Evaluation of Climate Models For Seasonal Forecasting in the MENA Region

Prepared by:

Berrahmouch Nohayla and Mohamed El-Badri

Hassania School of Public Works, Casablanca, Morocco

Supervised by:

Internal Supervisor:

Mr. Driss Bari

Hassania School of Public Works
and Direction Générale de la Météorologie, Morocco

External Supervisors:

Mrs. Wafae Badi

DGM, Morocco

Mr. Nicholas Savage

Met Office, Exeter, UK

2024 - 2025

Contents

ACKNOWLEDGMENTS

We would like to express our deepest gratitude to **Mrs. Wafae Badi** for her unwavering support and invaluable guidance throughout this project. Her insightful counsel and constant encouragement enabled us to overcome various challenges and maintain our focus. Her dedication to fostering progress, coupled with her constructive feedback and open-minded approach, were instrumental in shaping the direction of this work. Her mentorship has truly been a cornerstone of our journey, and we are profoundly grateful for her invaluable contributions.

Special thanks go to **Mr. Nicholas Savage** and his exceptional team at the UK Met Office. Their generosity in sharing their expertise and resources provided us with unparalleled opportunities to broaden our understanding of climate modeling. The engaging discussions and valuable insights shared by Mr. Savage and his team not only enriched this project but also fueled our motivation to explore innovative avenues. Their commitment to advancing climate science inspired us to aim higher and achieve more.

We extend our sincere thanks to **Mrs. Emma Dyer** for her thoughtful feedback and diligent follow-up throughout this project. Her valuable remarks and attention to detail were essential in refining our work and enhancing its quality. Her commitment to excellence and her dedication to supporting our efforts were greatly appreciated.

We are also immensely grateful to **Mr. Bari**, whose dedicated supervision, thoughtful suggestions, and constructive critiques significantly enhanced the quality of this work. His ability to balance critical feedback with motivating encouragement made a remarkable difference, guiding us through challenging moments and ensuring steady progress.

In addition, we would like to acknowledge the support received through the **WISER MENA project**. **Nicholas Savage's time was funded via the WISER MENA project.** The Weather and Climate Information Services (WISER) Programme is funded with UK International Development from the UK government and led by the Met Office in the UK. This work has been partially supported by UK International Development from the UK government; however, the views expressed do not necessarily reflect the UK government's official policies.

Lastly, we extend our heartfelt appreciation to all those who, directly or indirectly, contributed to this project. Your cooperation, guidance, and belief in our work have made this journey a fulfilling and enlightening experience. While this project is a testament to hard work and collaboration, it is also a reflection of the collective effort and support of everyone who believed in its success. To you, we owe our sincere thanks.

PREFACE

The MENA seasonal forecasting models have undergone both probabilistic and deterministic evaluations. This research study is regarded as the pioneering work and the first of its kind in this area which helps in situational context improvement in seasonal forecasting models. Given the alarming rate of increase in the impacts caused by extreme climatic events including severe droughts, and extreme heat and other climate sensitive issues in the MENA region, this work is a key contribution towards alleviating these issues=

Due to climatic extremes in the MENA region, agriculture, human livelihood, and natural resources are heavily affected. Consequently, it has become almost necessary to have forecasts of seasons that are credible so as to characterize the impacts, or to enhance preparedness. Although seasonal forecasting models have been widely researched and practiced in many parts of the world, their use in MENA countries' local level remains scarce. This gap is resolved in this study, providing new knowledge and tools for climate scientists working in the region.

In this work, we intend to broaden the knowledge fabric of climate change science by focusing on the climate change and variability vulnerability of the MENA region. The results obtained not only improve the comprehension of the dynamics of the local climate, but also lays a framework for specific approach to be employed for adaptation strategies.

We are immensely grateful to every individual or organization who has helped support this project and guided us through uncharted territory in the spectrum of MENA climate predictions.

CHAPTER 1

OVERVIEW AND RATIONALE OF THE STUDY

The last couple of decades have witnessed a surge in demand for seasonal climate forecasting. Global advancements in space science and technology have lead to the better anticipation of climate seasons up to a through range of 3-12 months. This is crucial for effective planning in major industries like agriculture or energy management, among others. These advancements breed an increased dependence on seasonal forecasting and in turn create a higher demand for accurate forecasting mechanisms. Therefore two central methodologies have witnessed prominence – deterministic and probabilistic methods. A hindsight understanding of these mechanisms is imperative, as they are useful for evaluating and understanding the shortcomings and effectiveness of different models employed in forecasting seasonal amps.

Probabilistic forecasts take one step forward, do not try to predict an ideal scenario and present different potential outcomes, each with a defined probability. Efforts, though different, instruct towards the same ends; meeting a specific operational/strategic need. Lorenz's butterfly effect presents the case for one such endeavor- it shows how a non-linear system's response can drastically alter depending on the initial conditions. Such chaos is especially present in weather and climate systems where even the slightest details can have large ramifications over longer periods.

In this context, the current study focuses on developing relationships that integrate conceptual advancements in seasonal forecasting efforts with practical and applicable methods. The geographical focus of this study is the **MENA region (Middle East and North Africa)**, a zone characterized by its diverse climatic conditions and critical socio-economic dependence on accurate seasonal forecasts. The MENA region is particularly vulnerable to climate variability due to its arid and semi-arid environments, limited water resources, and reliance on agriculture and energy. By analyzing seasonal forecasting within this region, the study seeks to address key challenges and contribute to sustainable climate-resilient solutions.

