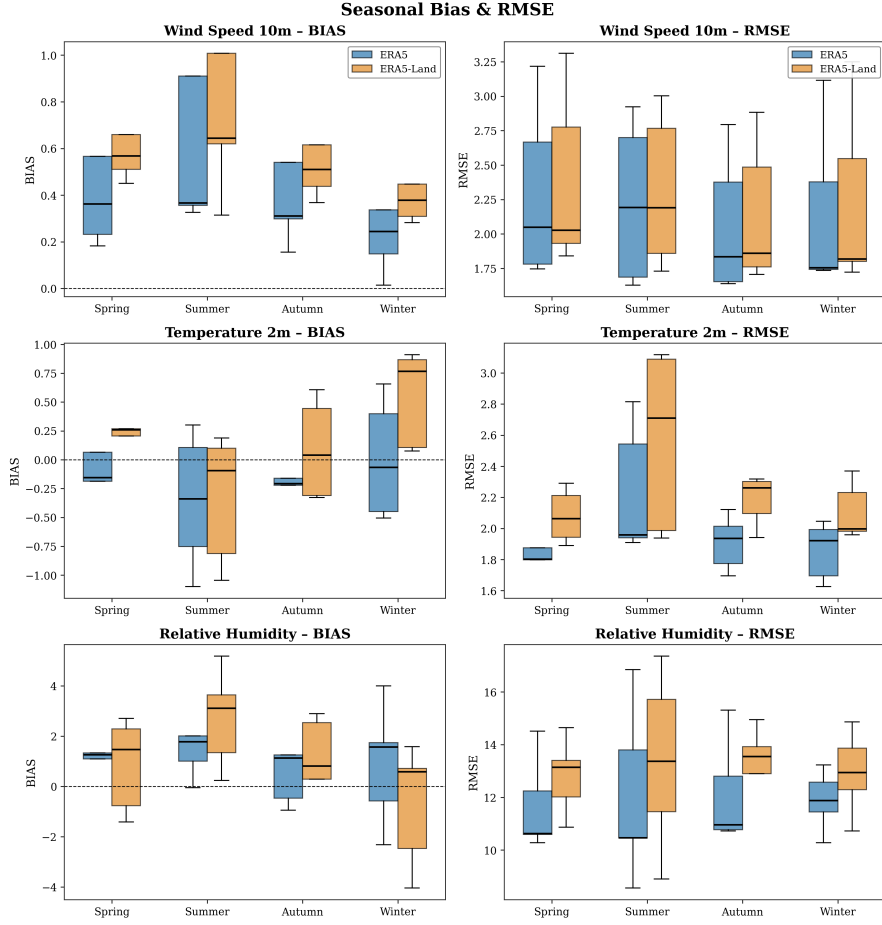


**Fig. 8** Seasonal bias and RMSE boxplots of ERA5 (blue boxes) and ERA5-Land (orange boxes) for T2M, WS10M, and RH2M.

## 5.2 Daily Extreme events verification

Figure 9 displays the global verifications scores of ERA5 and ERA5-Land for detecting extreme weather parameters including minimum and maximum temperature, max wind speed and total daily precipitation. Three scores are used the mean daily bias, RMSE and Pearson Correlation. Both reanalysis show very close scores for most of the extreme parameters. RMSE values of ERA5 and ERA5 Land are very close with a max difference of magnitude 0.27 m/s for wind speed max, for other parameter's this difference of errors is less than 0.02 unit. For correlation both ERA5 and ERA5-Land display very close and high correlations ( $>0.9$ ) for T2m\_MAX and T2m\_MIN, performance remains close but with low correlation (0.5) for WS\_MAX. For total precipitations, ERA5 displays higher values of correlation difference with

up to 0.18 compared with ERA5-Land. Regarding the mean bias ERA5 and ERA5-Land shows very close bias values for TP near zero (0.7mm) and for WS\_MAX with bias lower than -1.5m/s showing an underestimation of WS\_MAX. For temperature extremes, ERA5-Land shows better bias with -0.77 and 0.06 for T2m\_MAX and T2m\_MIN respectively. This affirms that both reanalysis are under-estimating maximum temperature and slightly over-estimates minimum temperatures.

