Response to Reviewer 2

Overall assessment:

The manuscript entitled "Convective Environments Over the Arabian Peninsula in Current and Future Climates: A Machine Learning Approach" presents itself as the first part of a study aiming at inferring precipitation extremes over the Arabian Peninsula in a future warmer climate through a machine learning and statistical downscaling framework. In this manuscript, the authors presents their methods to detect "convective environments" using three different Machine Learning algorithms trained on more than a hundred of usual convective parameters including CAPE, vertical wind shear, or vertical velocity from ERA5 reanalysis which is validated against the satellite precipitation product IMERG. They then apply their algorithm to CMIP6 model ensembles to deduce spatio-temporal properties of convective environments in CMIP6 historical and future climate projections. They discuss biases between CMIP6 and ERA5 convective environment and projected changes in CMIP6 convective environments compared to current climate.

This is a well thought and promising project. The manuscript appears to contain a substantial amount of relevant analyses. However, I find the overall presentation difficult to follow. This may be partly a matter of individual reading experience, but I think it is sufficiently pronounced to warrant substantial restructuring. Otherwise the manuscript presents a substantial amount of unclear formulations and arguments.

Please find below my major and specific comments.

We thank the reviewer for the comments and suggestions. We have carefully considered each comment and provided our responses below. Our responses are displayed in dark blue, with clickable links highlighted in contrasting dark orange.

Major Comment:

I found that the large number of analyses and the way they are sequentially presented makes it difficult to identify the main scientific storyline. In several places, the manuscript moves from one analysis or figure to another without making sufficiently clear the interpretation of the results, what scientific question motivates the transition, and how the new analysis relates to the preceding result. This creates the impression of a succession of analyses rather than a progressively developing argument. I think the manuscript would benefit considerably from a clearer hierarchy between the main results and supporting analyses.

I also found the separation between the Results and Discussion sections problematic in this respect. The Results section contains a large number of observations, but the motivation and broader significance of these observations only become clearer later in the Discussion. The reader therefore has to retain and connect a large amount of information before understanding how the individual results contribute to the overall conclusions. I would therefore strongly encourage the authors to reconsider the structure of the manuscript.

In my view, the manuscript would be much more effective if it would include some interpretation/discussion of the results inside the main Results section.

If we understand the reviewer's comment correctly, we will improve the scientific storyline by adding transition sentences between sections and analyses to establish a progressively developing argument in the revised manuscript. Additionally, we will include introductory sentences at the beginning of section as follows:

> **2 Data and methodology**
>
> In this section, we first present a brief description of IMERG precipitation data, ERA5 reanalysis data, and the CMIP6 models used in the study (Sect. 2.1). Second, we provide the definition of CEs and their related convection indices, as well as the preprocessing, estimation, and bias correction of convection indices (Sect. 2.2). Third, we briefly describe the ML models used in the study (Sect. 2.3). Fourth, we present the evaluation metrics used to assess the ability of ML models to classify CEs and estimate CE probabilities (Sect. 2.4). Finally, we describe the procedure for training and calibrating the ML models (Sect. 2.5).
>
> **2.1 Data**
>
> . . .
>
> **3 Results and discussion**
>
> In this section, we first evaluate the ML models in terms of their ability to classify CEs and estimate CE probabilities (Sect. 3.1). Second, we present and discuss the CE probabilities and CE frequency from IMERG, ERA5, and CMIP6 multi-model ensemble under the current climate, as well as the projected changes in CE probabilities and CE frequency under *FutI* and *FutIII* (Sect. 3.2). Third, we look at the changes in CE-related convection indices (Sect. 3.3). Finally, we investigate the biases in CE-related convection indices in CMIP6 models compared to ERA5 under the current climate (Sect. 3.4).

As shown above, we will merge the current Results and Discussion sections into a single section in the revised manuscript to improve the reader's experience.

Specific Comments:

L. 1-3: This sentence feels a bit too absolute, as if dynamical downscaling were the only solution for adaptation.

We will revise this sentence to read:

> Large ensembles of high-resolution projections are important for adaptation, but their generation through dynamical downscaling remains computationally expensive.

L. 9-10: " excluding winter" , "excluding spring" → what happens for these seasons? Do CE occurrence increases/decreases or remains the same?

We will revise these sentences based on the revised ML framework:

> At an additional +1 °C regional warming, changes in CE frequency are dominated by increases in winter, summer, and at all times of day, but by decreases in spring and autumn. At an additional +3 °C, changes in CE frequency are dominated by decreases in spring, autumn, and at most times of day, but by increases in winter, summer, and at 21:00 UTC+3.

L. 10-12: I find this assertion ("shift towards a higher contribution of extremes to total rainfall") a bit too speculative, when just an increase in moisture and convective instability are provided as argument.

We agree that this interpretation could be presented more clearly in the Abstract. The argument is based on the projected increase in CE-related moisture and instability and the decrease in CE frequency at *FutIII*. This combination suggests fewer but potentially more intense CEs, lending support to an increase in the contribution of extremes to total rainfall. This shift has also been reported in previous studies (Almazroui and Saeed 2020; Rao et al. 2024; Alkhawaga and Mohamed 2026; Pathak et al. 2025; Al-Sakkaf et al. 2024; Luong et al. 2020). Nevertheless, because this contribution is not directly quantified in our analysis, we will remove this interpretation from the Abstract.

L. 16-17: Not clear what is "combined with regional aridity, high temperatures, and limited freshwater resources", please reformulate.

We will revise this sentence as:

> This rapid warming, combined with regional aridity, high temperatures, and limited freshwater resources, makes the AP especially vulnerable to the impacts of climate change.

L. 36-41: The difference between explicit and implicit approaches is not so clear, when convective indices could also be used in CPMs to derive convective environments. Please precise.

We thank the reviewer for raising this point. We will revise this part as:

> Convective environments, hereafter CEs, are synoptic- or meso-scale atmospheric conditions favourable for the initiation and organisation of (deep) convection (e.g., combination of instability, moisture, vertical shear, lifting mechanisms). CEs, their associated precipitation, and their response to warming can be investigated using models that represent convection either explicitly or implicitly. Explicit approaches rely on convection-permitting models (CPMs), which resolve convection at O(1 km) resolution. Implicit approaches rely on coarser climate models operating at O(100 km) resolution, where convection is parameterised. However, climate models in which convection is parameterised struggle to represent the diurnal cycle, intensity, frequency, and duration of precipitation (e.g., Hohenegger et al. 2008). Yet, CPMs are computationally expensive and therefore limited in their spatial and temporal coverage. In addition, continuous convection-permitting simulations may be particularly inefficient in arid regions, where precipitation events are infrequent. Consequently, two options emerge: relying solely on implicit approaches or identifying situations of interest from the coarse climate model output and then downscaling them using CPMs. The latter is known as statistical–dynamical or case-selective downscaling (e.g., Beaulant et al. 2011; Meredith et al. 2018; Wang et al. 2026).

L. 42-44: Here, the reader would expect the authors to specify which approach is adopted in this manuscript.

We respectfully disagree with this suggestion. Our approach is already presented starting at L. 67, after the reader has been introduced to the information and context to understand the study objectives. Introducing our approach here would disrupt the flow for the reader. However, we will explicitly state that this study rely on models with implicit presentation of convection.

L. 80: The term "agglomeration" is somewhat unusual in this context and may be unclear to readers + the second part of this sentence is unclear with "resolved" referring to "questions" instead of "precipitation".

We will revise this sentence as:

> The present approach focuses on inferring the spatiotemporal distribution of CEs over the region, while leaving precipitation to be physically resolved using CPMs.

L. 106-107: Not clear to me, are pressure levels less numerous?

Yes. ERA5 pressure-level data are available on 37 levels (`https://cds.climate.copernicus.eu/datasets/reanalysis-era5-pressure-levels?tab=overview`), compared with the 137 hybrid model levels. We will add this information in the revised manuscript.

Section 2.1: It would be good to reconnect each data set to its purpose in this manuscript.

We respectfully disagree with this suggestion. The purpose of the datasets is already explicitly stated in the Introduction at L. 67.

L. 115-117: Please provide the number of hybrid levels (optionally compared to the number of pressure levels) to make this sentence clearer.

The number of vertical levels for each CMIP6 model is already provided in Table 1. We will add a reference to Table 1 in this sentence and use "hybrid model levels" consistently to improve clarity.

L. 121-124. With the current phrasing, it is not so clear whether you performed the calculation to retrieve the average regional warming or if you used a value from the literature

We will revise the phrasing to indicate that we estimated these values as follow:

> Since the IMERG data considered in the study extend to 2024, whereas CMIP6 historical simulations end in 2014, the datasets cannot be compared over the same full period. Moreover, differences in warming rates among CMIP6 models (e.g., Hausfather et al. 2022) and relative to observations mean that the same calendar periods may represent different climate states. Therefore, we followed a regional warming approach to minimise these differences by considering periods in CMIP6 simulations with average regional warming close to that observed during the IMERG period. We estimated the average regional warming of the AP for the period 2001 - 2024 relative to the preindustrial level (1850 - 1900) at +1.22 °C, based on the area-weighted average of Berkeley Earth (Rohde and Hausfather 2020), GISTEMP (GISTEMP Team 2025; Lenssen et al. 2024), HadCRUT5 (Morice et al. 2021), and NOAA Global

Temp (Huang et al. 2024) datasets (see supporting information Text S1). For CMIP6 data, we used the annual regional temperature anomalies relative to the preindustrial level to calculate a 24-year running average (matching the length of the IMERG record) and then selected periods with average regional warming closest to +1.22 °C. If the relevant period overlaps with the SSP5-8.5 scenario, only years from the historical simulation up to 2014 were used (Table 2), avoiding mixture of experiments, as noted by Lepore et al. (2021). Due to this truncation, the warming of the final period may differ from the target warming level; however, this is a trade-off between matching the target warming level and avoiding the combination of historical and SSP5-8.5 simulations.

In addition, we analysed two periods from the SSP5-8.5 scenario corresponding to an additional +1 °C (*FutI*) and +3 °C (*FutIII*) of regional warming relative to the current level, equivalent to regional warming levels of +2.22 °C and +4.22 °C above preindustrial levels, respectively. We chose a regional warming approach because regional warming rates are 1.4–2.1 times those in the tropics and 2.3–3.1 times the global mean warming rate (Cook et al. 2021). In addition, extreme precipitation (temperature) over the region is projected to increase at a slower (faster) rate than the global land (Seneviratne et al. 2016). According to Corre et al. (2025), regional warming levels are more suitable for adaptation planning when global warming levels differ from regional ones (Table S1[new order]). Finally, we excluded the MRI2 and CAN5 models from the analysis due to their short retained periods (Table 2) and large deviations of -0.277 °C and +0.554 °C from the target warming levels of +1.22 °C and +2.22 °C, respectively.

L. 124-125: Not clear in this sentence that you defined the periods so that they reach a predefined averaged warming value. In addition, the reason why this approach was adopted is missing here.

Please see our response to L. 121-124.

L. 125-126: When truncating at 2014, are the years post 2014 dropped from the calculation of the regional warming or not, implying that the final truncated period would not exactly reach the set warming of 1.22° C compared to preindustrial levels?

The years after 2014 were included in the initial calculation of the 24-year running average used to identify periods with regional warming closest to +1.22 °C. However, these years were excluded from the subsequent analysis to avoid combining historical and SSP5-8.5 simulations. Consequently, the average regional warming of the retained periods may differ from the target warming level. We will clarify this distinction and the associated trade-off in the revised manuscript, as shown in L. 121-124. In the revised manuscript and analysis, we also exclude two models from the revised analysis due to the large deviation from the target warming level and the short retained periods (see our response to L. 121-124).

L. 133: Table S2 is mentioned before table S1. Please correct here and in other instances where the order is not correct.

This will be corrected throughout the revised manuscript, as in our response to L. 121-124.

L. 142-146: It is not clear why the data was averaged over a 1° x 1° grid (when convection is usually more local), why the local median (both local and median) was chosen for detecting convective precipitation (why e.g. not another percentile defined over the whole AP). The argument about virga is also not clear as it is not precised that IMERG could potentially identify virga as precipitation reaching the surface, and probably refers more to the 1 mm/6h threshold than to the median.

We thank the reviewer for these questions and notes. If we upscaled IMERG data (0.1°, 0.5 h) only to ERA5 spatial and temporal resolution (0.25°, 1 h) and trained ML models at these resolutions, applying the ML models to the coarser and less frequent CMIP6 data would introduce a mismatch. Finer-resolution data better capture small-scale variability in convective environments, including localized extremes in instability and moisture (e.g., Taszarek et al. 2021). Consequently, these ML models could have learned from possibly more frequent and intense convective environments at the ERA5 scale that may not be present at the CMIP6 scale.

We use a local rather than AP-wide percentile because convective precipitation and its controlling mechanisms are spatiotemporally heterogeneous across the region. Luong et al. (2024) identified distinct modes of organized rainfall over Saudi Arabia. They reported northern, southern and dry modes of rainfall with different contributions from tropical, extratropical, and local processes and different precipitation characteristics; see their Fig. 1. The lifecycle and intensity of convective systems also vary substantially across the AP and seasons (Homoudi et al. 2026, Fig. 2, 3, and 7). A single AP-wide percentile would therefore assign the same precipitation threshold across regions with very different climatological regimes, whereas a local percentile (Fig. 1 in the manuscript) adapts the CE definition to the regional precipitation distribution.