CHAPTER 2

INTRODUCTION

2.1 Context

2.1.1 Overview of Climate Modeling and Seasonal Forecasting

Climate modeling is the process of using mathematical representations of the Earth's atmosphere, oceans, land surface, and ice systems to simulate and predict climate dynamics. These models are based on fundamental physical principles, such as the conservation of mass, energy, and momentum, and are implemented through numerical methods that solve complex equations governing the interactions between these systems.[1] Climate models range from global circulation models (GCMs), which simulate large-scale atmospheric and oceanic processes, to regional climate models (RCMs), which provide localized projections by incorporating finer-scale topographic and land-use details.[2] Seasonal forecasting, a subset of climate modeling, refers to the prediction of climate conditions, such as temperature and precipitation, over a period of one to six months. These forecasts rely on initial conditions (e.g., sea surface temperatures, soil moisture) and slowly varying components of the climate system, such as oceanic or atmospheric anomalies like the El Niño-Southern Oscillation (ENSO).[3] The basic principle behind seasonal forecasting is to leverage these slowly varying components, which have a predictable influence on regional weather patterns, using ensemble simulations to quantify uncertainties and provide probabilistic predictions.[4]

Seasonal forecasts play a crucial role in decision-making and planning across various sectors, including agriculture, water management, and climate risk mitigation. These forecasts provide early warnings of high-impact climate scenarios, enabling proactive decisions that result in financial savings, risk reduction, and optimized resource use. For instance, in agriculture, they assist farmers in selecting appropriate crops and determining optimal planting times based on anticipated water availability, thereby mitigating risks associated with droughts or excessive rainfall.[5]

[1] McGuffie, K. and Henderson-Sellers, A., 2014. A Climate Modelling Primer. https://doi.org/10.1002/9781118687853

[2] Flato et al., 2013. Evaluation of Climate Models. IPCC AR5 Chapter 9. https://www.ipcc.ch/report/ar5/wg1/chapter-9-evaluation-of-climate-models/

[3] Doblas-Reyes, F. J., García-Serrano, J., Lienert, F., Biescas, A. P., & Rodrigues, L. R., 2013. Seasonal climate predictability and forecasting: Status and prospects. https://doi.org/10.1038/ngeo1714

[4] Palmer, T. N., & Anderson, D. L., 1994. The prospects for seasonal forecasting—a review paper. https://doi.org/10.1256/smsqj.50402

[5] Werner, M. and Linés, C., 2024. Seasonal forecasts to support cropping decisions. https://doi.org/10.5194/egusphere-egu24-13436

Seasonal forecasts also support pre-harvest strategies, such as hedging decisions, which help shield farmers from price volatility, although their adoption is often hindered by perceptions of inaccuracy and complexity.[6] In water management, seasonal forecasts are vital for mitigating drought impacts, particularly in semi-arid regions, by enabling improved reservoir operations and efficient water allocation to reduce losses.[7] Additionally, these forecasts, when linked to hydrological models, improve predictions of water balance and inform critical decisions regarding water storage and distribution, despite occasional discrepancies between predicted and desired variables.[8] Seasonal forecasts are increasingly applied in climate risk management, where they help predict extreme weather events, providing decision-makers with tools to minimize societal and economic damages.[9] For example, accurate predictions of heatwaves or floods allow authorities to implement adaptive measures, reducing infrastructure damage and safeguarding public health. In economic sectors such as energy and water management, tailored seasonal forecasts enhance decision-making efficiency by aligning forecasts with user needs, thereby optimizing outcomes.[10] Despite their significant potential, the effectiveness of seasonal forecasts depends on their accuracy, relevance to user needs, and ease of use. Improved communication, stakeholder training, and efforts to bridge the gap between forecast complexity and user understanding are essential to maximize their utility.

2.1.2 Importance of Seasonal Climate Forecasts in MENA

Seasonal climate forecasts are critically important across the MENA region, where high temperatures, low water availability, and vulnerability to climate variability create substantial challenges for sustainable development. Forecasts provide early warnings of droughts, heatwaves, and other extreme weather events, enabling decision-makers to implement proactive measures to mitigate impacts on water resources, agriculture, and infrastructure.[11] In agriculture, these forecasts help farmers optimize crop selection and planting schedules, reducing the risks of crop failure in this water-scarce region.[12] In the water sector, seasonal forecasts guide reservoir management by predicting rainfall variability, improving water storage strategies, and ensuring more equitable water distribution.[13] With increasing climate risks, these forecasts also support disaster risk management by allowing governments to prepare for extreme events, such as heatwaves and floods, which are becoming more frequent in the region due to climate change.[14] Moreover, the economic benefits of using seasonal forecasts are significant. By enabling energy companies to anticipate peak demand periods driven by heatwaves, and by helping municipalities optimize water usage during droughts, these forecasts provide cost savings and efficiency gains.[15] However, challenges persist in ensuring the accuracy and usability of these forecasts. The arid and semi-arid nature of much of the MENA

[6] Hunt et al., 2020. Seasonal Forecast Based Preharvest Hedging. https://doi.org/10.22004/AG.ECON.309761

[7] Portele et al., 2021. Seasonal forecasts offer economic benefits for hydrological decision-making. https://doi.org/10.1038/s41598-021-89564-y

[8] MacLeod et al., 2023. Translating seasonal climate forecasts into water balance forecasts. https://doi.org/10.1371/journal.pclm.0000138

[9] Castino et al., 2023. Towards seasonal prediction of extreme temperature indices. https://doi.org/10.5194/ems2023-590

[10] Goodess et al., 2022. The Value-Add of Tailored Seasonal Forecast Information. https://doi.org/10.3390/cli10100152

[11] Dunn et al., 2020. The changing climate of MENA. https://pubs.giss.nasa.gov/abs/gu00200u.html

[12] Werner, M., and Linés, C., 2024. Seasonal forecasts to support cropping decisions. https://doi.org/10.5194/egusphere-egu24-13436

[13] Portele et al., 2021. Seasonal forecasts for hydrological decision-making. https://doi.org/10.1038/s41598-021-89564-y

[14] Castino et al., 2023. Towards seasonal prediction of extreme temperature indices. https://doi.org/10.5194/ems2023-590

[15] Goodess et al., 2022. Value-Add of tailored seasonal forecast information. https://doi.org/10.3390/cli10100152

region, coupled with complex interactions between regional climate drivers, makes it difficult to provide highly localized forecasts.[16] Addressing these challenges through improved modeling techniques and stakeholder engagement will be critical to maximizing the value of seasonal forecasts in the MENA region, ensuring better preparedness and resilience against a changing climate.

2.2 Objectives of the Work and Description of Report Content

The primary objective of this work is to evaluate the effectiveness of climate models, focusing specifically on their performance in predicting key climate variables such as temperature, precipitation. This evaluation incorporates both deterministic and probabilistic approaches to identify the most skillful models and their suitability for practical applications.