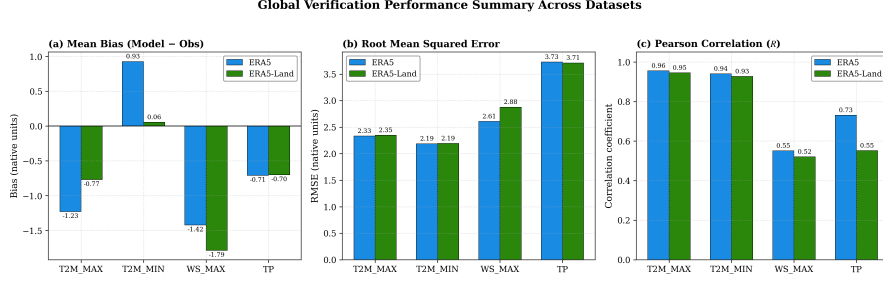


**Fig. 9** Overall verification barplots of ERA5 (blue bars) and ERA5-Land (green bars) for T2M\_MAX, T2M\_MIN, WS\_MAX and TP using mean bias, RMSE and Pearson Correlation

QQ-plots from Figure 10 confirms the previous findings, where ERA5-Land provides the closest curve to perfect for T2m\_MAX and T2m\_MIN with slight over-estimation of T2m\_Min and under-estimation of T2m\_MAX. Regarding maximum wind speed both reanalysis are under-estimating deeply the wind speed max.

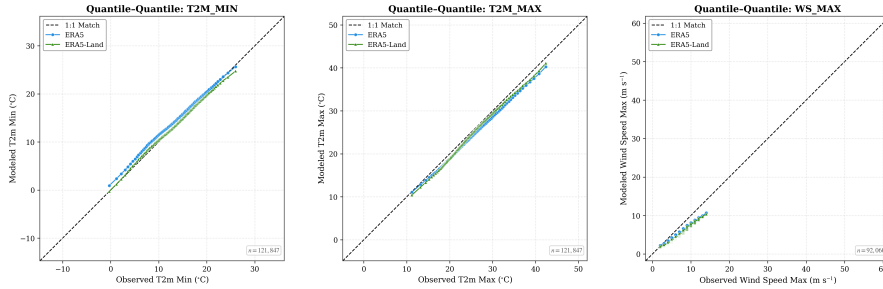


**Fig. 10** QQ-plots for ERA5(blue) and ERA5-Land (green) for T2M\_MIN, T2M\_MAX and WS\_MAX

Figure 11 displays the Roebber Performance Diagram evaluating the capability of ERA5 (circles) and ERA5-Land (triangles) to detect minimum and maximum 2-meter temperature (T2M\_MIN, T2M\_MAX) and maximum wind speed at 10m (WS10m) events across various temperature and wind speed thresholds (indicated by the color scale). A perfect model forecast would sit at the top-right corner (Probability of Detection POD = 1, Success Ratio = 1). The dashed lines represent