We agree that the discussion of virga was not clear. We will revise this paragraph as follows:

> We estimated the 6 h accumulated precipitation from IMERG data centred at timestamps matching the CMIP6 data (i.e., 00:00, 06:00, 12:00, and 18:00 UTC). Then we averaged the accumulated precipitation over a 1°×1°grid cell. Since ML models are used to identify CEs in CMIP6 models, the accumulation and averaging procedure ensures that the ML models are trained using data at approximately the same spatiotemporal scale as the CMIP6 data (∼1°×∼1°, 6 h). A grid cell was classified as convective if the accumulated average precipitation exceeded the local climatological median. Passive-microwave retrievals used in IMERG can potentially misidentify virga as precipitation reaching the surface (Wang et al. 2018). Therefore, to reduce this potential contamination, only precipitation values $\geq 1\,\mathrm{mm\,6\,h^{-1}}$ were included in percentile estimation. Figure 1 illustrates the median precipitation and the fraction of time rainfall exceeds $1\,\mathrm{mm\,6\,h^{-1}}$.
>
> We adopted a percentile-based approach because it provides a simple and adjustable method that can be used to investigate different intensities across the precipitation distribution. In principle, lightning or radar observations could provide a more direct identification of convection, but freely accessible, sufficiently long, and spatially consistent records are not available over the AP. Multiple percentiles between the 99$^{\text{th}}$ and 50$^{\text{th}}$ were tested, and the median showed the best performance (Text S2$^{\text{new order}}$ and Fig. S1). We constrained our search to the

50th percentile because we are mainly interested in CEs associated with precipitation above the local climatological median. These CEs can be of interest for a subsequent dynamical downscaling and impact studies, such as flash-flood assessments. Additionally, we centred the precipitation accumulations at the timestamps because we also tested accumulations starting at the timestamps, which yielded poorer classification results (Text S2$^{\text{new order}}$ and Fig. S1). Furthermore, we estimated the percentiles locally given the heterogeneous characteristics of convective precipitation and the controlling mechanisms in the region (Luong et al. 2024; Homoudi et al. 2026), thereby avoiding a mismatch between the local precipitation climatology and the associated atmospheric conditions.

L 149: Are the instantaneous convective indices computed at e.g. 0600 to predict convective rainfall occurrence centered at 0600 (that is, for IMERG, precipitation falling from 0300 to 0900) or there is a time offset between the predictor and the predictand? It should be considered that convective rainfall may feedback negatively on many convective predictors (e.g. CAPE).

Yes. A profile at 0600 is used to identify precipitation (above the median) occurring between 0300 and 0900, and there is no time offset between the predictor and predictand. We considered the potential negative feedback of convection on the predictors when defining the predictand in Fig. S1. Specifically, we tested precipitation accumulations centred on the profile timestep (M) and accumulations starting at the profile timestep (S). For example, a profile at 0600 is paired with precipitation accumulated from 0600 to 1200 for the S definition. Across the tested precipitation percentiles, the M accumulation yielded more favourable evaluation metrics than the corresponding S accumulation. However, the potential negative feedback of convective rainfall on the predictors remains physically plausible.

L 149-151 : Here as well I am not totally convinced with the method using relative thresholds instead of absolute thresholds and of the justification to enhance the signal of convection.

Take for example the lifted index (LI), if the monthly-diurnal average is e.g. +10 °C (so overally stable), but one day it is +5 °C then it is a strong negative anomaly of -5 °C but it is still stable. Then this day will be given the same LI predictor than a day with -10 °C in a very unstable month/hour of average -5 °C. Yet, in the former configuration, I see hardly any possibilities for deep convection compared to the later configuration.

Besides, if I am not mistaking, the convective precipitation threshold (defined as the median) is defined from the whole climatology, and is not made relative to the hour of the day or the month in the year. Therefore predictor and predictand do not really match in my opinion.

We thank the reviewer for this comment, and we agree with the physical interpretation illustrated by the LI example. To address this concern, we repeated the entire ML framework using the absolute values of the predictors instead of anomalies. The revised manuscript and analyses will be based on absolute predictor values. This also removes the mismatch noted by the reviewer between the predictors and predictand.

For XGB, training on absolute predictor values slightly worsens FAR, HSS, and AUCPR, while POD and BSS remain unchanged (Table 1). For RF, training on absolute predictors slightly worsens POD and HSS but improves FAR, while BSS and AUCPR remain unchanged. DL generally benefited from training

on anomalies, which was the original motivation for removing the monthly-diurnal cycle. Comparing the revised percentile analysis using absolute predictors (Fig. 1) with the previous percentile analysis based on anomalies (Fig. 2) shows that DL generally performs better with anomalies, particularly at the 99th percentile with centred precipitation accumulation (M), where POD and HSS decrease to zero when absolute predictors are used. Although our application is not a forecasting problem, Nelson et al. (1999) showed that neural networks provide more accurate forecast when the seasonal cycle is removed from the input data.

Table 1: Evaluation metrics for the optimised DL, RF, and XGB models; trained on absolute and anomaly predictors with the local climatological median as the convective precipitation threshold. Metrics are estimated from the adjusted and calibrated probabilities from the ERA5 test data using 300 bootstrap samples.

	DL		RF		XGB	
Metric	Absolute	Anomalies	Absolute	Anomalies	Absolute	Anomalies
POD	0.55	0.56	0.55	0.56	0.58	0.58
FAR	0.55	0.52	0.51	0.53	0.47	0.46
HSS	0.49	0.51	0.50	0.51	0.55	0.56
BSS	0.33	0.36	0.35	0.35	0.39	0.39
AUCPR	0.50	0.54	0.53	0.53	0.59	0.60

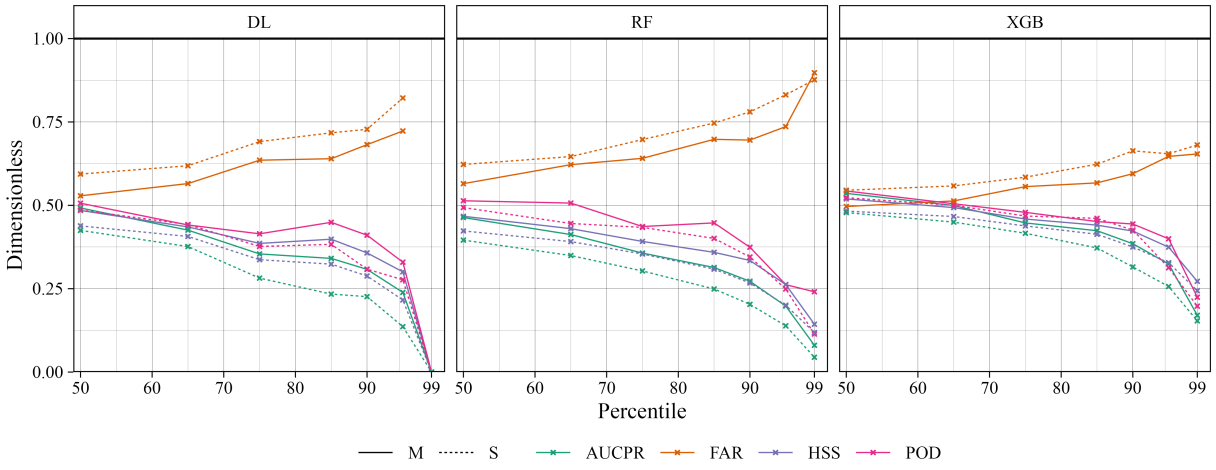


Figure 1: (new) Evaluation metrics of DL (left panel), RF (middle panel), and XGB (right panel) across 14 label variations for CEs (7 precipitation percentiles × 2 accumulation options). Precipitation accumulation is defined either at the middle of the time window (M; solid line with cross markers) or at the start of the time window (S; dashed line with cross markers). Performance metrics are AUCPR (teal), FAR (orange), HSS (purple), POD (pink). Missing points are undefined values.

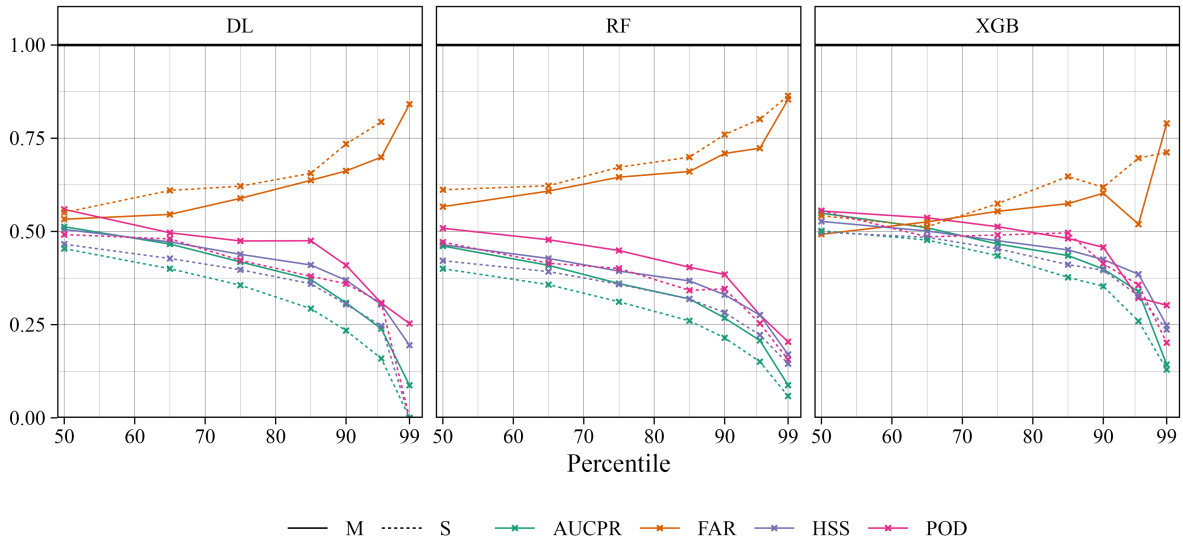


Figure 2: (old) Performance metrics of DL, RF, and XGB across 14 label variations for CEs. Considering precipitation accumulation at the middle (M) or the start (S) of the time window.

L. 153: Please specify the interpolation method.

We will revise this part and add the vertical and spatial interpolation method as follows:

> For CMIP6 data on hybrid model levels, we first vertically interpolated the data to pressure levels at 10 hPa increments using the methodology described in Trenberth et al. (1993). Then, we calculated and regridded convection indices onto a 1°×1°grid using bilinear interpolation from the Climate Data Operators (`CDO`; Schulzweida 2023). Generally, the atmospheric profiles are restricted to pressure levels between 1000 and 100 hPa for both ERA5 and CMIP6, unless explicitly stated otherwise (Table S2new order). Our `R`-package `MeteoMate` performs vertical interpolation and estimates convection indices efficiently by performing its core calculations in Fortran with OpenMP parallelism (Homoudi and Barfus 2025).

L 157: The text does not say the used reference data set for the MBCn (I assume ERA5).

We will add that ERA5 is the reference data in the revised manuscript.

L. 182-83: The threshold used to convert predicted probabilities to a binary classifier could be precised here (I see it is only defined later in the text) for more clarity.

We will revise this part as follows:

> This is obtained by converting predicted probabilities to binary predictions using probability thresholds of 0.25, 0.27, and 0.28, for DL, RF, and XGB, respectively.

L. 188-190: The letters “h”, “q”, “f”, and “m” need to be defined.

Although this information is already provided in Table 3, we will also include them at L. 183:

The contingency table (Table 3) compares these binary predictions with actual labels from IMERG data, where hit (h) is the number of CEs correctly classified as CEs, false alarm (f) is the number of non-CEs incorrectly classified as CEs, miss (m) is the number of CEs incorrectly classified as non-CEs, and quiescent (q) is the number of non-CEs correctly classified as non-CEs (e.g., Wilks 2019).

L. 198: As mentioned, this could have been precised earlier in the text. Please also precise the data set used to calculate the HSS (whole climatology, training, test, or validation).

We thank the reviewer for this suggestion. We acknowledge that the information at L. 198 appears in the wrong place and is incomplete; however, we respectfully disagree that the complete procedure for defining the threshold probability should be specified at L. 198 or earlier. This information will be moved from L. 198 to Sect. 2.5, after the relevant concepts of calibration, the validation data, and HSS are introduced, as in our response to L. 210. The threshold values themselves will nevertheless be reported earlier in the text, as in our response to L. 182-83.

L. 208: It is not clear which imbalance do you refer to, and the reason why it is mentioned here. Please specify.

The imbalance clause will be removed in the revised manuscript, as in our response to L. 210.

L. 210: Please clarify if the frequency of CEs in each parts (a, b, c) is the same as the whole dataset (i.e. equal to 0.66%).

As already stated in the manuscript, the splitting is stratified to approximately preserve the original distribution of CEs across the subsets. Nevertheless, we will provide the exact CE prevalence in each subset in the revised manuscript. Sect. 2.5 will read as follows:

In the IMERG-ERA5 dataset, the prevalence (i.e., the climatological probability) of CEs is 0.66%. We split the dataset into three subsets: a) 60% for training (5.2 million samples), b) 20% for validation and hyperparameter tuning (1.7 million), and c) the remaining 20% for final testing and evaluation (1.7 million). The dataset was split using stratified sampling to approximately preserve the original prevalence of CEs (Dankers et al. 2019); therefore, the prevalence is 0.66%, 0.66%, and 0.65% in the training, validation, and testing data, respectively.