2.2.1 Specific aims of evaluating deterministic and probabilistic models.

The evaluation of deterministic and probabilistic models is essential for understanding their unique strengths, limitations, and potential applications in diverse fields. Deterministic models, which generate a single, precise outcome based on initial conditions, are widely used when exactness and reproducibility are critical, such as in engineering and physical simulations.[17] Their evaluation focuses on assessing accuracy and reliability under specific conditions, providing clarity in cause-and-effect relationships. In contrast, probabilistic models incorporate uncertainty by assigning probabilities to various potential outcomes, enabling the representation of real-world complexities and variability.[18] These models are particularly beneficial for strategic planning and risk management, where understanding a range of possible scenarios is crucial. The evaluation of both types of models includes conducting sensitivity analyses to determine how changes in input variables affect outcomes, which helps in identifying key drivers of uncertainty and improving model performance.[19] Additionally, risk assessment is a vital component, with deterministic approaches offering straightforward estimations for defined scenarios, while probabilistic approaches address uncertainties by simulating a spectrum of possible outcomes.[20] These evaluations also aim to support decision-making processes by identifying which type of model is more appropriate for specific contexts—deterministic models for precise predictions and probabilistic models for flexible planning under uncertainty.[21] Finally, probabilistic models are often recognized for their adaptability in dynamic environments, as they can incorporate new data and adjust probability distributions to reflect evolving conditions, making them indispensable for complex systems where deterministic models may fall short.[22] Together, the evaluation of deterministic and probabilistic models provides invaluable insights into their suitability for addressing specific challenges, supporting informed decision-making, and advancing model development.

[16]Latif et al., 2011. ENSO predictability and regional climate impacts. https://doi.org/10.1175/2010JCLI3405.1

[17]McGuffie, K., and Henderson-Sellers, A., 2014. *A Climate Modelling Primer*. Wiley. https://doi.org/10.1002/9781118687870

[18]Palmer, T., and Hagedorn, R., 2006. *Predictability of Weather and Climate*. Cambridge University Press. https://doi.org/10.1017/CBO9780511617652

[19]Seneviratne, S.I., et al., 2021. *Metrics for climate model evaluation: A review*. Nature Communications. https://doi.org/10.1038/s43247-021-00094-x

[20]PreventionWeb, 2021. *Deterministic and Probabilistic Risk*. https://www.preventionweb.net/understanding-disaster-risk/key-concepts/deterministic-probabilistic-risk

[21]Goodess, C.M., et al., 2022. *The Value-Add of Tailored Seasonal Forecast Information for Industry Decision Making*. Climate. https://doi.org/10.3390/cli10100152

[22]Latif, M., and Keenlyside, N., 2011. *El Niño/Southern Oscillation Predictability*. Journal of Climate. https://doi.org/10.1175/2010JCLI3405.1

2.2.2 Description of Content

This report is designed to provide a comprehensive analysis of climate model evaluation, focusing on both deterministic and probabilistic approaches. The structure of the report follows a logical progression, starting with an introduction to the fundamental concepts behind climate models. The first section lays the groundwork for understanding the key differences between deterministic and probabilistic models, describing how each approach is used to simulate climate systems and predict future outcomes. The methodology chapter follows, detailing the specific techniques employed to assess the models. This includes the use of both deterministic and probabilistic metrics such as Root Mean Square Error (RMSE), Anomaly Correlation Coefficient (ACC), and Brier Score, which are critical for evaluating the models' accuracy and performance in predicting climate variables like temperature and precipitation.

Next, the report moves on to the results and analysis, where the performance of the selected models is presented and compared. This section highlights the models' strengths and weaknesses, providing insight into how well they predict climate patterns across various geographical regions and time periods. Special attention is given to the models' skill in forecasting extreme weather events, which are particularly relevant to sectors like agriculture, water resource management, and disaster risk reduction.

The final section of the report provides conclusions and recommendations based on the analysis. This chapter synthesizes the findings, offering practical suggestions for improving the accuracy, usability, and application of climate forecasts. Recommendations also address how future developments in climate modeling can better meet the needs of decision-makers and stakeholders. The report as a whole seeks to contribute valuable insights into the ongoing development of climate prediction systems, aiming to enhance their effectiveness in real-world applications.

CHAPTER 3

LITERATURE REVIEW

3.1 Overview of Climate Models

3.1.1 Deterministic Models

Deterministic models rely on mathematical equations that describe the physical processes of the atmosphere. These models use fixed initial conditions to provide precise predictions, making them suitable for short-term forecasting. However, due to the chaotic nature of atmospheric systems, as demonstrated by Lorenz's theorem, deterministic models are limited in their ability to predict long-term outcomes. Small errors in initial conditions can lead to significant differences in results, reducing their reliability for seasonal or long-term forecasting.[1]

Deterministic climate models operate based on fixed initial conditions and mathematical equations that simulate physical processes in the atmosphere. These models are particularly useful for short-term predictions as they provide precise and singular forecasts. However, deterministic models are significantly limited when forecasting over extended periods. This limitation arises due to the inherent sensitivity of atmospheric systems to initial conditions—a concept known as the *butterfly effect*, introduced by Edward Lorenz in 1963. His research demonstrated that even minute changes in the initial conditions of a system could lead to vastly different outcomes over time, emphasizing the chaotic nature of weather systems.

For seasonal forecasting, deterministic models often fail because minor errors in the initial conditions can amplify, resulting in inaccurate predictions for longer timescales. Despite these challenges, deterministic models are vital for understanding specific phenomena over shorter durations with high spatial and temporal resolution.

3.1.2 Probabilistic Models

Probabilistic models address the limitations of deterministic approaches by incorporating uncertainty into forecasts. Instead of producing a single outcome, these models generate a range of possible scenarios, each with an associated probability, using ensemble simulations or statistical techniques. This makes probabilistic models particularly useful for medium- to long-term forecasts and risk assessment in climate-sensitive sectors such as agriculture, water management, and disaster

[1]Lorenz, E. N. (1963). Deterministic Nonperiodic Flow. *Journal of the Atmospheric Sciences, 20*(2), 130–141.

mitigation.[2]

The evaluation of probabilistic models relies on metrics that assess their ability to represent uncertainty and provide actionable insights:

- **Reliability:** Measures how well predicted probabilities align with observed frequencies.
- **Resolution:** Assesses the model's ability to distinguish between different outcomes.
- **Discrimination:** Evaluates the model's ability to separate events from non-events.[3]

Probabilistic models are especially valuable for decision-making under uncertainty, as they provide stakeholders with a clearer understanding of risks and potential scenarios, enabling proactive measures to mitigate impacts.