the Critical Success Index (CSI, labeled 0.1 to 0.9), and the straight diagonal lines represent the Frequency Bias (Bias or B, where  $B=1$  is unbiased). The performance diagrams for maximum and minimum temperature (T2M) and maximum wind speed (WS10m) provide a starkly different view from the continuous metrics. Focusing on the most quantitatively robust data—the T2M\_max threshold of  $35^{\circ}\text{C}$ —both ERA5 and ERA5 Land exhibit nearly identical and notably acceptable categorical skill. At this threshold, both reanalyses yielded a Probability of Detection (POD) of 0.6 and a Success Ratio ( $1 - \text{False Alarm Ratio}$ ) higher than 0.8. This results in a Frequency Bias of 0.5, indicating a slight systematic under forecasting of warm events (the models predict far fewer heat days than observed), and a Critical Success Index (CSI) of 0.6—a value reflecting good predictive utility for absolute station level extremes. The complete overlap of ERA5 and ERA5 Land points in the low skill quadrant of the diagram (low POD, low SR) for very extreme heat events (above  $45^{\circ}\text{C}$ ) is not an indictment of the reanalyses themselves, but rather a direct consequence of representativeness error. The verification of 2-meter minimum temperature over Morocco reveals a clear operational trade-off driven by the horizontal resolution differences between ERA5 and ERA5-Land. Evaluation via a Roebber performance diagram demonstrates that while both reanalysis products perform exceptionally well for mild night and morning minimums ( $10^{\circ}\text{C}$  to  $15^{\circ}\text{C}$ ), their accuracy significantly degrades for colder events near freezing ( $0^{\circ}\text{C}$  to  $5^{\circ}\text{C}$ ), where both configurations systematically over-predict the frequency of sharp cold drops. Intriguingly, the coarser 31 km ERA5 grid consistently achieves a higher Success Ratio by minimizing false alarms. In contrast, the 9km ERA5-Land grid yields a higher Probability of Detection (POD) across almost all temperature thresholds; however, this aggressive capability to resolve localized microclimates and topographic cold-air pooling comes at the expense of a lower Success Ratio, as the finer-scale drops generate a higher rate of spatial or temporal false alarms when validated against ground observation stations. Regarding wind speed maximum, ERA5 often shows low POD near zero for extreme wind mean max thresholds  $15\text{m/s}$  and moderate SR (0.2–0.4), placing it in the lower left quadrant. Bias is usually  $< 1$  (under forecasting extremes). For moderate wind mean max, ERA5 often shows higher POD (0.3–0.4) but slightly higher SR (0.7–0.8), placing it further up. Bias may be closer to 0.4. ERA5 is likely the better model for maximum wind speed because it detects more moderate wind events without a catastrophic increase in false alarms, leading to a higher CSI. Temperature and wind extremes at individual stations are overwhelmingly dictated by local micro topography, elevation, and land use contrasts—factors that cannot be resolved by either the 31 km ERA5 or even the higher resolution 9 km ERA5 Land grid. Crucially, this categorical deficiency stands in direct contrast to the continuous verification shown earlier. Those analyses demonstrated that ERA5 and ERA5 Land capture the mean state, diurnal cycles, and spatial variability of temperature with high correlation (often  $>0.9$ ) and low RMSE. Regarding ROC curves of daily total precipitation (bottom left corner illustration), while the ROC curve yields a high AUC score approaching 0.9 for both reanalysis, this strong performance is largely driven by the high frequency of non-occurrence observations. Because Morocco’s climate during the study period was characterized by prolonged, multi-year drought episodes, the

overwhelming abundance of true negative events artificially inflates the AUC-ROC, masking the model’s true skill in capturing distinct onset or termination transitions.