The rarity of CEs in the IMERG-ERA5 dataset results in a significant imbalance between CEs and non-CEs. This imbalance complicates predicting CE occurrence, as ML models struggle to learn from imbalanced datasets (e.g., Sun et al. 2009; Vahid Yousefnia et al. 2024). Furthermore, the size of the training dataset exceeds available memory for RF and XGB; therefore, we sampled 50% of the non-CEs and oversampled CEs with replacement (e.g., Hoens and Chawla 2013), increasing the CE prevalence to 2.9% and 25.9% in the training data for RF and XGB, respectively. These values were obtained from a sensitivity experiment to determine the optimal fraction of CEs to include in the training data. For further details, see supporting information (Text S4$^{(tbc)}$ and Fig. S2).

The computational and memory requirements differ among the ML models. For RF, these requirements increase with the number and size of trees, with tree size generally inversely proportional to the minimum node size and proportional to the number of training samples (Breiman 2001; Wright and Ziegler 2017). For XGB, these requirements depend on the size of the training data and the number of trees, while the size of individual trees is generally controlled by the `max_depth` parameter (Chen and Guestrin 2016), potentially resulting in lower model memory requirements than RF. In contrast, DL uses a fixed network architecture and processes the training data in batches; therefore, memory requirements mainly depend on the network architecture and batch size rather than the total number of samples, allowing DL to process the complete training data. Training in batches allows DL to learn from all available CEs, whereas resampling during training for RF and XGB may result in some trees having insufficient CEs for adequate learning. In addition, we used the `BinaryFocalCrossentropy` loss function (Lin et al. 2018) to down-weight easy examples, particularly the abundant non-CEs, and focus training on difficult examples.

Hyperparameters are the values that control the model's learning process. Optimising these parameters is necessary since the default values are not universally optimal and can lead to suboptimal results. We tuned the hyperparameters and model structures for RF, XGB, and DL using Bayesian optimisation (Jones et al. 1998). The optimal hyperparameters for RF and XGB are shown in Tables S3 and S4, respectively, and the DL model architecture and hyperparameters are shown in Table S5.

Since we altered the CE prevalence in the training data for RF and XGB, the predicted probabilities may reflect the altered CE prevalence rather than the original CE prevalence (0.66%). Nonetheless, this mismatch can be corrected through an analytical adjustment of the probabilities (Pozzolo et al. 2015; Vahid Yousefnia et al. 2024). The adjusted probability estimate is given as

$$f(\xi, g) = \frac{f(\xi, \hat{g})}{f(\xi, \hat{g}) + \frac{1-g}{g}\frac{\hat{g}}{1-\hat{g}}(1 - f(\xi, \hat{g}))}, \tag{2}$$

where g is the climatological probability of CEs (0.66%), $\hat{g}$ is the CE prevalence in the training data (2.9% and 25.9% for RF and XGB, respectively), and $f(\xi, \hat{g})$ is the model output probability. While this adjustment accounts for the altered CE prevalence, it does not guarantee that the adjusted probabilities are well calibrated. Therefore, we applied beta calibration (Kull et al. 2017) as a post-processing step to further refine the adjusted probabilities

$$f(\xi, g)_{calibrated} = \frac{1}{1 + 1/(e^{c}\frac{f(\xi,g)^{a}}{(1-f(\xi,g))^{b}})}, \tag{3}$$

where a and b are the shape parameters, and c is the location parameter. Probabilities obtained from DL were directly beta-calibrated since CE prevalence was not altered during training. Beta calibration can outperform other calibration approaches because it can model non-linear transformations. In addition, if the input probabilities are already well calibrated, beta calibration learns a function approximating identity, thereby preserving well-calibrated predictions (Kull et al. 2017). Calibrated probabilities are evaluated visually with reliability diagrams

and quantified using the Expected Calibration Error (ECE). The ECE is the weighted average difference between predicted probabilities and observed frequencies across probability bins (Naeini et al. 2015, supporting information Text S5(tbc)).

To summarise, we trained RF and XGB using the resampled training data and DL using the complete training data. We then obtained the predicted probabilities from the validation data and adjusted probabilities of RF and XGB (Eq. 2) to account for the difference between the original CE prevalence and the prevalence in the resampled training data. Next, we calibrated the adjusted probabilities from RF and XGB and the raw probabilities from DL using beta calibration (Eq. 3), fitted individually for each ML model. The threshold probability for each ML model was determined by maximising HSS using the calibrated probabilities from the validation data. This procedure ensures independent evaluation using the ERA5 test data as it remains completely untouched during training, calibration, and threshold optimisation. For the ERA5 test and CMIP6 data, we applied the same procedure: 1) obtaining predicted probabilities from the trained ML models, 2) adjusting the probabilities for RF and XGB to account for the altered training CE prevalence (Eq. 2), and 3) calibrating the probabilities using the previously fitted beta calibration functions (Eq. 3).

L. 212-214: Please cite relevant literature for this approach.

We will add Hoens and Chawla (2013) as in our response to L. 210.

L. 214: I wonder why the complete train dataset was still used for DL despite its large size and imbalance.

Please see the third paragraph in the revised Sect. 2.5 in our response to L. 210.

L. 216: As mentioned above, this could have been precised earlier in the text. Looking at Fig. S1, since the performance of the ML models were not tested for percentiles below the 50th percentile, I am not convinced that the optimal percentile is 50. Could it be expanded to lower percentiles or could it be justified why the lower percentiles were not considered?

This information will appear earlier in the revised manuscript, including the justification for not including lower percentiles. See our response to L. 142–146.

L. 218-219: The same remark holds here: why sampling ratios < 15% were not included in the plot to illustrate the optimality at 15%?

This was an oversight on our side. In the revised ML framework, we tested additional sampling ratios of 1%, 3%, and 5% with anomalies as predictors. Based on the AUCPR, the optimal sampling ratio for XGB remained unchanged (i.e., 15%), while the optimal ratio for RF became 3%. Since we are now training the ML models using absolute predictor values, we repeated the analysis, resulting in optimal sampling ratios of 3% and 35% for RF and XGB, respectively (Fig. 3).

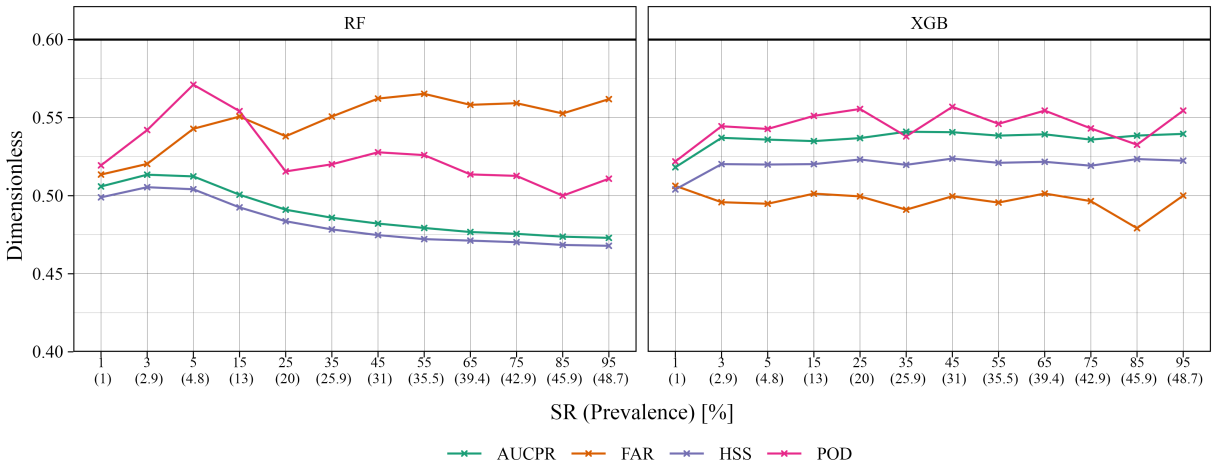


Figure 3: (new) Evaluation metrics of RF (left panel) and XGB (right panel) across different CE sampling ratios (prevalence) in the training dataset. Evaluation metrics are AUCPR (teal), FAR (orange), HSS (purple), and POD (pink), shown as solid lines with cross markers.

Eq. 2, 3: Both the adjusted probability estimates and the calibration appear to use the CE frequency of the climatological dataset. Since the data are split using stratified sampling that preserves the climatological CE distribution, the prevalence of CEs in the test set is therefore known and consistent with the frequency used in the probability adjustment/calibration. This may result in performance that is more favorable than would be expected when the model is applied to an independent dataset with a different or unknown CE frequency, such as a future-climate dataset. Please discuss this issue and clarify how the proposed probability adjustment and calibration are expected to perform when the underlying CE frequency changes.

We interpret the reviewer's reference to changes in "CE frequency" as changes in the overall **CE prevalence** in the dataset, since CE frequency is generally time-relative and is ambiguous in this context. For example, an increase in CE frequency over the southern AP and a decrease over the northern AP may compensate for each other, resulting in little or no change in the overall CE prevalence. We respectfully disagree with several points raised by the reviewer.

First, stratified sampling is appropriate to prevent random over- or under-representation of CEs in the training, validation, and test data (Dankers et al. 2019; Xu et al. 2026). This ensures that the original CE prevalence is preserved across the three data parts before training (resampling) and that the evaluation metrics are estimated using the real-world CE prevalence (this is particularly important in operational settings).

Second, probability adjustment is necessary when ML models are trained on resampled training data that do not reflect the original CE prevalence. Resampling the training data changes the prior probabilities of CEs and consequently introduces bias in the posterior probabilities of CEs produced by the ML model (Saerens et al. 2002; Pozzolo et al. 2015). The adjustment is a monotonic transformation of the predicted probabilities (Eq. 2) and therefore does not change their ranking (Pozzolo et al. 2015).

Third, calibration is applied to improve the reliability of the (adjusted) probabilities (Kull et al. 2017; Ojeda et al. 2023), since ML models can exhibit bias in their predicted probabilities; for example, boosted

trees tend to push probabilities away from 0 and 1 (e.g., Niculescu-Mizil and Caruana 2005). In addition, the applied calibration is a monotonic transformation (Eq. 3) and therefore does not change their ranking (Kull et al. 2017). The goal of calibration is to make the ML model probabilities reflect the true event probabilities, so that a calibrated probability of 90% indicates that CE occurred in 90% of similar cases.

Moreover, the reported evaluation metrics and their confidence intervals were estimated from 300 (revised: 1000) non-stratified bootstrap resamples of the test data, accounting for sampling variability in the historical evaluation metrics, including slight variations in CE prevalence.

To conclude, probability adjustment and calibration do not materially improve the discrimination skill of the ML models. Table 2 shows that POD, FAR, HSS, and AUCPR remain nearly unchanged across raw, adjusted, and calibrated XGB probabilities when using their respective threshold probabilities, whereas the most pronounced change is observed in BSS. These results indicate that probability adjustment and calibration primarily benefit probabilistic skill (BSS) rather than the ability of the models to distinguish CEs from non-CEs (AUCPR) or their threshold-based classification skill (POD, FAR, and HSS). Similar results are obtained for RF and DL.

Table 2: Threshold probabilities and evaluation metrics for XGB using raw, adjusted, and calibrated probabilities. The threshold probabilities were determined from the validation data. Evaluation metrics are reported as medians and 95% confidence intervals from 1000 bootstrap resamples of the test data.

	Raw	Adjusted	Calibrated
Threshold probability	0.68	0.04	0.28
POD	0.588 (0.580–0.598)	0.586 (0.578–0.595)	0.586 (0.576–0.594)
FAR	0.478 (0.470–0.486)	0.476 (0.467–0.484)	0.475 (0.467–0.484)
HSS	0.550 (0.543–0.558)	0.551 (0.543–0.558)	0.550 (0.542–0.558)
BSS	0.147 (0.129–0.163)	0.315 (0.306–0.323)	0.393 (0.385–0.401)
AUCPR	0.588 (0.579–0.597)	0.588 (0.579–0.597)	0.588 (0.579–0.597)

Our study follows the *Perfect Prognosis* approach (Maraun and Widmann 2018), which assumes that the statistical relationship between atmospheric predictors and CE occurrence remains applicable under a changing climate. Under this assumption, changes in atmospheric conditions can alter the overall CE prevalence without invalidating the historical relationship between predictors and CE probabilities. Retaining the historical probability adjustment, calibration mapping, and classification threshold does not constrain the predicted future CE prevalence to its historical value.

Nevertheless, we acknowledge the uncertainty of applying the trained and calibrated ML models to future climates. For example, the threshold probability may change even without a change in CE prevalence to achieve optimal classification skill in the future. In addition, future CE prevalence may change due to changes in atmospheric conditions without a corresponding change in the threshold probability. These changes, combined with the unavailability of true future CE prevalence and labels, make it difficult to quantify the reliability of the predicted probabilities and the classification skill under future climate con-

ditions. However, if the true future CE prevalence and labels become available, the CE probabilities could be reassessed and, where appropriate, corrected following approaches such as Horsch et al. (2008). We will acknowledge this uncertainty in the revised manuscript.

L. 237: Please add the missing parentheses.

They will be added in the revised manuscript as in our response to L. 210.

L. 241-242: Please specify how the probabilities from CMIP6 were adjusted and calibrated.

These information will be added to the manuscript as in our response to L. 210

L. 262: Please include the reference to Fig. S5 at the end of this sentence (instead of the next one) as this sentence appears to already be commenting on this figure.