Comparison of Deterministic and Probabilistic Models

Deterministic and probabilistic models serve complementary roles in climate modeling and forecasting. Their distinct features and applications are summarized in Table **??**.

Table 3.1: Comparison of Deterministic and Probabilistic Models

Feature	**Deterministic Models**	**Probabilistic Models**
Predictability	Produces a single fixed outcome based on initial conditions	Generates a range of outcomes with associated probabilities
Sensitivity to Initial Conditions	Highly sensitive, leading to reduced accuracy over long timeframes	Less sensitive due to ensemble techniques reducing error amplification
Application Domain	Suitable for short-term, high-resolution tasks, e.g., extreme event analysis	Ideal for medium- and long-term decision-making under uncertainty
Use of Historical Data	Limited emphasis on historical variability	Extensively relies on historical data for statistical projections
Examples	Global Circulation Models (GCMs), Regional Climate Models (RCMs)	Ensemble forecasting, statistical downscaling

While deterministic models are preferred for precise and short-term predictions, probabilistic models provide critical insights into the likelihood of various scenarios, making them indispensable for managing climate-related risks.

[2]World Meteorological Organization (2024). *Guidance on Verification of Operational Seasonal Climate Forecasts.* https://library.wmo.int/records/item/56227-guidance-on-verification-of-operational-seasonal-climate-forecasts

[3]Rapport de projet 2024–2025, *3rd Year Meteorology Modeling Project.*

3.1.3 STUDIES IN "MENA" REGION

The current and changing climate in MENA

Much [4] of the MENA region is characterised by high temperature and low water availability, a combination of variables that have the potential to lead towards the environmental limits/threshold for safe human habitation. This makes the region particularly vulnerable to climate change and climate variability, as small variations in climate can easily produce high temperatures or extensive droughts that are harmful to human lives and livelihoods.

Changes in temperature and rainfall patterns have already been observed in the region and are expected to change further in the near future, especially if global warming exceeds 1.5 to 2 ℃ above the pre-industrial level. Annual mean temperatures across the MENA region have increased between 0.3–0.5℃ per decade1 over the period 1980–2015 [5]. Since the 1950s, hot and cold extremes have become warmer, the number of cold days has decreased, and the number of warm days has increased (Dunn et al., 2020). There has been an increase in heat waves intensity, frequency and duration [6]. Annual mean precipitation shows a high level of spatial variability over the MENA region. During the period 1980–2015 there have been downward trends in mean annual precipitation [7] . Dry conditions, drought intensity and frequency has increased in the past over the region [8]

Impact-Based Evaluation

Impact-based forecasting refers to a type of weather or climate forecasting that goes beyond predicting the meteorological parameters (e.g., temperature, rainfall, wind speed) and instead focuses on predicting the potential impacts of those conditions on society, infrastructure, and ecosystems. The goal is to provide actionable insights that help communities and decision-makers prepare for and mitigate the effects of extreme weather and climate events.

Evaluation of Seasonal Forecast Models

An impact-based evaluation[9] [10] was conducted as global study on five seasonal forecast models to identify the most effective for extreme precipitation forecasting (focuses on regions which were vulnerable to wildfire and flooding). The models assessed included:

- Centro Euro-Mediterraneo sui Cambiamenti Climatici (CMCC: version 35),
- Deutscher Wetterdienst (DWD: version 21),
- Environment and Climate Change Canada (ECCC: version 3),
- Météo-France (version 8),
- UK Met Office (UK-Met: version 601).

The findings highlighted the ***UK-Met*** and ***Météo-France*** models as consistently superior across all four seasons. Meanwhile, the ECCC and CMCC models exhibited strong performance on specific indices and in particular regions, ranking just below the top two models.

[4] Met Office WISER Report
[5] (Gutiérrez et al., 2021)
[6] (Perkins-Kirkpatrick and Lewis, 2020)
[7] (Gutiérrez et al., 2021)
[8] (Seneviratne et al., 2021).
[9] https://agupubs.onlinelibrary.wiley.com/doi/full/10.1029/2024EF004936
[10] Zahir Nikraftar, Rendani Mbuvha, Mojtaba Sadegh, Willem A. Landman

ROC Scores and Regional Performance

The ROC scores indicate that forecast models perform exceptionally well in tropical and subtropical regions. This result is consistent with our study and can be attributed to the general predictability of oceanic conditions and the influence of climate drivers such as the El Niño-Southern Oscillation (ENSO). The Météo-France and UK-Met models exhibited superior performance during the SON and MAM seasons.

However, the prevalence of grids with no discrimination ROC categories is more common in extratropical regions. This can be attributed to:

- Lower predictability of extratropical variations,
- Model limitations in capturing interactions between tropical and extratropical regions,
- Challenges in representing land surface processes (De Andrade et al., 2019).

The CMCC, DWD, and ECCC models often fail to detect extreme events in many extratropical areas, underscoring the stronger performance of the UK-Met and Météo-France models in these scenarios.

Percent Bias Analysis

The analysis of Percent Bias across four seasons demonstrates a consistent underestimation by forecast models for most extreme wet precipitation indices. Key observations include:

- Forecast models underestimate extreme wet precipitation indices while overestimating light precipitation.
- Models perform better in capturing the intensity and magnitude of extreme events (e.g., highest daily and multi-day rainfall) compared to the frequency of wet or dry days.

In tropical and subtropical regions, models like ***UK-Met*** and ***Météo-France*** exhibit strong performance due to their ability to capture large-scale climate patterns. In contrast, extratropical regions show higher biases, reflecting challenges in modeling complex interactions and seasonal variations.

Global Model Comparison

The ***UK-Met*** model consistently demonstrates lower biases and stronger performance globally compared to the ***Météo-France*** model, highlighting its effectiveness in representing climate patterns. However, all models show limitations in accurately modeling persistent extreme wet and dry periods, particularly in extratropical areas.

SYSTEM 7 FRANCE

seasonal forecasting evaluation has been the subject of numerous studies, with a focus on improving the accuracy and reliability of predictions related to precipitation and other weather parameters. One such study[11] conducted a probabilistic evaluation of seasonal precipitation re-forecasting from May to November over a period of 23 years (1993–2015). The study utilized the Brier Score (BS) and its decomposition to assess forecast performance, with the aim of providing more reliable and actionable predictions for extreme weather events.