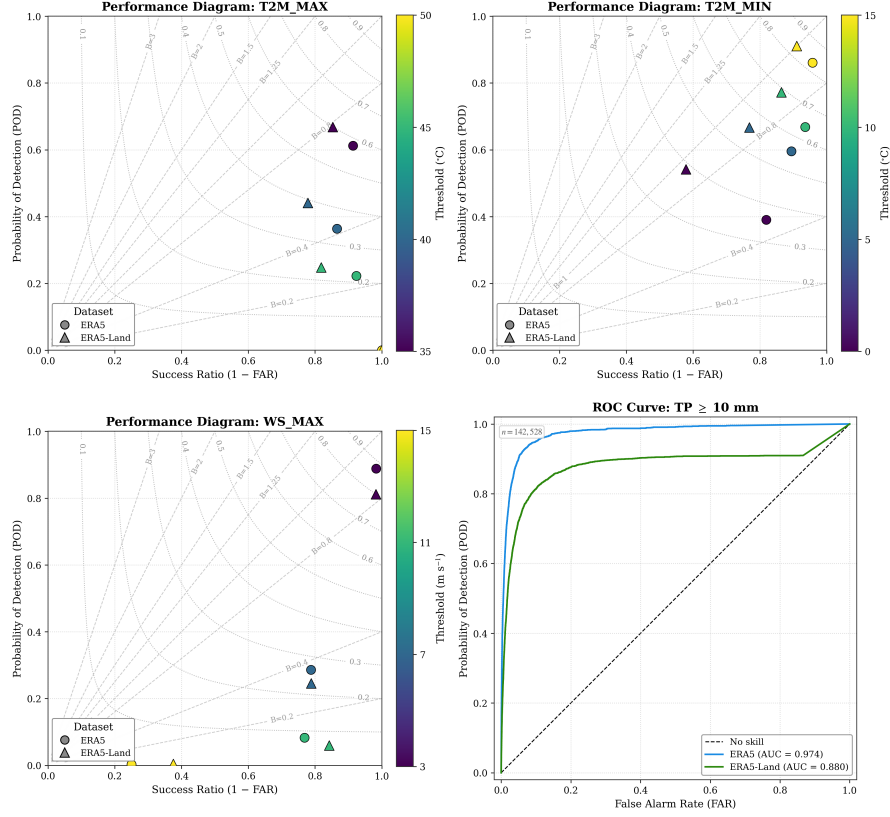


**Fig. 11** Roebber Performance Diagram evaluating the capability of ERA5 (circles) and ERA5-Land (triangles) to detect minimum and maximum 2-meter temperature (T2M\_MIN (up-right), T2M\_MAX (up-left)) and maximum wind speed at 10m (WS10m\_MAX (bottom-left)) events across various temperature and wind speed thresholds and ROC curve for total precipitation with 10mm threshold (bottom-right) (ERA5 (blue) and ERA5-Land (green))

In conclusion, while both reanalyses are equally robust for climatological and mean state applications, they are structurally unsuitable for deterministic, absolute threshold extreme event detection at the station scale over complex terrain. For such applications, we strongly recommend either using anomaly based thresholds (e.g., percentiles of the local climatology) or relying solely on continuous error metrics to evaluate model skill. Three threads run through the results above. First, resolution helps some variables and not others: ERA5-Land’s finer  $\sim 9$  km grid measurably reduces WS10m RMSE over the Atlas Mountains and the coast relative to ERA5’s  $\sim 31$  km grid, but confers no advantage—and in the cold-desert (BWk) class actively adds nocturnal noise—for T2m and RH2m, whose dominant

error is rooted in shared boundary-layer and soil-moisture-coupling parametrisations rather than in horizontal grid spacing per se. Second, both products are structurally limited, not by a flaw unique to either, but by station-scale representativeness: individual stations at Ifrane, Taroudant, Errachidia and Midelt show persistently high rRMSE across the whole diurnal cycle, and neither the coarse nor the fine grid resolves the local micro-topography responsible. Third, and most consequential for downstream use, categorical skill for daily extremes is not uniformly poor—it degrades specifically and monotonically as the verified threshold becomes rarer, for T2M\_MAX, T2M\_MIN and WS\_MAX alike. This is the expected signature of traditional POD/SR/CSI scores approaching their degenerate limits as the event frequency shrinks [42], and it means that a single reported skill value at an arbitrary threshold can be highly misleading about a reanalysis’ fitness for genuinely extreme-event applications; threshold-resolved reporting, as done here, or explicitly rare-event-robust scores (e.g. the Extremal/Symmetric Extremal Dependence Index), should be preferred in future work [42].