Fig. S5 will be added to this sentence and removed from the next one.

Fig. S4: Please clarify why the AUCPR baseline is missing on this plot.

The baseline AUCPR values of 0.0055, 0.0035, 0.0079, and 0.0089 for 03:00 UTC+3, 09:00 UTC+3, 15:00 UTC+3, and 21:00 UTC+3, respectively, were mistakenly omitted and will appear in the revised plot.

Fig. S5: Please consider adding the frequency map of CEs for comparison/discussion.

We agree that a CE frequency map can help interpret the spatial evaluation metrics. We will add the observed CE frequency at each grid point from the test data as a reference panel for comparison.

L 267-268: If I understand well, the “climatological” data mentioned in L. 229 is only the training and the validation data set? It is not so clear in my opinion until this passage (L. 267).

We are not entirely sure that we have correctly understood the reviewer’s comment. Nevertheless, we hope that our response to L. 210 also addresses this concern.

L 271 and Fig. 3: The term “Multivariate Model” is not sufficiently clear, as it does not indicate that the model is based on CMIP6 data. Given that MBCn is applied to the CMIP6 convection indices, I suggest explicitly including “CMIP6” in the label to make the distinction clearer throughout the results.

We believe there is a misunderstanding here. The term used in the manuscript is "multi-model" at L. 271, not "multivariate model". "Multi-model" refers to the use of multiple CMIP6 models, whereas "multivariate" refers to multiple variables and is relevant only to MBCn as a multivariate bias-correction method in L. 157. Therefore, "multivariate model" would not be an appropriate description in this context.

Nevertheless, we will revise the labels to "CMIP6 multi-model" to make the reference to the CMIP6 models explicit.

L. 273: Please define FutI and FutIII at their first occurrence in the manuscript.

We will define *FutI* and *FutIII* as in our response to L. 121-124.

L 273-274: I do not see an obvious increase in occurrence for probabilities below 10-2. Please reformulate this sentence.

Based on the revised analysis, we will revise the sentence as:

> At *FutI*, the CMIP6 multi-model density ratio relative to the current climate indicates an increase in the density over the CE-probability ranges of 0.017–0.86, 0.005–0.74, 0.21–0.77, and 0.003–0.48 for DJF, MAM, JJA, and SON, respectively. In addition, DJF and JJA exhibit additional increases over lower CE-probability ranges of 0.0010–0.0012 and 0.0018–0.014, respectively. Outside these specified ranges the CMIP6 multi-model density ratio decreases, with the largest decrease in SON (32.7%) and the smallest decrease in winter (4.55%). Noteworthy, the projected changes at *FutI* vary among individual CMIP6 models, with some models showing increases and others showing decreases in CE-probability density.

Please note that the projected changes in FutI are not consistent across all ensemble members, with some members showing increases and others showing decreases.

We will add this information to the revised manuscript, as in our response to L 273-274.

L 280-281: Please clarify what "it" refers to, as it appears not to refer to the diurnal cycle. Also, is the underestimation exactly 0.001, or is this an approximate value? If it is approximate, please indicate this accordingly.

This sentence will be removed in the revised manuscript.

L 283 : The term "earlier" (or "afterwards" in L. 275) is unclear, particularly in the context of diurnal-cycle plots, where it could be interpreted as referring to an earlier time of day. Please specify what is meant by "earlier" (e.g., at a lower value of the relevant variable or threshold).

We will revise this part as follows:

> At *FutI*, the CMIP6 multi-model density ratio indicates an increase in the density of CE probabilities relative to the current climate over the CE-probability ranges of 0.003–0.52, 0.003–0.44, 0.02–0.78, and 0.004–0.85 at 03:00, 09:00, 15:00, and 21:00 UTC+3, respectively. In addition, 15:00 UTC+3 exhibits a separate increase over the lower CE-probability range of 0.004–0.007. Outside these specified ranges, the CMIP6 multi-model density ratio decreases, with the largest decrease at 09:00 UTC+3 (28.3%) and the smallest decrease at 21:00 UTC+3 (5.02%). Regarding *FutIII*, the CMIP6 multi-model density ratio indicates an increase in the density of CE probabilities relative to the current climate over the CE-probability ranges of 0.002–0.061, 0.002–0.033, 0.002–0.40, and 0.003–0.55 at 03:00, 09:00, 15:00, and 21:00 UTC+3, respectively. Outside these specified ranges, the CMIP6 multi-model density ratio decreases, with the largest decrease at 09:00 UTC+3 (48.1%) and the smallest decrease at 21:00 UTC+3 (28.1%).

The term "afterwards" in L 275 is removed, as in our response to L 273-274.

L. 302-304: Could the authors briefly explain why the fact that the ERA5 test data represent only 20% of the total period (2001–2024) is expected to result in an underestimation of CE frequency?

The seasonal CE frequency is estimated by summing CEs for each season and year and then averaging across years; thus, with 80% of the samples excluded, a lower seasonal CE frequency is expected. To verify this, we compared the seasonal CE frequency (Fig. 4) calculated from the complete IMERG record (2001–2024), the original IMERG labels in the ERA5 test data (ERA5_ORG), and the XGB predictions for the ERA5 test data (ERA5_XGB). The differences between ERA5_ORG and IMERG are large while the differences between ERA5_XGB and ERA5_ORG are minimal, indicating that the underestimation primarily arises from the test data containing only 20% of the original samples rather than from XGB missing CEs.

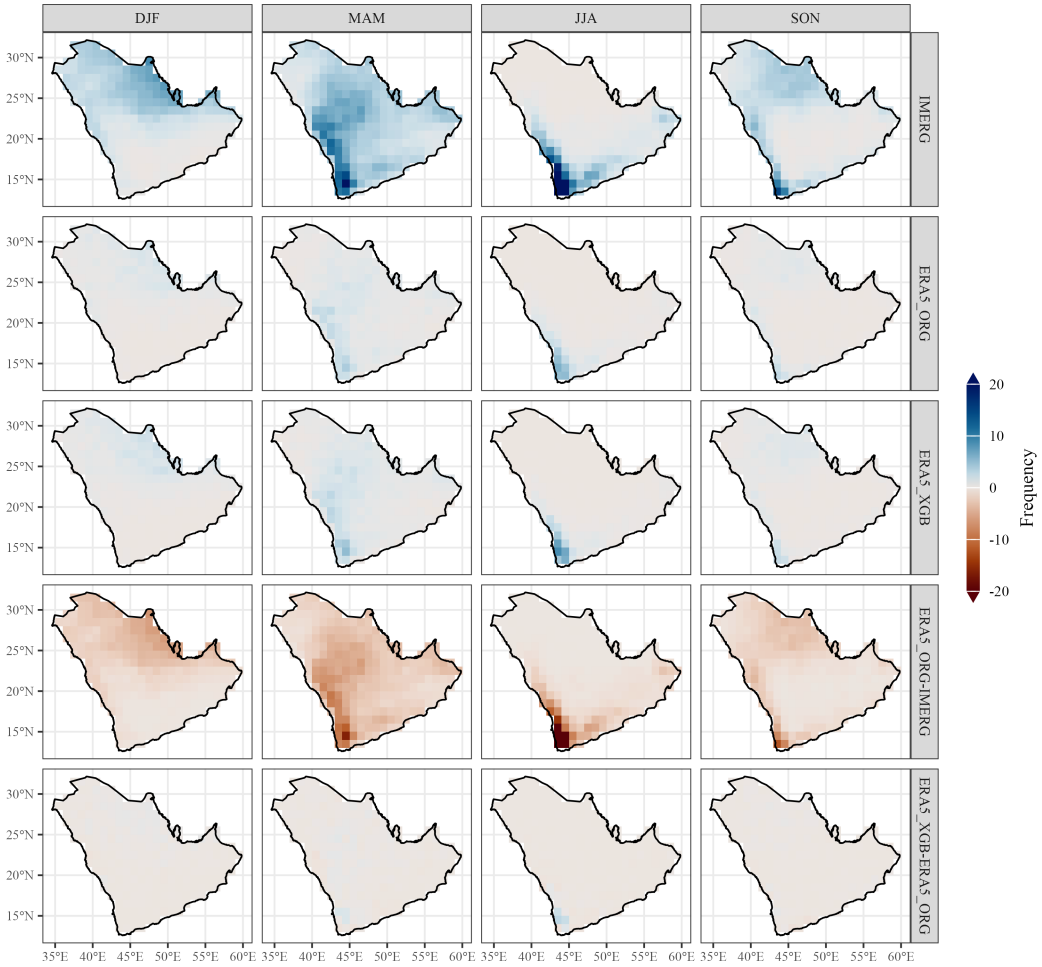


Figure 4: Seasonal CE frequency calculated from the complete IMERG record (IMERG; first row), the true IMERG labels in the ERA5 test data (ERA5_ORG; second row), and the CEs identified by XGB using the ERA5 test data (ERA5_XGB; third row). The fourth and fifth rows show the differences ERA5_ORG − IMERG and ERA5_XGB − ERA5_ORG, respectively

Sec 3.3: The section on biases in the CMIP6 convective indices (Section 3.3) feels somewhat disconnected

from the subsequent results. I suggest considering an exchange of Sections 3.2 and 3.3, so that the discussion of projected changes in CE frequency (new Section 3.3) is followed directly by the projected changes in convective indices conditioned on CE (Section 3.4). This ordering would provide a more natural flow between the occurrence of CEs and their associated convective environments.

We agree with this suggestion. In the revised manuscript, we will reorder these sections accordingly, as shown in our response to the **Major Comment**.

L. 327-328: Please clarify how the CMIP6 models with grid spacing coarser than 1° were regridded to the 1° grid.

The same bilinear interpolation described in our response to L. 153 was applied to all CMIP6 models, including those with native grid spacing coarser than 1°.

L331 332: "Dry bias" is unclear here. Please specify what the PW and 2-m dew-point temperature biases are relative to (e.g., ERA5).

We will replace the term "dry bias" with "negative bias" in the revised manuscript. Regarding the reference dataset, we respectfully disagree. The immediately preceding sentence, which opens this section, already establishes that CE-associated atmospheric profiles from CMIP6 are evaluated relative to ERA5. Repeating this information at every instance would be redundant. The Section's revised opening sentences will read as:

> Here, we assess the conditional fidelity of CMIP6 in reproducing CE components at 1°×1°in the current climate by comparing atmospheric profiles from ERA5 that are labelled as CEs based on precipitation from IMERG against atmospheric profiles from CMIP6 that are classified as CEs by XGB.

L. 350: Fig. 8a appears to show the mean bias in vertical profile not the profiles themselves. Please correct.

The "mean bias" can be misleading here as it may suggest biases were estimated for paired profiles from CMIP6 models and ERA5 and then averaged, which is not the case. We aggregated the profiles seasonally for each model and subtracted the corresponding seasonal mean profile of ERA5. Thus, for example, "Bias in the mean RH profile" is appropriate. However, we acknowledge that the wording at L. 350 incorrectly refers to the mean profiles rather than their biases and will revise it accordingly.

L 381-382: The reported increase in w500 is not readily apparent from Fig 10c.

The changes in ω_{500} are relatively small compared with the spread of the distributions shown in Fig. 10c, making them difficult to distinguish visually. However, Table S7 (referenced in the same sentence) quantifies the changes. We will apply a transformation to the y-axis in Fig. 10c to make the changes more apparent.

L. 388-389: If the diurnal-cycle plots are included, it would be useful to discuss the relevant features of the diurnal cycle in the text. Otherwise, the purpose of presenting these plots is unclear.

We will expand the discussion of the diurnal cycle in the revised manuscript.

L. 417: "the the" → Please correct.

It will be corrected in the revised manuscript.

L 417 – 420: Could the authors briefly explain why would the ML model performs the best when "these convective situations are influenced by large-scale environments"?

Paragraph at L. 417–428 will be revised as follows:

> Large-scale atmospheric systems influencing convection over the AP, particularly the northern and central regions, include mid-latitude frontal systems (Almazroui et al. 2015; Almazroui et al. 2016; Luong et al. 2024; Charabi and Al-Hatrushi 2010), the Red Sea Convergence Zone, the Red Sea Trough (Dasari et al. 2018; de Vries et al. 2013; Khan et al. 2018), and the subtropical jet (Taraphdar et al. 2025), whereas localised convection, primarily over the southern AP, is influenced by orographic forcing (Abdullah and Al-Mazroui 1998; Khan et al. 2018; Cook et al. 2021; Branch et al. 2020; Farquharson et al. 1996), interaction between sea-breeze and surface heating or cold pools (Fonseca et al. 2025; Nelli et al. 2022; Eager et al. 2008; Gopalakrishnan et al. 2023), and the remnants of low-pressure systems from the ISM (Charabi 2009; Babu et al. 2016).
>
> ML model evaluation metrics are more favourable in the northern AP and during cold months or cold periods of the day likely because convection is influenced by large-scale systems, whereas evaluation metrics are less favourable where convection is localised, such as in the southern AP, over complex terrain or during summer and afternoon hours. This contrast likely reflects the coherent environmental signatures of large-scale systems, which can be captured by predictors at 1°and 6 h, while the smaller spatiotemporal scales of localised convection and its associated processes make them more difficult to represent in these predictors, thereby limiting the ability of ML models to identify the associated CEs. Similar findings are reported by Gómez-Navarro et al. (2022) over complex terrain in Switzerland. Furthermore, an instantaneous atmospheric profile is labelled convective based on accumulated precipitation; however, this profile may not accurately represent the atmospheric state when localised convection occurs far from the profile timestamp.