The evaluation was conducted on the operational seasonal forecasting system of Meteo-France, which used 25 ensemble members, perturbed model dynamics, and initial conditions. The system aimed to provide a more detailed probabilistic forecast, in addition to existing deterministic metrics,

[11] https://www.mdpi.com/2674-0494/1/3/16

for both seasonal and intra-seasonal forecasts. The BS was estimated using tercile probabilities and a non-parametric counting estimator, with the GPCP[12] observation data serving as the reference.

Multiple analyses were performed to evaluate the robustness of the BS score, revealing that spatial distributions of the BS can vary significantly based on the sampling methods, reference data, and ensemble types used. The analysis showed that large errors, especially in the tropical ocean, could be reduced by using hindcast ensemble climatological samples. In particular, errors over the Nino region in the Pacific Ocean could be mitigated using these methods. This highlights the importance of employing various ensemble data sources and reference climatology to enhance the reliability of seasonal forecasts.

A notable finding was the reduction in BS when using ensemble observations, especially in the tropical ocean, suggesting that increasing ensemble size can improve forecast accuracy up to a point. However, this was not the case in all regions, as some areas, such as the tropical Indian Ocean, exhibited high BS even with different analysis methods. The study also found that intra-seasonal analyses showed similar patterns to seasonal hindcasts, but with higher BS due to reduced sample sizes, highlighting the need for higher-resolution models and improved initial conditions.

The study concluded that, despite improvements, probabilistic forecasting still faces challenges, particularly in the tropical regions, where errors fluctuate with lead time. The study emphasized the need for continued development of forecasting methods, particularly in reducing uncertainties in evaluation scores. Future evaluations should expand beyond the BS to include other metrics, such as the forecast skill score and the relative operating characteristic (ROC), to better assess forecast performance and identify system deficiencies.

This study's findings underline the importance of ensemble forecasting and the use of diverse data sources to improve the accuracy of seasonal precipitation forecasts, particularly in tropical regions where predictability remains challenging.

3.1.4 Evaluation Approaches

In the WMO[13] . Guide, several criteria are provided for evaluating a good forecast. Each criterion offers insight into specific aspects of the model but cannot, on its own, fully determine the forecast's quality. By combining all the criteria, we can comprehensively assess the performance of the model.

Resolution

Resolution measures whether the outcome differs given different forecasts, while discrimination measures whether the forecasts differ given different outcomes.

Discrimination looks at how well your forecast separates cases when the event (outcome) happens (pass) from when it doesn't happen (fail). It's about telling apart the events. Resolution looks at how well your forecast adapts to different situations, giving distinct probabilities for different cases. It's about adjusting to the situation.

Resolution measures how well a forecast distinguishes between different outcomes. A forecast has high resolution if the predicted probabilities vary significantly depending on the actual outcome. In other words, resolution tells us whether the forecast changes (e.g., gives different probabilities) when the actual outcome changes. High resolution: The forecast gives distinct and varying probabilities when different events (outcomes) occur. For example, if in one case the forecast predicts a high probability of rain and it rains, and in another case predicts a low probability and it doesn't rain, the forecast shows good resolution. Low resolution: If the forecast probabilities don't change

[12]Global Precipitation Climatology Project (GPCP)

[13]https://library.wmo.int/records/item/56227-guidance-on-verification-of-operational-seasonal-climate-forecasts

much regardless of whether it rains or not (e.g., always predicting a 50% chance of rain), the forecast has poor resolution because it fails to capture the differences in actual outcomes. Resolution can be determined by measuring how strongly the outcome is conditioned upon the forecast. If the outcome is independent of the forecast, the forecast has no resolution and is useless Forecasts with no resolution are neither "good" nor "bad", but are useless. Metrics of resolution distinguish between potentially useful and useless forecasts, but not all these metrics distinguish between "good" and "bad" forecasts.

The following equation represents the "resolution" component of the Brier Score (BS) decomposition, which quantifies how well a set of probability forecasts differentiates between events and non-events:

$$\textbf{Resolution} = \frac{1}{n} \sum_{k=1}^{d} n_k \left(\bar{y}_k - \bar{y}\right)^2 \tag{3.1}$$

where:

$$\bar{y}_k = \frac{1}{n_k} \sum_{i=1}^{n_k} y_{k,i} \tag{3.2}$$

- n is the total number of forecasts,
- d is the number of discrete probability bins,
- n_k is the number of forecasts in the k-th bin,
- $\bar{y}_k$ is the observed relative frequency for the k-th probability bin,
- $\bar{y}$ is the overall observed relative frequency.

The term $(\bar{y}_k - \bar{y})^2$ captures the variance between individual forecast categories and the overall event frequency. Higher resolution indicates that forecasts better differentiate between events and non-events.
so the resolution tells us how the model change with different situations.
the scores used to evaluate resolution are Brier Score and Reliability.

Discrimination

Discrimination measures how well the forecast separates cases where the event occurs from cases where it does not. In other words, it examines whether the forecast probabilities differ for events that happen versus those that don't. High discrimination: A forecast has high discrimination if, for example, when rain occurs, the forecast consistently predicts a high probability of rain, and when rain doesn't occur, it predicts a low probability. It means the forecast is good at distinguishing between rain and no-rain days. Low discrimination: If the forecast provides similar probabilities regardless of whether it rains or not (e.g., predicting a 60% chance of rain every day), it has poor discrimination because it doesn't effectively differentiate between days with and without rain. The score used to evaluate descrimination is ROC[14].

[14]Relative operating characteristics

Reliability

A forecast is reliable if the predicted probabilities match the actual frequencies. For instance: If you forecast a 40% probability for below-normal rainfall, below-normal rainfall should occur in 40% of the cases where you make that prediction. Similarly, if you forecast a 25% chance of above-normal rainfall, above-normal rainfall should happen 25% of the time when you give that probability. If this relationship holds consistently over many forecasts, the forecasts are well-calibrated (or reliable). A Reliable but Uninformative Forecast A forecast that always gives the climatological probability (e.g., always predicting a 33% chance for each category: below, normal, above normal) would be reliable because the climatological average matches the observed frequencies. However, this forecast wouldn't provide any information about changing conditions from case to case—it doesn't adapt to the current situation, making it uninformative.

$$\textbf{Reliability} = \frac{1}{n} \sum_{k=1}^{d} n_k \left(\bar{p}_k - \bar{y}_k \right)^2 \tag{3.3}$$

- n is the total number of forecasts,
- d is the number of discrete probability bins,
- n_k is the number of forecasts in the k-th bin,
- $\bar{y}_k$ is the observed relative frequency for the k-th probability bin,
- $\bar{p}_k$ is relative frequency for the k-th probability.