## 6 Conclusion and Perspectives

ERA5 and ERA5-Land both reproduce the mean state, diurnal cycle and spatial structure of near-surface variables over Morocco with high skill, and are broadly interchangeable for climatological applications, with ERA5 marginally ahead for temperature, humidity and wind bias, and ERA5-Land ahead via its finer grid, for wind-speed RMSE in complex terrain. Categorical verification of daily extremes shows that skill is threshold-dependent: moderate warm, cold and wind extremes are captured reasonably well by both products, but skill collapses specifically at the rarest, most extreme thresholds, a pattern attributable to the intrinsic behavior of standard contingency-table scores for rare events as much as to any deficiency of either reanalysis. We recommend that downstream applications – including the use of ERA5/ERA5-Land as a verification baseline for out-of-sample AI weather models – (i) adopt anomaly- or local-percentile-based extreme definitions, or rare-event-robust scores, rather than reporting skill at a single absolute threshold; (ii) always report the mean-normalised rRMSE (Eq. 3) alongside the absolute bias/RMSE, particularly for WS10m at low-wind stations, where a small denominator can inflate the percentage without a proportional change in absolute error; and (iii) explicitly audit and report ERA5-Land’s land-mask coverage before using it as a station-matched baseline over Morocco’s coastline. Planned extensions include a weather-regime stratification of the verification and a comparison against the World Meteorological Organization’s WIGOS/GBON station-density targets to quantify remaining representativeness gaps [19].

## References

- [1] Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz-Sabater, J., Nicolas, J., Peubey, C., Radu, R., Schepers, D., Simmons, A., Soci, C., Abdalla, S., Abellan, X., Balsamo, G., Bechtold, P., Biavati, G., Bidlot, J., Bonavita, M., De Chiara, G., Dahlgren, P., Dee, D., Diamantakis, M., Dragani,

- R., Flemming, J., Forbes, R., Fuentes, M., Geer, A., Haimberger, L., Healy, S., Hogan, R.J., Hólm, E., Janisková, M., Keeley, S., Laloyaux, P., Lopez, P., Lupu, C., Radnoti, G., Rosnay, P., Rozum, I., Freestone, F., Agustí-Panareda, A., Abreu, R., Albergel, C., Balsamo, G., Thépaut, J.N.: The ERA5 global reanalysis. *Q. J. R. Meteorol. Soc.* **146**(730), 1999–2049 (2020)
- [2] Muñoz-Sabater, J., Dutra, E., Agustí-Panareda, A., Albergel, C., Arduini, G., Balsamo, G., Boussetta, S., Choulga, M., Harrigan, S., Hersbach, H., Martens, B., Miralles, D.G., Piles, M., Rodríguez-Fernández, N.J., Zsoter, E., Buontempo, C., Thépaut, J.N.: ERA5-Land: A state-of-the-art global reanalysis dataset for land applications. *Earth Syst. Sci. Data* **13**(9), 4349–4383 (2021)
- [3] Lam, R., Sanchez-Gonzalez, A., Willson, M., Wyrnsberger, P., Fortunato, M., Alet, F., Ravuri, S., Ewalds, T., Aitchison, L., Blair, A., Mohamed, S., Battaglia, P.: Learning skillful medium-range global weather forecasting. *Science* **382**(6677), 1416–1421 (2023)
- [4] Bodnar, C., Bruinsma, W.P., Lucic, A., Stanley, M., Brandstetter, J., Garvan, P., Perdikaris, P.: Aurora: A foundation model of the atmosphere. Preprint at <https://arxiv.org/abs/2405.13063> (2024)
- [5] Bi, K., Xie, L., Zhang, H., Chen, X., Gu, X., Tian, Q.: Accurate medium-range global weather forecasting with 3D neural networks. *Nature* **619**(7970), 533–538 (2023)
- [6] Pathak, J., Subramanian, S., Harrington, P., Raja, S., Chattopadhyay, A., Mardani, M., Kurth, T., Hall, A., Li, Z., Kashinath, K., Anandkumar, A.: Four-CastNet: A global data-driven high-resolution weather model using adaptive Fourier neural operators. Preprint at <https://arxiv.org/abs/2202.11214> (2022)
- [7] Lang, S., Alexe, M., Chantry, M., Dramsch, J., Pinault, F., Raoult, B., Rabier, F.: AIFS–ECMWF’s data-driven forecasting system. Preprint at <https://arxiv.org/abs/2406.01465> (2024)
- [8] Keisler, R.: Forecasting global weather with graph neural networks. Preprint at <https://arxiv.org/abs/2202.07575> (2022)
- [9] Chen, S., Long, G., Jiang, J., Liu, D., Zhang, C.: Foundation models for weather and climate data understanding: A comprehensive survey. Preprint at <https://arxiv.org/abs/2312.03014> (2023)
- [10] Rasp, S., Hoyer, S., Meroze, A., Langmore, I., Battaglia, P., Russell, T., Sha, F.: WeatherBench 2: A benchmark for the next generation of data-driven global weather models. *J. Adv. Model. Earth Syst.* **16**(6), 2023–004019 (2024)
- [11] Bouallègue, Z.B., Clare, M.C., Magnusson, L., Gascon, E., Maier-Gerber, M., Janoušek, M., Pappenberger, F.: The rise of data-driven weather forecasting:

- A first statistical assessment of machine learning-based weather forecasts in an operational-like context. *Bull. Am. Meteorol. Soc.* **105**(6), 864–883 (2024)
- [12] Bouallègue, Z.B.: "What is a realistic forecast?" Assessing data-driven weather forecasts, a journey from verification to falsification. Preprint at <https://arxiv.org/abs/2602.00622> (2026)
  - [13] He, Y.H., Chan, P.W.: A study of the extended-range weather forecasts by artificial intelligence models Pangu-Weather and FengWu. *Weather* **80**(4), 134–138 (2025)
  - [14] Casati, B., Loveday, N., Bouallègue, Z.B., Bröcker, J., Wahl, S., Elía, R., McTaggart-Cowan, R.: Operational verification of artificial intelligence weather prediction. *Bull. Am. Meteorol. Soc.* **107**, (2026)
  - [15] Charlton-Perez, A.J., Dacre, H.F., Driscoll, S., Gray, S.L., Harvey, B., Harvey, N.J., Volonté, A.: Do AI models produce better weather forecasts than physics-based models? A quantitative evaluation case study of Storm Ciarán. *npj Clim. Atmos. Sci.* **7**(1), 93 (2024)
  - [16] Watson, P.A.: Machine learning applications for weather and climate need greater focus on extremes. *Environ. Res. Lett.* **17**(11), 111004 (2022)
  - [17] Olivetti, L., Messori, G.: Do data-driven models beat numerical models in forecasting weather extremes? A comparison of IFS HRES, Pangu-Weather, and GraphCast. *Geosci. Model Dev.* **17**(21), 7915–7962 (2024)
  - [18] Burgh-Day, C.O., Leeuwenburg, T.: Machine learning for numerical weather and climate modelling: a review. *Geosci. Model Dev.* **16**(22), 6433–6477 (2023)
  - [19] World Meteorological Organization: Guide to Instruments and Methods of Observation, WMO-No. 8. WMO, Geneva (2023)
  - [20] Merz, B., Kuhlicke, C., Kunz, M., Pittore, M., Babeyko, A., Bresch, D.N., Wurpts, A.: Impact forecasting to support emergency management of natural hazards. *Rev. Geophys.* **58**(4), 2020–000704 (2020)
  - [21] Kron, W., Löw, P., Kundzewicz, Z.W.: Changes in risk of extreme weather events in Europe. *Environ. Sci. Policy* **100**, 74–83 (2019)
  - [22] Liu, R., Zhang, X., Wang, W., Wang, Y., Liu, H., Ma, M., Tang, G.: Global-scale ERA5 product precipitation and temperature evaluation. *Ecol. Indic.* **166**, 112481 (2024)
  - [23] Molina, M.O., Gutiérrez, C., Sánchez, E.: Comparison of ERA5 surface wind speed climatologies over Europe with observations from the HadISD dataset. *Int. J. Climatol.* **41**(10), 4864–4878 (2021)