L 432-435: I do not fully understand the relevance of these references in this context. The authors identify CEs based on environmental variables rather than convective precipitation from the models. Therefore, the finding that CMIP6 models require more moisture to transition from drizzle to intense convection does not necessarily imply that these models underestimate moisture in convective environments. Please clarify how the cited studies support the interpretation made here.

We thank the reviewer for raising this point. We acknowledge that the current wording does not clearly communicate our hypothesis. First, we would like to note that convection parameterisations in climate models modify atmospheric temperature and moisture profiles through convective heating, vertical transport, and entrainment and detrainment processes (e.g., Prein et al. 2015). Therefore, although we are

not using CMIP6 precipitation to identify CEs, the predictors from CMIP6 are influenced by the model's representation of convection.

Our original hypothesis for the negative PW and T_{d2m} biases builds on the documented tendency of convection parameterisation in climate models to produce precipitation that is too frequent, less intense, and longer-lasting (i.e., drizzle bias Martinez-Villalobos et al. 2022; Chen et al. 2021). The drizzle bias has been attributed to premature and overly frequent moist convection occurring in relatively dry or weakly unstable environments, whereas in reality atmospheric instability often accumulates before intense convection develops (e.g., Chen and Dai 2019; Dai 2006; Chen et al. 2020). Furthermore, Hagos et al. (2021) showed that most of the examined CMIP6 models operate more frequently in the drizzle regime and require higher moisture accumulation than ERA5 for the transition to intense precipitation. The drizzle bias has also been identified over the AP (Pathak et al. 2025). We hypothesise that overly frequent weak moist convection and precipitation may remove atmospheric moisture and reduce instability before sufficient accumulation occurs to support intense convection. This could contribute to the negative moisture and instability biases identified in our analysis.

Nevertheless, we revised Fig. 7 and examined whether the environmental biases are limited to ML-identified CEs or are generally apparent in CMIP6 models over the AP. We analysed the biases relative to ERA5 for all atmospheric profiles (Fig. 5) and for CEs only (Fig. 6). For all atmospheric profiles, the results show that PW is predominantly negatively biased over the region, while $T_{[}$d2m] exhibits negative biases across all CMIP6 models. Most models prominently underestimate MUCAPE at higher percentiles. These results indicate that the moisture and instability biases are not exclusively associated with CE classification. We will revise the discussion to include this diagnostic analysis and briefly clarify the proposed hypothesis.

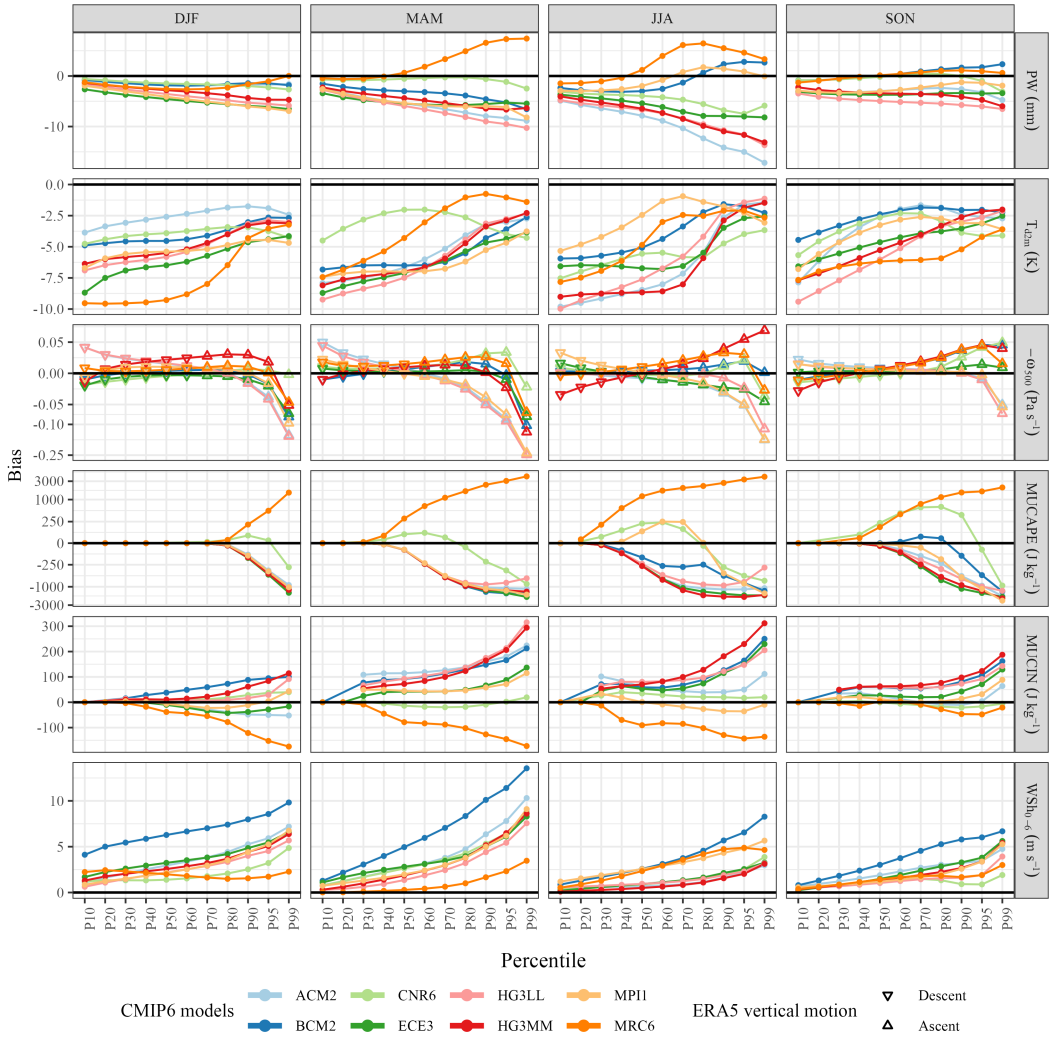


Figure 5: Seasonal biases at different percentiles for precipitable water (PW; first row), 2 m dew-point temperature (Td_{2m}; second row), negative vertical velocity at 500 hPa ($-\omega_{500}$; third row), most unstable CAPE (MUCAPE; fourth row), most unstable CIN (MUCIN; fifth row), and 0–6 km wind shear ($WSh_{0\text{-}6}$; sixth row) in **all atmospheric profiles** from CMIP6 models at the current regional warming level relative to ERA5. The axes for MUCAPE and ω_{500} are pseudo-log transformed with sigma values of 75 and 0.025, respectively. Triangles in the ω_{500} panels indicate whether the corresponding ERA5 percentile represents descent or ascent. CMIP6 models are shown as solid lines with point markers: ACM2 (light blue), BCM2 (blue), CNR6 (light green), ECE3 (green), HG3LL (light pink), HG3MM (red), MPI1 (light orange), and MRC6 (orange). The horizontal black line indicates zero bias.

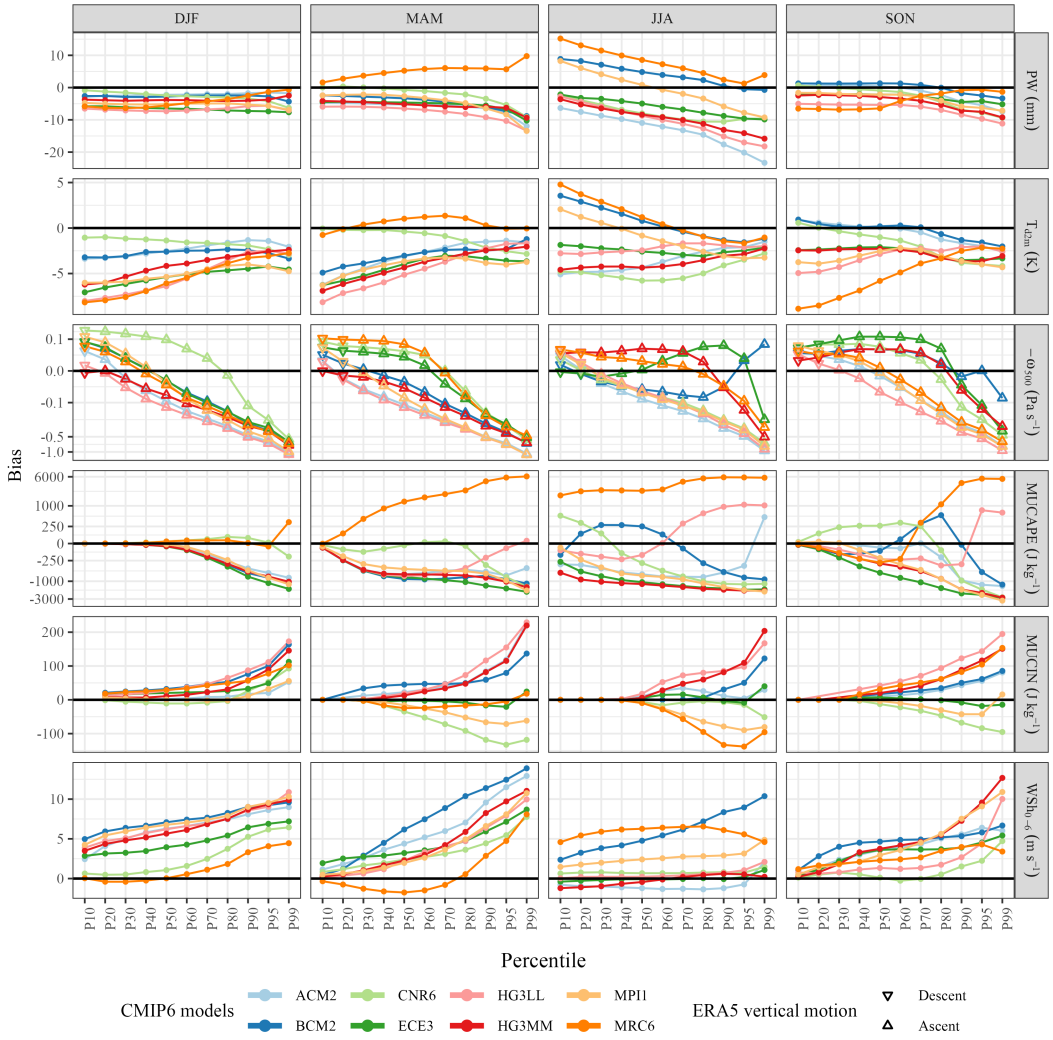


Figure 6: Seasonal biases at different percentiles for precipitable water (PW; first row), 2 m dew-point temperature (Td_{2m}; second row), negative vertical velocity at 500 hPa ($-\omega_{500}$; third row), most unstable CAPE (MUCAPE; fourth row), most unstable CIN (MUCIN; fifth row), and 0–6 km wind shear ($WSh_{0\text{-}6}$; sixth row) **in CEs** from CMIP6 models at the current regional warming level relative to ERA5. The axes for MUCAPE and ω_{500} are pseudo-log transformed with sigma values of 100 and 0.025, respectively. Triangles in the ω_{500} panels indicate whether the corresponding ERA5 percentile represents descent or ascent. CMIP6 models are shown as solid lines with point markers: ACM2 (light blue), BCM2 (blue), CNR6 (light green), ECE3 (green), HG3LL (light pink), HG3MM (red), MPI1 (light orange), and MRC6 (orange). The horizontal black line indicates zero bias.

L 435-438: A similar comment applies here. The fact that CMIP6 models struggle to represent elevated convection does not necessarily imply that this is due to biases in the environmental variables. These are two distinct issues, and evidence would be needed to establish that the misrepresentation of elevated convection results specifically from biases in the environmental conditions. Please clarify the basis for this interpretation.

The text in L. 435–438 reads as follows:

> Additionally, convection in cold months over the AP exhibits a secondary nocturnal peak (Homoudi et al., 2026). Most CMIP6 models struggle to represent nocturnal elevated convection above the boundary layer (Tang et al., 2021), providing an additional plausible explanation for the underestimation of CEs' occurrence in CMIP6.

In this text, we have not claimed that biases in environmental variables cause CMIP6 models to struggle with representing elevated convection. But we suggest that the difficulty in representing nocturnal elevated convection may contribute to the underestimation of CEs' occurrence in CMIP6. We acknowledge that this statement is misplaced and should appear where we discuss the underestimation of CE occurrence. We will revise the statement as follows:

> Convection in cold months over the AP exhibits a secondary nocturnal peak associated with MCSs (Homoudi et al. 2026). Most CMIP6 models struggle to represent nocturnal elevated convection above the boundary layer (Tang et al. 2021). Song et al. (2024) explained that convection parameterisations in GCMs rely on local large-scale atmospheric conditions to trigger convection at individual grid cells, whereas observed MCSs are long-lived and exhibit propagating features, meaning that convection at a particular location may have been initiated upstream. Consequently, such models may fail to represent convection associated with propagating MCSs. Deficiencies in representing nocturnal elevated convection may hinder the development and propagation of associated MCSs and the establishment of their mesoscale circulations (Houze Jr. 2004), potentially contributing to the underestimation of subsequent CE occurrence.

L. 457: The statement "The consistent increase in trends during DJF" is unclear. Please specify what is increasing and whether this increase is consistent across the models. Moreover, since this study does not analyse temporal trends, the use of the term "trends" seems inappropriate here. Please clarify or revise this statement.