Sharpness

Sharp forecasts provide a strong signal about the expected outcome. For example, a sharp forecast might assign a 70% chance to a certain outcome, like above-normal rainfall. This high probability communicates more confidence in that specific outcome. On the other hand, when the forecast probabilities are close to the climatological values (like assigning a 40% chance to above-normal, 35% to normal, and 25% to below-normal), the forecast is not very sharp, meaning the forecaster isn't very confident in predicting any one outcome. The climatological probabilities are reliable, but aren't sharp.

CHAPTER 4

METHODOLOGY

4.1 DATA

The hindcast data used in this study was obtained using the OSOP package[1], a tool developed by the UK Met Office to facilitate the retrieval of climate and meteorological data. The dataset comprises monthly mean seasonal forecasts for temperature over the MENA (Middle East and North Africa) region.

The hindcast data spans the common period 1993–2016 and was downloaded from the Copernicus Climate Change Service (C3S) platform.

The data was retrieved for the following configurations:

- Variable: 2-meter air temperature (t2m).
- Forecast Range: Lead times of interest (1–3 months), it includes DJF[2], JJA[3], MAM[4], SON[5]
- Geographical Area: MENA region.
- Temporal Coverage: 1993–2016
- the used centers are $UKMO, ECMWF, ECCC_2, ECCC_3, CMCC, Meteo-France_8, DWD$

In addition to the hindcast data, this study utilized ERA5 reanalysis data, a state-of-the-art atmospheric reanalysis product produced by the European Centre for Medium-Range Weather Forecasts (ECMWF).

4.2 Deterministic Evaluation Metrics

To evaluate the performance of seasonal climate prediction models, a range of deterministic measures is employed. These metrics offer valuable insights into the accuracy and reliability of forecasts by assessing the alignment between predicted and observed values. In this section, we present the main deterministic measures, which will also be applied in our study.

[1] https://github.com/OSFTools/osop/tree/main/scripts
[2] December,January,February
[3] June,July,August
[4] March,April,May
[5] September,October,November

4.2.1 Deterministic Measures

Deterministic measures are statistical metrics used to assess the accuracy and reliability of climate forecasting models. These measures quantify the degree of correspondence between the predicted (Hindcast) and observed (Observation) values. Below, we present some widely used deterministic measures.

4.2.2 Anomaly Correlation Coefficient (ACC)

The Anomaly Correlation Coefficient (ACC) evaluates the skill of seasonal forecasts by comparing anomalies (deviations from climatology) between hindcasts and observations. It considers both the direction and magnitude of the anomalies, making it a robust metric for assessing forecast accuracy in climate studies.

$$ACC = \frac{\sum_{i=1}^{n}(H_i - \bar{H})(O_i - \bar{O})}{\sqrt{\sum_{i=1}^{n}(H_i - \bar{H})^2 \cdot \sum_{i=1}^{n}(O_i - \bar{O})^2}}$$

where:

- ACC: Anomaly Correlation Coefficient.
- H_i: Forecast anomaly (Hindcast).
- O_i: Observed anomaly (Observation).
- $\bar{H}$: Mean of the forecast anomalies.
- $\bar{O}$: Mean of the observed anomalies.
- n: Number of valid time points.

The interpretation of ACC provides insights into forecast accuracy:

- A higher ACC (closer to 1) indicates that the model successfully captures the direction and magnitude of anomalies, showing a high level of forecast skill.
- An ACC near 0 suggests that the model does not capture the anomalies accurately, indicating poor forecast skill.
- A negative ACC indicates an inverse relationship between the forecast and observed anomalies, suggesting that the model consistently misrepresents the direction of deviations.

4.2.3 Root Mean Square Error (RMSE)

The RMSE measures the average magnitude of error between the hindcast and observed values. It is a robust metric to assess the overall predictive accuracy of a model.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(H_i - O_i)^2}$$

where:

- H: The Hindcast.

- O: The Observation.
- i: The valid time (index of the data point).

The interpretation of RMSE provides insights into the model's accuracy:

- A lower RMSE indicates higher accuracy, meaning the hindcast values are closer to the observed values.
- A higher RMSE suggests larger discrepancies between predictions and observations, indicating lower model accuracy.
- RMSE is useful for comparing the performance of different models or forecasting methods.

4.2.4 Coefficient of Determination (R^2)

The coefficient of determination, R^2, evaluates the goodness of fit of a model. It measures the proportion of variance in the observed data that is explained by the model's predictions. An R^2 value closer to 1 indicates better predictive performance, while values near 0 suggest a weak relationship.

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(O_i - H_i)^2}{\sum_{i=1}^{n}(O_i - \bar{O})^2}$$

where:

- R^2: Coefficient of determination.
- H_i: Predicted value (Hindcast).
- O_i: Observed value (Observation).
- $\bar{O}$: Mean of the observed values.
- $\sum_{i=1}^{n}(O_i - H_i)^2$: Residual sum of squares (unexplained variance).
- $\sum_{i=1}^{n}(O_i - \bar{O})^2$: Total sum of squares (total variance).

The interpretation of R^2 provides insights into the model's performance:

- A higher R^2 (closer to 1) indicates that a large proportion of the variance in the observations is explained by the model, signifying better predictive accuracy.
- A lower R^2 (closer to 0) suggests that the model fails to explain much of the variance, indicating poor model performance.
- R^2 is a useful metric for comparing different models and understanding how well each model captures the variability in the observed data.

4.2.5 Conclusion on Deterministic Measures

Deterministic measures such as RMSE, Spearman correlation, R^2, and ACC provide essential tools for evaluating the accuracy and reliability of seasonal forecasting models. Together, they address different aspects of model performance, offering a comprehensive framework for improving forecasting methodologies.