- [24] Urraca, R., Huld, T., Gracia-Amillo, A., Martinez-de-Pison, F.J., Kaspar, F., Sanz-Garcia, A.: Evaluation of global horizontal irradiance estimates from ERA5 and COSMO-REA6 reanalyses using ground and satellite-based data. *Sol. Energy* **164**, 339–354 (2018)
- [25] Li, D., Liu, Y., Luo, R., Tang, W., Gao, J., Tan, Z.: Validation and revision of low latitudes cloud base height from ERA5. *Atmos. Res.* **309**, 107595 (2024)
- [26] Dullaart, J.C., Muis, S., Bloemendaal, N., Aerts, J.C.: Advancing global storm surge modelling using the new ERA5 climate reanalysis. *Clim. Dyn.* **54**(1), 1007–1021 (2020)
- [27] Yavashi, D.D., Erlat, E.: Tropical nights in the Mediterranean: A spatiotemporal analysis of trends from 1950 to 2022. *Int. J. Climatol.* **44**(4), 1472–1488 (2024)
- [28] Aragão, L., Porcù, F.: Cyclonic activity in the Mediterranean region from a high-resolution perspective using ECMWF ERA5 dataset. *Clim. Dyn.* **58**(5), 1293–1310 (2022)
- [29] Vannucchi, V., Taddei, S., Capecchi, V., Bondoni, M., Brandini, C.: Dynamical downscaling of era5 data on the north-western Mediterranean sea: From atmosphere to high-resolution coastal wave climate. *J. Mar. Sci. Eng.* **9**(2), 208 (2021)
- [30] Nogueira, M., Boussetta, S., Balsamo, G., Albergel, C., Trigo, I.F., Johannsen, F., Miralles, D.G., Dutra, E.: Upgrading land-cover and vegetation seasonality in the ecmwf coupled system: Verification with fluxnet sites, meteosat satellite land surface temperatures, and era5 atmospheric reanalysis. *Journal of Geophysical Research: Atmospheres* **126**(15), 2020–034163 (2021)
- [31] El Khalki, E.M., Trambly, Y., Saidi, M.E., Marchane, A., Chehbouni, A.: Hydrological assessment of different satellite precipitation products in semi-arid basins in Morocco. *Front. Water* **5**, 1243251 (2023)
- [32] Ait Dhmane, L., Moustadraf, J., Rachdane, M., Saidi, M.E., Benjmel, K., Amraoui, F., Hadri, A.: Spatiotemporal assessment and correction of gridded precipitation products in North Western Morocco. *Atmosphere* **14**(8), 1239 (2023)
- [33] Chakri, A., Abakarim, S., Rodrigues, J.C.A., Laftouhi, N.E., Ibouh, H., Zouhri, L., Zaitseva, E.: Spatial bias correction of ERA5<sub>ag</sub> reanalysis precipitation using machine learning models in semi-arid region of Morocco. *Atmosphere* **16**(11), 1234 (2025)
- [34] El Kihel, B., El Kadri Elyamani, N.E., Chillali, A.: Evaluation of wind potential and economic prospects in the southern regions of Morocco with MERRA-2 and ERA5. *Wind Eng.* **49**(2), 475–498 (2025)