We will revise the text to read as:

> The projected increase in CE frequency during DJF under *FutI* and *FutIII* could be due to the southward shift and intensification of the subtropical jet, the rise in transient activity, and moisture influx from adjacent water bodies that are expected under a warming climate (Pathak et al. 2025).

Regarding the consistency across the models, we only report and discuss patterns and changes supported by model agreement of at least 75% (revised from 70% following the exclusion of two CMIP6 models);

otherwise, we explicitly state when this criteria is not met. Therefore, adding this information specifically here would provide no additional benefit, while maintaining consistency would require repeating it throughout the manuscript, resulting in redundancy.

L. 461: "Both futures" is somewhat vague. Please reformulate this statement to explicitly specify the two future scenarios being referred to.

In the revised manuscript, we will explicitly state the two future warming levels, i.e., *FutI* and *FutIII*.

L. 465: Hypothesis d. Please see my previous comment (L. 435-438)

We are discussing the possible reasons for the underestimation of CE occurrence here, not the biases in environmental variables, as explained in our response to L 435-438.

L. 475-476: Please check the stated relationship between low-level moisture and LCL. At constant RH (as stated in the next sentence), an increase in specific humidity requires an increase in temperature, and it is not clear that this would result in a lower LCL. Please clarify the physical basis for this relationship.

We acknowledge that Figs. 11 and S10 incorrectly showed relative changes instead of absolute changes. We have revised both figures to show absolute changes using results from the revised ML framework. The revised seasonal results show that RH changes at *FutI* and *FutIII*, with increases, decreases, and model agreement depending on the season and atmospheric level (Fig. 7b). However, the revised diurnal results show that RH increases monotonically with warming between 1000 and 700 hPa (Fig. 8b). The contrasting seasonal RH changes warrant further in-depth analysis.

The LCL pressure of the most unstable parcel generally increases under warming in both seasonal (Fig. 9a) and diurnal analyses (Fig. 9b), indicating a likely decrease in LCL altitude. This is consistent with the projected decrease in MUCIN and the increase in T_{d2m} reported in the manuscript.

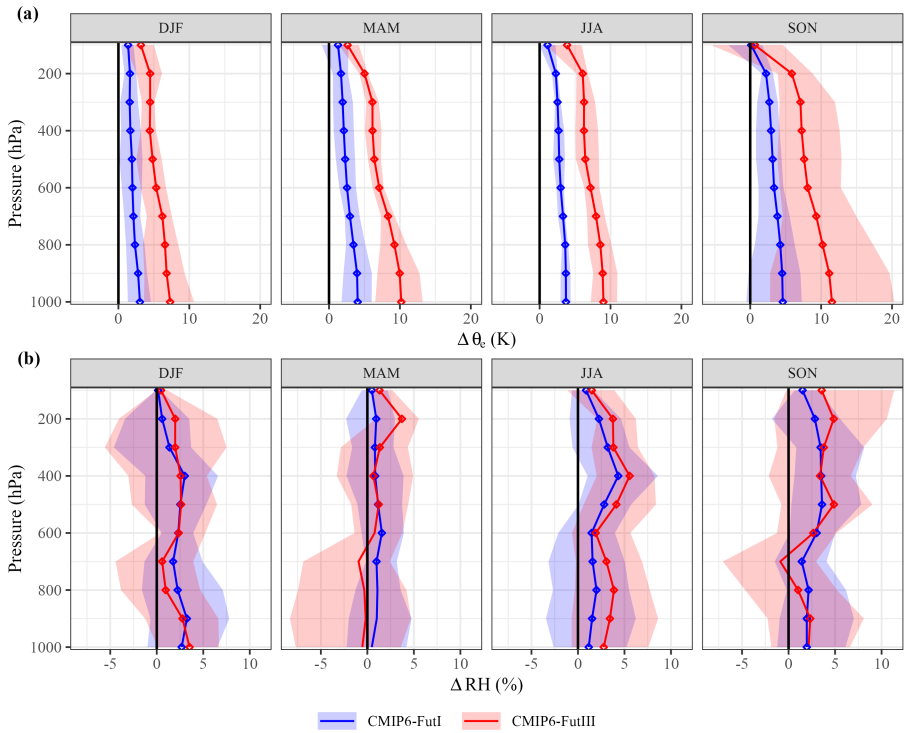


Figure 7: Seasonal changes in the CMIP6 multi-model mean (a) θ_e and (b) RH profiles for CEs at *FutI* (blue colour) and *FutIII* (red colour) relative to current regional warming level. Shaded areas indicate the 2.5th to 97.5th percentile range of the CMIP6 ensemble, and diamond shapes indicate at least 75% model agreement on the sign of change.

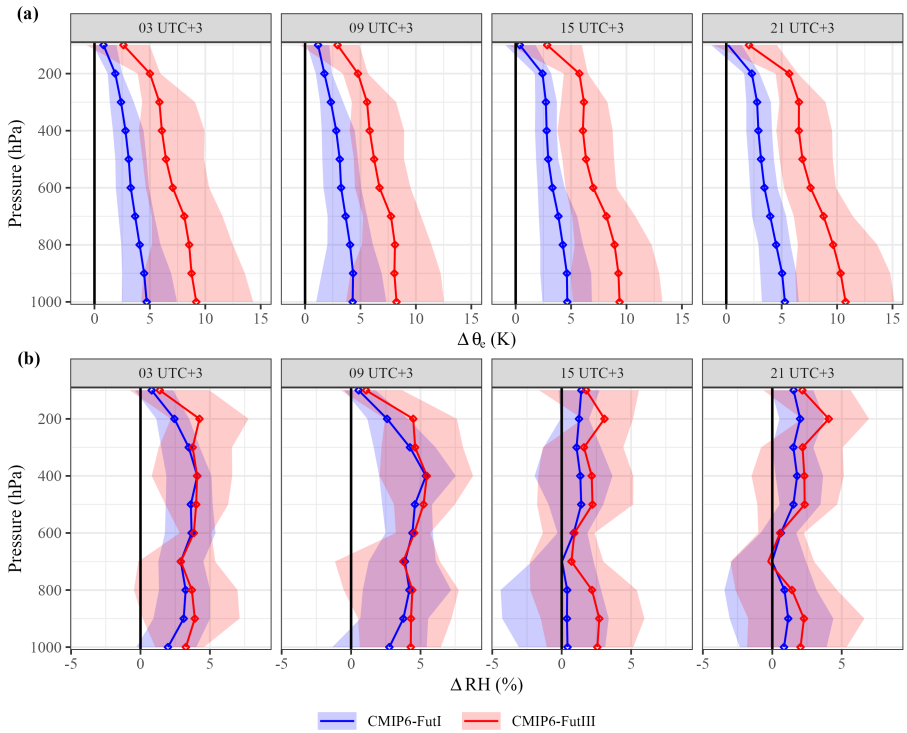


Figure 8: Diurnal changes in the CMIP6 multi-model mean (a) θ_e and (b) RH profiles for CEs at *FutI* (blue colour) and *FutIII* (red colour) relative to current regional warming level. Shaded areas indicate the 2.5th to 97.5th percentile range of the CMIP6 ensemble, and diamond shapes indicate at least 75% model agreement on the sign of change.

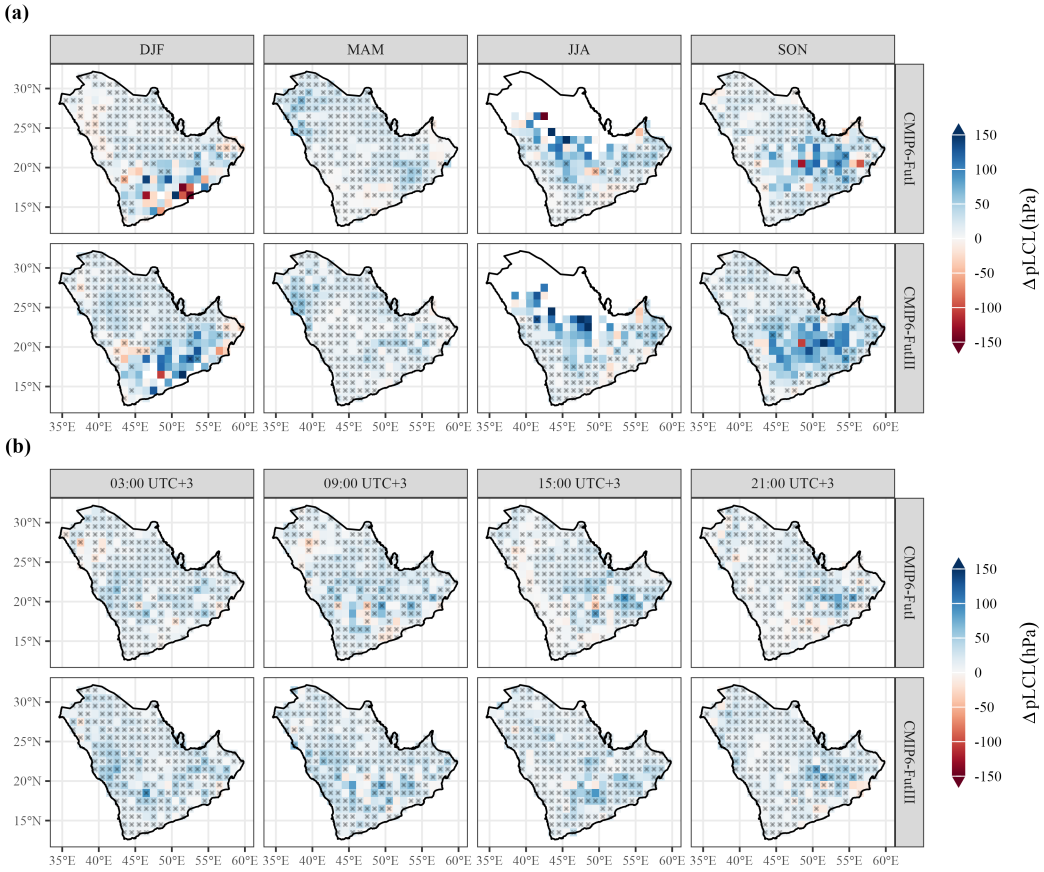


Figure 9: Seasonal (a) and diurnal (b) changes in the CMIP6 multi-model mean LCL pressure of the most unstable parcel (hPa) for CEs at *FutI* and *FutIII* relative to current regional warming level. Stippling indicates at least 75% models' agreement on the sign of change.

L 471-481: The link between CE intensification and the subsequent statements about a decrease in the number of wet days and an increased contribution of extreme precipitation to the total precipitation is not clear. CE intensification does not directly imply changes in either wet-day frequency or the relative contribution of extreme precipitation. Please provide evidence supporting these relationships or clarify the basis for these conclusions.

We agree the current wording is ambiguous. We will revise this passage considering the information in our response to L. 10-12.

L 481-483: I find the argument difficult to follow, as this statement and the one in L. 475 appear to be contradictory. The authors first state that tropospheric RH generally does not change with warming, but subsequently argue that slower ocean warming limits moisture supply needed to maintain continental RH, resulting in a drying influence. If these statements refer to different atmospheric layers or spatial domains (e.g., tropospheric versus near-surface/continental RH), this distinction should be clearly explained. Otherwise, please clarify how these two arguments are consistent.

We will remove this reasoning and the associated reference from the revised manuscript, since they are no longer relevant to the updated results and discussion.

L. 500: Please add the missing parentheses.

They will be added in the revised manuscript.

I would suggest adding as a limitation the method used to identify convective events. In particular, precipitation exceeding the median over a 6-hour period at a $1° \times 1°$ grid cell does not necessarily imply that the precipitation is convective. Future versions of the approach could consider incorporating a more direct indicator of convection, such as lightning observations.

We thank the reviewer for this suggestion. While precipitation exceeding the median does not necessarily imply convection, given the predominantly convective nature of precipitation over the AP (e.g., Luong et al. 2024) and our focus on precipitation exceeding the median of wet periods, we question the extent to which this limits our approach.

To examine this concern, we estimated the maximum precipitation intensity within each $1°$ grid cell and 6 h window. We used a published Z–R relationship for Jeddah, Saudi Arabia (Albar et al. 2015) to convert reflectivity thresholds of 25 and 30 dBZ into equivalent rainfall intensities. We found that 99.6% and 90.8% of the identified CEs, respectively, included precipitation exceeding these intensity thresholds. These reflectivity thresholds are intended to account for weaker convective systems, consistent with Queiroz and Biscaro (2026), who adopted 25 dBZ to identify and track convection in northeastern Brazil, including its semi-arid regions. We acknowledge that the Z–R relationship was developed for Jeddah and may not be spatially and temporally representative of precipitation across the AP. Nevertheless, the substantial overlap supports the use of our precipitation-based criteria to label CEs.

As already indicated in our response to L. 142–146, lightning observations or radar reflectivity could provide more direct indicators of convection. We will reiterate their potential use in future implementations when sufficiently long, freely accessible, and spatially consistent observational records become available.

L 505: The ending “is recommended” is awkward and appears grammatically inconsistent with the beginning of the sentence (“Future research should focus on. . . ”). Please revise the sentence for clarity.

The phrase "is recommended" will be removed in the revised manuscript.