4.3 Probabilistic Evaluation Metrics

In addition to deterministic measures, probabilistic metrics play a crucial role in evaluating forecast performance by assessing the ability of models to predict the likelihood of various outcomes. These metrics provide a deeper understanding of how well forecast probabilities align with observed outcomes.

4.3.1 The Brier Score (BS)

The Brier Score (BS)[6] is the mean squared difference between pairs of forecast probabilities p and the binary observations y. N is the total number of forecasts. It measures the total probability error, considering that the observation is 1 if the event occurs, and 0 if the event does not occur (dichotomous events).

$$BS_j = \frac{1}{N}\sum_{i=1}^{N}(y_{j,i} - p_{j,i})^2$$

where:

- N: The number of forecasts.
- $y_{j,i}$: 1 if the i^{th} observation was in category j, and 0 otherwise.
- $p_{j,i}$: The i^{th} forecast probability for category j.

The interpretation of the Brier Score (BS) provides insights into forecast accuracy:

- A BS closer to 0 indicates better forecast accuracy, meaning the predicted probabilities are closer to the actual binary outcomes (observations).
- A BS closer to 1 indicates poorer forecast accuracy, meaning the forecast probabilities significantly deviate from the actual outcomes.
- A perfect forecast (where the predicted probability matches the observed outcome perfectly) receives a BS of 0.
- Higher BS values indicate less accurate forecasts, with errors increasing as the score moves toward 1.

4.3.2 Reliability

The reliability[7] measures the degree of correspondence between the forecast probability and the observed frequency for an event or outcome that is being predicted. It summarizes the conditional bias of the forecasts for a given event and is equal to the weighted average of squared differences between the forecast and conditional observed probabilities. If the reliability is 0, the forecast is perfectly reliable. To observe the frequency distribution, the forecast probability, from 0 to 1, is divided into 5 bins (0.1, 0.3, 0.5, 0.7, 0.9) to compare to the observed frequency in each of the same bins in this study.

[6] wmo guidance verification
[7] wmo guidance verification

$$Reliability = \frac{1}{n}\sum_{k=1}^{d} n_k(\bar{p_k} - \bar{y_k})^2$$

where:

- n_k: The number of forecasts for the k^{th} probability value ($\bar{p_k}$).
- $\bar{y_k}$: The observed relative frequency for that value.

The interpretation of reliability provides insights into the forecast's performance:

- A reliability score closer to 0 indicates that the forecast probability matches the observed frequency well, meaning the forecast is highly reliable.
- A higher reliability score suggests a discrepancy between forecast probabilities and observed outcomes, indicating a less reliable forecast.
- Perfect reliability (where forecast probability exactly matches observed frequency) results in a score of 0.
- Reliability is useful for evaluating whether forecast probabilities are over- or under-estimating the actual occurrence of events.

4.3.3 The ranked probability score (RPS)

The Ranked Probability Score (RPS) is a performance metric used in probabilistic forecasting to assess how well the predicted probability distribution matches the observed outcome distribution. It is particularly useful when there are multiple categories (e.g., terciles such as lower, middle, and upper) and is commonly applied in fields such as meteorology, climatology, and economics.

$$RPS = \frac{1}{n(m-1)}\sum_{i=1}^{n}\sum_{k=1}^{m-1}\left(\sum_{j=1}^{k}(y_{j,i} - p_{j,i})\right)^2$$

where:

- n: The number of forecasts.
- m: The number of categories.
- $y_{j,i}$: 1 if the i^{th} observation was in category j, and 0 otherwise.
- $p_{j,i}$: The i^{th} forecast probability for category j.

The interpretation of the Ranked Probability Score (RPS) is as follows:

- A lower RPS indicates better forecast performance, where the predicted probabilities closely match the observed categories.
- A perfect forecast, where the probability distribution exactly matches the observed outcome, results in an RPS of 0.
- A higher RPS indicates a poor forecast, where the predicted probabilities significantly deviate from the actual observations.

- The score ranges from 0% (perfect forecast) to 100% (completely incorrect forecast), with the latter occurring when all observations are in the outermost categories and the forecasts are maximally incorrect.

4.3.4 Relative Operating Characteristics

The ROC[8] can be used in forecast verification to measure ***the ability of the forecasts to distinguish an event from a non-event***. For seasonal forecasts with three or more categories, the first problem is to define the "event". One of the categories must be selected as the current category of interest, and an occurrence of this category is known as an event. An observation in any of the other categories is defined as a non-event and no distinction is made as to which of these two categories does occur. So, for example, if below normal is selected as the event, normal and above normal are treated equally as non-events.

The score indicates the probability of successfully discriminating below-normal observations from normal and above-normal observations. It indicates how often the forecast probability for below normal is higher when below normal actually does occur compared to when either normal or above normal occurs.

The interpretation of the ROC is as follows:

- A higher ROC score indicates better discrimination ability, meaning the forecast is more successful in distinguishing between an event (e.g., below normal) and a non-event (e.g., normal or above normal).
- A perfect forecast that can always correctly distinguish between events and non-events would have an ROC of 1, indicating perfect discrimination.
- A ROC score of 0.5 indicates a random forecast, where the forecast does not provide any more useful information than random guessing.
- Lower ROC values indicate that the forecast performs poorly at distinguishing between events and non-events.

4.3.5 Relative Operating Characteristics Skill Score

The Relative Operating Characteristic Skill Score (ROCSS) is a measure used in forecast verification to assess the ability of probabilistic forecasts to discriminate between events and non-events. It builds on the Relative Operating Characteristic (ROC) curve, which plots the hit rate (true positive rate) against the false alarm rate (false positive rate) at various forecast probability thresholds.

- The ROC curve evaluates the discrimination capability of a forecast, i.e., how well the forecast can separate occurrences of an event (e.g., below-normal temperature) from non-events (e.g., normal or above-normal temperature).
- The ROC Skill Score quantifies the area under the ROC curve (AUC) and compares it to a no-skill forecast.

$$ROCSS = \frac{AUC - AUC_{\text{no-skill}}}{1 - AUC_{\text{no-skill}}}$$

where:

[8] wmo guidance verification

- AUC: Area Under the ROC Curve for the forecast being evaluated.
- $AUC_{\text{no-skill}}$: Area Under the Curve for a no-skill forecast, typically 0.5.