- [35] Rezkallah, K., Sebbar, A., Maugis, P., Krakauer, N.Y., Hammoudy, W., Badri, W.: Rainfall Variability and Droughts in the Sebou Basin (Morocco): ERA5 Bias Correction by Quantile Mapping and Trend Analysis over 1980–2024. *Appl. Sci.* **16**(14), 7023 (2026)
- [36] Rasp, S., Dueben, P.D., Scher, S., Weyn, J.A., Mouatadid, S., Thuerey, N.: WeatherBench: A benchmark data set for data-driven weather forecasting. *J. Adv. Model. Earth Syst.* **12**(11), 2020–002203 (2020)
- [37] Sarpong, R., Nazemi, A.: Benchmarking Snow Fields of ERA5-Land in the Northern Regions of North America. Preprint at <https://doi.org/10.5194/egusphere-2025-1> (2025)
- [38] Beck, H.E., Zimmermann, N.E., McVicar, T.R., Vergopolan, N., Berg, A., Wood, E.F.: Present and future Köppen-Geiger climate classification maps at 1-km resolution. *Sci. Data* **5**(1), 180214 (2018)
- [39] World Meteorological Organization: WIGOS / GBON: Global Basic Observing Network Technical Regulations and gap-analysis reports. WMO, Geneva (2021)
- [40] Wilks, D.S.: A skill score based on economic value for probability forecasts. *Meteorol. Appl.* **8**(2), 209–219 (2001)
- [41] Roebber, P.J.: Visualizing multiple measures of forecast quality. *Weather Forecast.* **24**(2), 601–608 (2009)
- [42] Ferro, C.A., Stephenson, D.B.: Extremal dependence indices: Improved verification measures for deterministic forecasts of rare binary events. *Weather Forecast.* **26**(5), 699–713 (2011)
- [43] Wang, Y., Warder, S.C., Benmoufok, E.F., Wynn, A., Buxton, O.R., Staffell, I., Piggott, M.D.: Geographic variability in reanalysis wind speed biases: A high-resolution bias correction approach for UK wind energy. *Energy Convers. Manage.* **352**, 121066 (2026)
- [44] North, R., Schumacher, C., Epifanio, C.C., Deranadyan, G., Putra, A.W., Adiyasa, A.: The land–sea breeze circulation over the West Coast of Sumatra. *Mon. Weather Rev.* **153**(11), 2295–2312 (2025)
- [45] Wood, R.R., Janzing, J., Hamel, A., Götte, J., Schumacher, D.L., Brunner, M.I.: Comparison of high-resolution climate reanalysis datasets for hydro-climatic impact studies. *Hydrol. Earth Syst. Sci.* **29**(17), 4153–4178 (2025)
- [46] Wei, Y., Sun, Y., Ma, Y., Tan, Y., Ren, X., Peng, K., Xin, J.: Deviations of Boundary Layer Height and Meteorological Parameters Between Ground-Based Remote Sensing and ERA5 over the Complex Terrain of the Mongolian Plateau. *Remote Sens.* **17**(3), 393 (2025)

- [47] Alkhalidi, M., Al-Dabbous, A., Al-Dabbous, S., Alzaid, D.: Evaluating the accuracy of the ERA5 model in predicting wind speeds across coastal and offshore regions. *J. Mar. Sci. Eng.* **13**(1), 149 (2025)
- [48] Patra, A., Oueslati, B., Chevallier, T., Renaud, P., Kervella, Y., Dubus, L.: Evaluation of ERA5, COSMO-REA6 and CERRA in simulating wind speed along the French coastline for wind energy applications. *Adv. Sci. Res.* **22**, 69–85 (2025)
- [49] Caton Harrison, T., King, J.C., Bracegirdle, T.J., Lu, H.: Dynamics of extreme wind events in the marine and terrestrial sectors of coastal Antarctica. *Q. J. R. Meteorol. Soc.* **150**(762), 2646–2666 (2024)
- [50] Cuxart, J., Holtslag, A.A., Beare, R.J., Bazile, E., Beljaars, A., Cheng, A., Xu, K.M.: Single-column model intercomparison for a stably stratified atmospheric boundary layer. *Boundary-Layer Meteorol.* **118**(2), 273–303 (2006)