L. 515: The statement that the study “addresses a critical gap in the literature by focusing on convectively favourable conditions rather than the precipitation itself” seems to overstate the novelty of the study. As written, it suggests that previous studies have primarily focused on precipitation and have not investigated environments favourable for convection, which is not the case (e.g., Lepore et al., 2021). Please clarify what specifically is novel about the present approach and revise this statement accordingly.

We will revise this passage and restate our motivation from the Introduction as follows:

> In doing so, the study addresses a critical regional and methodological gap by applying an ML framework to identify convectively favourable conditions over the AP, rather than relying on precipitation simulated by CMIP6 models (e.g., Nelli et al. 2025; Pathak et al. 2025), which is largely governed by convection parameterisation schemes and therefore of limited reliability

over the region (e.g., Pathak et al. 2025; Zhang et al. 2025). Lepore et al. (2021) investigated CEs globally using predefined severe-weather proxies and thresholds under a global warming framework, whereas our study focuses on the AP using ML models trained on predictors particularly relevant for the variable spatiotemporal nature of convection over the AP, avoiding reliance on heuristic thresholds for individual predictors or an all-encompassing predictor e.g., Doswell III and Schultz 2006. Additionally, our study adopts a regional warming approach, which is more suitable for regional adaptation planning (Corre et al. 2025), particularly as the AP is warming more rapidly than the global average (Cook et al. 2021).

In addition, The accompanying CEs dataset (Homoudi 2026) lays the foundation for a statistical-dynamical downscaling framework (e.g., Beaulant et al. 2011; Meredith et al. 2018; Wang et al. 2026), wherein identified CEs can be dynamically downscaled at convection-permitting resolution to produce physically consistent regional projection ensembles over the AP while reducing the computational cost. This approach provides an efficient pathway for further scrutiny of the impact of climate change on convective precipitation systems over the AP.

L 518: The statement "it is still possible that a CE does not rain" is unclear, please clarify.

We will revise the sentence to:

Although the ML models are trained to identify atmospheric environments associated with precipitation exceeding the local median, their predictions can include false alarms, as favourable atmospheric conditions do not guarantee that rainfall will occur or that its amount will exceed the local median.

Figures:

Fig. 2c: "Basline" should be corrected to "Baseline." In addition, the caption does not describe what each curve/color represents. Please provide a clear description of the different curves (here and in the other figures where necessary).

We thank the reviewer for noting this. We will correct it. Furthermore, we will revise the captions to clearly describe the meaning of each curve and colour. For example, the caption of Fig. 2 will be revised as follows:

(a) Reliability diagrams for DL, RF, and XGB evaluated on ERA5 test data, showing the raw, analytically adjusted, and calibrated probabilities (black solid lines with point, star, and cross markers, respectively). The red dashed diagonal line indicates perfect reliability. (b) Bin-wise resolution (dashed line) and reliability (solid line), (c) the precision-recall curves (solid lines) with AUCPR baseline of 0.0066 (dashed red line), and (d) performance evaluation metrics, for DL (dark purple), RF (teal), and XGB (orange). The shaded areas and error bars denote the symmetric 95% confidence intervals obtained from 1000 bootstrap resamples.

Fig. 9a: Both the x- and y-axes are labelled MUCAPE, whereas MUCIN is referred to in the text.

The y-axis represents MUCIN and the x-axis represents MUCAPE. The typo will be corrected in the

revised manuscript.

Supplementary Document

L. 45: Please remove the extra space.

It will be removed in the revised Supplementary Document.

Figure captions: Please provide sufficient information in the captions to make the figures understandable independently. In particular, describe the meaning of the different lines/colors and add y-axis labels where necessary.

We will revise the figure captions in the revised Supplementary Document.

References

Abdullah, Ma and Ma Al-Mazroui (1998). "Climatological Study of the Southwestern Region of Saudi Arabia. I. Rainfall Analysis". In: *Climate Research* 9, pp. 213–223. ISSN: 0936-577X, 1616-1572. DOI: 10.3354/cr009213.

Albar, Ayman Mohammed Hashem, Abdulrahman Khalaf AL-Khalaf, and Heshmat Abdel-Basset Mohamed (2015). "Radar Rainfall Estimation of a Severe Thunderstorm over Jeddah". In: *Atmospheric and Climate Sciences* 5.3, pp. 302–316. DOI: 10.4236/acs.2015.53023.

Alkhawaga, Abdalmonem and Mohamed M. Mohamed (2026). "Precipitation Analysis under Climate Change Impacts over the United Arab Emirates". In: *Water Cycle*. ISSN: 2666-4453. DOI: 10.1016/j.watcyc.2026.06.002.

Almazroui, Mansour and Sajjad Saeed (2020). "Contribution of Extreme Daily Precipitation to Total Rainfall over the Arabian Peninsula". In: *Atmospheric Research* 231, p. 104672. ISSN: 0169-8095. DOI: 10.1016/j.atmosres.2019.104672.

Almazroui, Mansour et al. (2015). "Atmospheric Circulation Patterns in the Arab Region and Its Relationships with Saudi Arabian Surface Climate: A Preliminary Assessment". In: *Atmospheric Research* 161–162, pp. 36–51. ISSN: 01698095. DOI: 10.1016/j.atmosres.2015.03.014.

Almazroui, Mansour et al. (2016). "Climatology of the 500-hPa Mediterranean Storms Associated with Saudi Arabia Wet Season Precipitation". In: *Climate Dynamics* 47.9, pp. 3029–3042. ISSN: 1432-0894. DOI: 10.1007/s00382-016-3011-0.

Babu, C. A., P. R. Jayakrishnan, and Hamza Varikoden (2016). "Characteristics of Precipitation Pattern in the Arabian Peninsula and Its Variability Associated with ENSO". In: *Arabian Journal of Geosciences* 9.3, p. 186. ISSN: 1866-7538. DOI: 10.1007/s12517-015-2265-x.

Beaulant, A-L. et al. (2011). "Statistico-dynamical Downscaling for Mediterranean Heavy Precipitation". In: *Quarterly Journal of the Royal Meteorological Society* 137.656, pp. 736–748. ISSN: 0035-9009, 1477-870X. DOI: 10.1002/qj.796.

Branch, Oliver et al. (2020). "Convection Initiation over the Eastern Arabian Peninsula". In: *Meteorologische Zeitschrift* 29.1, pp. 67–77. ISSN: 0941-2948. DOI: 10.1127/metz/2019/0997.

Breiman, Leo (2001). "Random Forests". In: *Machine Learning* 45.1, pp. 5–32. ISSN: 1573-0565. DOI: 10.1023/A:1010933404324.

Charabi, Yassine (2009). "Arabian Summer Monsoon Variability: Teleconexion to ENSO and IOD". In: *Atmospheric Research* 91.1, pp. 105–117. ISSN: 0169-8095. DOI: 10.1016/j.atmosres.2008.07.006.

Charabi, Yassine and Salim Al-Hatrushi (2010). "Synoptic Aspects of Winter Rainfall Variability in Oman". In: *Atmospheric Research* 95.4, pp. 470–486. ISSN: 01698095. DOI: 10.1016/j.atmosres.2009.11.009.

Chen, Di and Aiguo Dai (2019). "Precipitation Characteristics in the Community Atmosphere Model and Their Dependence on Model Physics and Resolution". In: *Journal of Advances in Modeling Earth Systems* 11.7, pp. 2352–2374. DOI: https://doi.org/10.1029/2018MS001536. eprint: https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1029/2018MS001536.

Chen, Di, Aiguo Dai, and Alex Hall (2021). "The Convective-To-Total Precipitation Ratio and the "Drizzling" Bias in Climate Models". In: *Journal of Geophysical Research: Atmospheres* 126.16, e2020JD034198. ISSN: 2169-8996. DOI: 10.1029/2020JD034198.

Chen, Jiao et al. (2020). "Changes in Convective Available Potential Energy and Convective Inhibition under Global Warming". In: *Journal of Climate* 33.6, pp. 2025–2050. ISSN: 0894-8755, 1520-0442. DOI: 10.1175/JCLI-D-19-0461.1.

Chen, Tianqi and Carlos Guestrin (2016). "XGBoost: A Scalable Tree Boosting System". In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 785–794. DOI: 10.1145/2939672.2939785. arXiv: 1603.02754 [cs].

Cook, Kerry H. et al. (2021). "Greenhouse-Gas Induced Warming Amplification over the Arabian Peninsula with Implications for Ethiopian Rainfall". In: *Climate Dynamics* 57.11-12, pp. 3113–3133. ISSN: 0930-7575, 1432-0894. DOI: 10.1007/s00382-021-05858-x.

Corre, Lola et al. (2025). "Using Regional Warming Levels to Describe Future Climate Change for Services and Adaptation: Application to the French Reference Trajectory for Adaptation". In: *Climate Services* 38, p. 100553. ISSN: 24058807. DOI: 10.1016/j.cliser.2025.100553.

Dai, Aiguo (2006). "Precipitation Characteristics in Eighteen Coupled Climate Models". In: *Journal of Climate* 19.18, pp. 4605–4630. ISSN: 0894-8755, 1520-0442. DOI: 10.1175/JCLI3884.1.

Dankers, Frank J. W. M. et al. (2019). "Prediction Modeling Methodology". In: *Fundamentals of Clinical Data Science*. Ed. by Pieter Kubben, Michel Dumontier, and Andre Dekker. Cham: Springer International Publishing, pp. 101–120. ISBN: 978-3-319-99713-1. DOI: 10.1007/978-3-319-99713-1_8.

Dasari, Hari Prasad et al. (2018). "ENSO Influence on the Interannual Variability of the Red Sea Convergence Zone and Associated Rainfall". In: *International Journal of Climatology* 38.2, pp. 761–775. ISSN: 0899-8418, 1097-0088. DOI: 10.1002/joc.5208.

de Vries, A. J. et al. (2013). "Extreme Precipitation Events in the Middle East: Dynamics of the Active Red Sea Trough: ACTIVE RED SEA TROUGH DYNAMICS". In: *Journal of Geophysical Research: Atmospheres* 118.13, pp. 7087–7108. ISSN: 2169897X. DOI: 10.1002/jgrd.50569.

Doswell III, Charles A. and David M. Schultz (2006). "On the Use of Indices and Parameters in Forecasting Severe Storms". In: *E-Journal of Severe Storms Meteorology* 1.3, pp. 1–22. ISSN: 1559-5404. DOI: 10.55599/ejssm.v1i3.4.

Eager, Rebecca E. et al. (2008). "A Climatological Study of the Sea and Land Breezes in the Arabian Gulf Region". In: *Journal of Geophysical Research: Atmospheres* 113.D15. ISSN: 2156-2202. DOI: 10.1029/2007JD009710.

Farquharson, F. A. K., D. T. Plinston, and J. V. Sutcliffe (1996). “Rainfall and Runoff in Yemen”. In: *Hydrological Sciences Journal* 41.5, pp. 797–811. ISSN: 0262-6667. DOI: 10.1080/02626669609491546.

Fonseca, Ricardo et al. (2025). “Summertime Secondary Convection and Interaction with Sea-Breeze Circulations”. In: *Quarterly Journal of the Royal Meteorological Society* 151.766, e4907. ISSN: 1477-870X. DOI: 10.1002/qj.4907.

GISTEMP Team (2025). *GISS Surface Temperature Analysis (GISTEMP), version 4*. Accessed: 2025-11-13.

Gómez-Navarro, J. J. et al. (2022). “Event Selection for Dynamical Downscaling: A Neural Network Approach for Physically-Constrained Precipitation Events”. In: *Climate Dynamics* 58.9, pp. 2863–2879. ISSN: 1432-0894. DOI: 10.1007/s00382-019-04818-w.

Gopalakrishnan, Deepak et al. (2023). “Anatomy of a Summertime Convective Event over the Arabian Region”. In: *Monthly Weather Review* 151.4, pp. 989–1004. ISSN: 0027-0644, 1520-0493. DOI: 10.1175/MWR-D-22-0082.1.

Hagos, Samson M. et al. (2021). “The Relationship between Precipitation and Precipitable Water in CMIP6 Simulations and Implications for Tropical Climatology and Change”. In: *Journal of Climate* 34.5, pp. 1587–1600. ISSN: 0894-8755, 1520-0442. DOI: 10.1175/JCLI-D-20-0211.1.

Hausfather, Zeke et al. (2022). “Climate Simulations: Recognize the ‘Hot Model’ Problem”. In: *Nature* 605.7908, pp. 26–29. DOI: 10.1038/d41586-022-01192-2.

Hoens, T. Ryan and Nitesh V. Chawla (2013). “Imbalanced Datasets: From Sampling to Classifiers”. In: *Imbalanced Learning*. John Wiley & Sons, Ltd. Chap. 3, pp. 43–59. ISBN: 9781118646106. DOI: https://doi.org/10.1002/9781118646106.ch3. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/9781118646106.ch3.

Hohenegger, Cathy, Peter Brockhaus, and Christoph Schär (2008). “Towards Climate Simulations at Cloud-Resolving Scales”. In: *Meteorologische Zeitschrift* 17.4, pp. 383–394. ISSN: 0941-2948. DOI: 10.1127/0941-2948/2008/0303.

Homoudi, Ahmed (2026). *Dataset for "Convective Environments Over the Arabian Peninsula in Current and Future Climates: A Machine Learning Approach"*. Version 1.0.0. Accessed: 2026-06-05. DOI: 10.5281/zenodo.18514414.

Homoudi, Ahmed and Klemens Barfus (2025). *MeteoMate: Calculating Convective Indices for Climate Model Outputs*. R package version 1.0.0.

Homoudi, Ahmed et al. (2026). “Climatology of Convective Precipitation Systems Over the Arabian Peninsula Using Object-Tracking”. In: *International Journal of Climatology* 46.1, e70158. ISSN: 1097-0088. DOI: 10.1002/joc.70158.

Horsch, Karla, Maryellen L. Giger, and Charles E. Metz (2008). “Prevalence Scaling: Applications to an Intelligent Workstation for the Diagnosis of Breast Cancer”. In: *Academic Radiology* 15.11, pp. 1446–1457. ISSN: 1076-6332. DOI: https://doi.org/10.1016/j.acra.2008.04.022.

Houze Jr., Robert A. (2004). “Mesoscale convective systems”. In: *Reviews of Geophysics* 42.4. DOI: https://doi.org/10.1029/2004RG000150. eprint: https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1029/2004RG000150.

Huang, Boyin et al. (2024). *NOAA Global Surface Temperature Dataset (NOAAGlobalTemp), Version 6.0*. Accessed: 2025-11-27. DOI: 10.25921/rzxg-p717.

Jones, Donald R., Matthias Schonlau, and William J. Welch (1998). "Efficient Global Optimization of Expensive Black-Box Functions". In: *Journal of Global Optimization* 13.4, pp. 455–492. ISSN: 1573-2916. DOI: 10.1023/A:1008306431147.

Khan, Basit et al. (2018). "Climatology of Sea Breezes along the Red Sea Coast of Saudi Arabia". In: *International Journal of Climatology* 38.9, pp. 3633–3650. ISSN: 0899-8418, 1097-0088. DOI: 10.1002/joc.5523.

Kull, Meelis, Telmo M. Silva Filho, and Peter Flach (2017). "Beyond Sigmoids: How to Obtain Well-Calibrated Probabilities from Binary Classifiers with Beta Calibration". In: *Electronic Journal of Statistics* 11.2. ISSN: 1935-7524. DOI: 10.1214/17-EJS1338SI.

Lenssen, Nathan et al. (2024). "A NASA GISTEMPv4 Observational Uncertainty Ensemble". In: *Journal of Geophysical Research: Atmospheres* 129.17, e2023JD040179. ISSN: 2169-8996. DOI: 10.1029/2023JD040179.

Lepore, Chiara et al. (2021). "Future Global Convective Environments in CMIP6 Models". In: *Earth's Future* 9.12, e2021EF002277. ISSN: 2328-4277. DOI: 10.1029/2021EF002277.

Lin, Tsung-Yi et al. (2018). *Focal Loss for Dense Object Detection*. DOI: 10.48550/arXiv.1708.02002. arXiv: 1708.02002 [cs.CV].

Luong, Thang M., Hari P. Dasari, and Ibrahim Hoteit (2020). "Extreme Precipitation Events Are Becoming Less Frequent but More Intense over Jeddah, Saudi Arabia. Are Shifting Weather Regimes the Cause?" In: *Atmospheric Science Letters* 21.8, e981. ISSN: 1530-261X, 1530-261X. DOI: 10.1002/asl.981.

Luong, Thang M. et al. (2024). "Organized Precipitation and Associated Large-Scale Circulation Patterns over the Kingdom of Saudi Arabia". In: *International Journal of Climatology* 44.10, pp. 3295–3314. ISSN: 1097-0088. DOI: 10.1002/joc.8524.

Maraun, Douglas and Martin Widmann (2018). *Statistical Downscaling and Bias Correction for Climate Research*. 1st ed. Cambridge: Cambridge University Press. ISBN: 978-1-107-06605-2. DOI: 10.1017/9781107588783.

Martinez-Villalobos, Cristian, J. David Neelin, and Angeline G. Pendergrass (2022). "Metrics for Evaluating CMIP6 Representation of Daily Precipitation Probability Distributions". In: *Journal of Climate* 35.17, pp. 5719–5743. ISSN: 0894-8755, 1520-0442. DOI: 10.1175/JCLI-D-21-0617.1.

Meredith, Edmund P., Henning W. Rust, and Uwe Ulbrich (2018). "A Classification Algorithm for Selective Dynamical Downscaling of Precipitation Extremes". In: *Hydrology and Earth System Sciences* 22.8, pp. 4183–4200. ISSN: 1607-7938. DOI: 10.5194/hess-22-4183-2018.

Morice, C. P. et al. (2021). "An Updated Assessment of Near-Surface Temperature Change From 1850: The HadCRUT5 Data Set". In: *Journal of Geophysical Research: Atmospheres* 126.3, e2019JD032361. ISSN: 2169-8996. DOI: 10.1029/2019JD032361.

Naeini, Mahdi Pakdaman, Gregory Cooper, and Milos Hauskrecht (2015). "Obtaining Well Calibrated Probabilities Using Bayesian Binning". In: *Proceedings of the AAAI Conference on Artificial Intelligence* 29.1. ISSN: 2374-3468. DOI: 10.1609/aaai.v29i1.9602.

Nelli, Narendra et al. (2022). "Characterization of the Atmospheric Circulation near the Empty Quarter Desert during Major Weather Events". In: *Frontiers in Environmental Science* 10. ISSN: 2296-665X. DOI: 10.3389/fenvs.2022.972380.

Nelli, Narendra et al. (2025). "Drivers and Trends of Summertime Convection Over the Southeastern Arabian Peninsula". In: *Geophysical Research Letters* 52.23, e2025GL118960. ISSN: 0094-8276, 1944-8007. DOI: 10.1029/2025GL118960.

Nelson, Michael et al. (1999). "Time series forecasting using neural networks: should the data be deseasonalized first?" In: *Journal of Forecasting* 18.5, pp. 359–367. ISSN: 1099-131X. DOI: 10.1002/(SICI)1099-131X(199909)18:5<359::AID-FOR746>3.0.CO;2-P.

Niculescu-Mizil, Alexandru and Rich Caruana (2005). "Predicting Good Probabilities with Supervised Learning". In: *Proceedings of the 22nd International Conference on Machine Learning*. ICML '05. New York, NY, USA: Association for Computing Machinery, pp. 625–632. ISBN: 978-1-59593-180-1. DOI: 10.1145/1102351.1102430.

Ojeda, Francisco M. et al. (2023). "Calibrating Machine Learning Approaches for Probability Estimation: A Comprehensive Comparison". In: *Statistics in Medicine* 42.29, pp. 5451–5478. ISSN: 1097-0258. DOI: 10.1002/sim.9921.

Pathak, Raju et al. (2025). "Dynamics of Intensification of Extreme Precipitation Events over the Arabian Peninsula Derived from CMIP6 Simulations". In: *npj Climate and Atmospheric Science* 8.1, p. 126. ISSN: 2397-3722. DOI: 10.1038/s41612-025-01016-w.

Pozzolo, Andrea Dal et al. (2015). "Calibrating Probability with Undersampling for Unbalanced Classification". In: *2015 IEEE Symposium Series on Computational Intelligence*. Cape Town, South Africa: IEEE, pp. 159–166. ISBN: 978-1-4799-7560-0. DOI: 10.1109/SSCI.2015.33.

Prein, Andreas F. et al. (2015). "A Review on Regional Convection-permitting Climate Modeling: Demonstrations, Prospects, and Challenges". In: *Reviews of Geophysics* 53.2, pp. 323–361. ISSN: 8755-1209, 1944-9208. DOI: 10.1002/2014RG000475.

Queiroz, Amanda Carolina da Silva and Thiago Souza Biscaro (2026). "Seasonal Variability of Precipitating Systems in Four Radar Domains of Northeast Brazil". In: *International Journal of Climatology* 46.6, e70295. DOI: https://doi.org/10.1002/joc.70295. eprint: https://rmets.onlinelibrary.wiley.com/doi/pdf/10.1002/joc.70295.

Rao, K. Koteswara et al. (2024). "Future Changes in the Precipitation Regime over the Arabian Peninsula with Special Emphasis on UAE: Insights from NEX-GDDP CMIP6 Model Simulations". In: *Scientific Reports* 14.1, p. 151. ISSN: 2045-2322. DOI: 10.1038/s41598-023-49910-8.

Rohde, Robert A. and Zeke Hausfather (2020). "The Berkeley Earth Land/Ocean Temperature Record". In: *Earth System Science Data* 12.4, pp. 3469–3479. ISSN: 1866-3508. DOI: 10.5194/essd-12-3469-2020.

Saerens, Marco, Patrice Latinne, and Christine Decaestecker (2002). "Adjusting the Outputs of a Classifier to New a Priori Probabilities: A Simple Procedure". In: *Neural Computation* 14.1, pp. 21–41. ISSN: 0899-7667. DOI: 10.1162/089976602753284446.

Al-Sakkaf, Ali Salem et al. (2024). "Assessing Exposure to Climate Extremes over the Arabian Peninsula Using ERA5 Reanalysis Data: Spatial Distribution and Temporal Trends". In: *Atmospheric Research* 300, p. 107224. ISSN: 01698095. DOI: 10.1016/j.atmosres.2024.107224.

Schulzweida, Uwe (2023). *CDO User Guide*. Version 2.3.0. DOI: 10.5281/zenodo.10020800.

Seneviratne, Sonia I. et al. (2016). "Allowable CO2 Emissions Based on Regional and Impact-Related Climate Targets". In: *Nature* 529.7587, pp. 477–483. ISSN: 1476-4687. DOI: 10.1038/nature16542.

Song, Jinyan et al. (2024). "Realistic Precipitation Diurnal Cycle in Global Convection-Permitting Models by Resolving Mesoscale Convective Systems". In: *Geophysical Research Letters* 51.13, e2024GL109945. ISSN: 1944-8007. DOI: 10.1029/2024GL109945.

Sun, Yanmin, Andrew K. C. Wong, and Mohamed S. Kamel (2009). "Classification of Imbalanced Data: A Review". In: *International Journal of Pattern Recognition and Artificial Intelligence* 23.04, pp. 687–719. ISSN: 0218-0014, 1793-6381. DOI: 10.1142/S0218001409007326.

Tang, Shuaiqi et al. (2021). "Evaluating the Diurnal and Semidiurnal Cycle of Precipitation in CMIP6 Models Using Satellite- and Ground-Based Observations". In: *Journal of Climate* 34.8, pp. 3189–3210. ISSN: 0894-8755, 1520-0442. DOI: 10.1175/JCLI-D-20-0639.1.

Taraphdar, Sourav et al. (2025). "Subtropical Jet Regulates Arabian Winter Precipitation: A Viable Mechanism". In: *Journal of the Atmospheric Sciences* 82.4, pp. 713–732. ISSN: 0022-4928, 1520-0469. DOI: 10.1175/JAS-D-23-0213.1.

Taszarek, Mateusz et al. (2021). "Global Climatology and Trends in Convective Environments from ERA5 and Rawinsonde Data". In: *npj Climate and Atmospheric Science* 4.1, pp. 1–11. ISSN: 2397-3722. DOI: 10.1038/s41612-021-00190-x.

Trenberth, Kevin E E, Jeffrey C Berry, and Lawrence E Buja (1993). *Vertical Interpolation and Truncation of Model-Coordinate Data*. Tech. rep. NCAR/TN-396+STR. Boulder, CO, USA: University Corporation for Atmospheric Research, p. 60. DOI: 10.5065/D6HX19NH.

Vahid Yousefnia, Kianusch et al. (2024). "A Machine-Learning Approach to Thunderstorm Forecasting through Post-Processing of Simulation Data". In: *Quarterly Journal of the Royal Meteorological Society* 150.763, pp. 3495–3510. ISSN: 0035-9009. DOI: 10.1002/qj.4777.

Wang, Songning et al. (2026). "Future Climate Projections of Hazardous Convective Weather Using an Ensemble of Environment-informed, Convection-permitting Dynamical Downscaling Simulations". In: (). DOI: 10.1029/2025JD045354.

Wang, Yu, Yalei You, and Mark Kulie (2018). "Global Virga Precipitation Distribution Derived From Three Spaceborne Radars and Its Contribution to the False Radiometer Precipitation Detection". In: *Geophysical Research Letters* 45.9, pp. 4446–4455. ISSN: 1944-8007. DOI: 10.1029/2018GL077891.

Wilks, Daniel S. (2019). "Chapter 9 - Forecast Verification". In: *Statistical Methods in the Atmospheric Sciences (Fourth Edition)*. Ed. by Daniel S. Wilks. Fourth Edition. Elsevier, pp. 369–483. ISBN: 978-0-12-815823-4. DOI: 10.1016/B978-0-12-815823-4.00009-2.

Wright, Marvin N. and Andreas Ziegler (2017). "Ranger: A Fast Implementation of Random Forests for High Dimensional Data in C++ and R". In: *Journal of Statistical Software* 77, pp. 1–17. ISSN: 1548-7660. DOI: 10.18637/jss.v077.i01.

Xu, Shuo et al. (2026). "LLM-based annotation and token-augmented modeling for emotional tone classification in online cancer peer-support posts". In: *PLOS Digital Health* 5.5, pp. 1–17. DOI: 10.1371/journal.pdig.0001235.

Zhang, Binglin et al. (2025). "Evaluation of the Performance of CMIP6 Models in Simulating Extreme Precipitation and Its Projected Changes in Global Climate Regions". In: *Natural Hazards* 121.2, pp. 1737–1763. ISSN: 1573-0840. DOI: 10.1007/s11069-024-06850-4.