Interpretation of ROCSS:

- 1: Perfect discrimination ability, where the forecast can perfectly distinguish between events and non-events.
- 0: No skill, meaning the forecast performs no better than random guessing.
- Negative values: The forecast performs worse than random guessing, indicating a forecast that is worse than a no-skill forecast.

4.3.6 Summary of Probabilistic Forecast Metrics

Probabilistic metrics are essential in evaluating seasonal forecasts, as they provide insights into various aspects of model performance. The main properties typically assessed are *Reliability*, *Discrimination*, *Sharpness*, and *Resolution*. These aspects help to understand the quality of probabilistic forecasts and guide improvements in forecasting models.

- **Reliability**: This metric evaluates how closely the predicted probabilities match the observed frequencies. For example, if a model predicts a 70% probability of an event, reliability measures whether that event occurs approximately 70% of the time. A perfectly reliable model would show a diagonal line on the reliability diagram, meaning predicted probabilities correspond exactly to the observed frequencies.
- **Discrimination**: Discrimination examines the model's ability to distinguish between different outcomes, such as events (e.g., below-normal temperatures) and non-events (e.g., normal or above-normal temperatures). It assesses whether the model can correctly classify the conditions. The Area Under the Curve (AUC) of the ROC curve is typically used to evaluate discrimination.
- **Sharpness**: Sharpness concerns the model's confidence in its predictions, particularly how close the forecast probabilities are to the extremes (0 or 1). A forecast with high sharpness indicates a model that makes bold predictions, while low sharpness suggests that the model gives more uncertain forecasts, close to the middle (e.g., 0.5).
- **Resolution**: This measures how well the model can distinguish between different forecast categories. A model with high resolution provides more specific information about the forecast and captures subtle variations. Brier score decomposition is a common method for assessing resolution.

CHAPTER 5

RESULTS

This chapter presents the results of our study, divided into two main sections: temperatures and precipitation. In each section, we provide a comprehensive analysis of the deterministic and probabilistic evaluation of forecast performance across the MENA region. By examining both temperature and precipitation metrics, we aim to highlight the strengths and limitations of the forecasting models, offering valuable insights into their reliability and applicability in this climatically diverse area.

5.1 Temperature

In the temperature session, the use of heatmaps and temperature metrics maps will allow for a visual interpretation of model performance across various metrics. By analyzing these heatmaps, we can identify the most effective models based on metrics such as Anomaly Correlation (ACC), RMSE (Root Mean Square Error), and the Coefficient of Determination (R-squared). These visualizations will help in understanding the relationships between predicted and observed temperatures, aiding in the selection of models with the highest predictive accuracy. Additionally, the use of probabilistic evaluation metrics such as the Brier Score, Reliability, Ranked Probability Score (RPS), and Relative Operating Characteristics (ROC) will provide insights into the model's performance in terms of forecast quality, focusing on calibration, discrimination, and sharpness. These methods will be used to refine and improve predictive models for temperature forecasting in the MENA region.

5.1.1 Deterministic evaluation results

Anomaly Correlation Coefficient

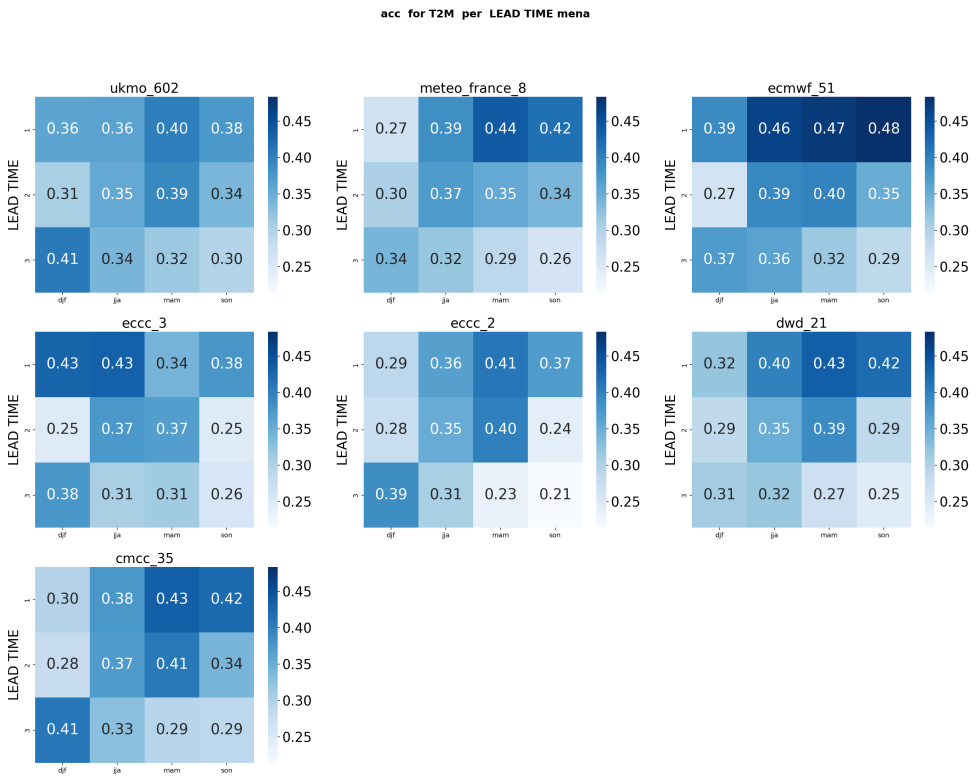

Figure 5.1: The Heatmap of acc for the mena region for every period ***(1 for perfect ACC)***

The acc is moderate for all centers; however, the best models are ***ECMWF, meteo-france and eccc-3***. There is no clear variability in performance over time. For SON and JJA, performance is good at time 1 for all centers but decreases with increasing lead-time. Then, for DJF, ***UKMO,ECCC-3 and ECMWF*** are the best with values between 0.39 and 0.43 for the first led-time and a little decrease along lead-time, the ***CMCC-35*** show good acc with a value of 0.41 for the 3d lead-time. Moreover, for MAM and SON, we have good acc for the 1st lead-time, especially ***ECMWF, METEO-FRANCE, DWD and CMCC-35***, with values between 0.42 and 0.48, but this performance decrease greatly along lead-time to reach values around 0.2. The JJA show good and stable ACC for all centers and lead-times with values between 0.43 and 0.31.

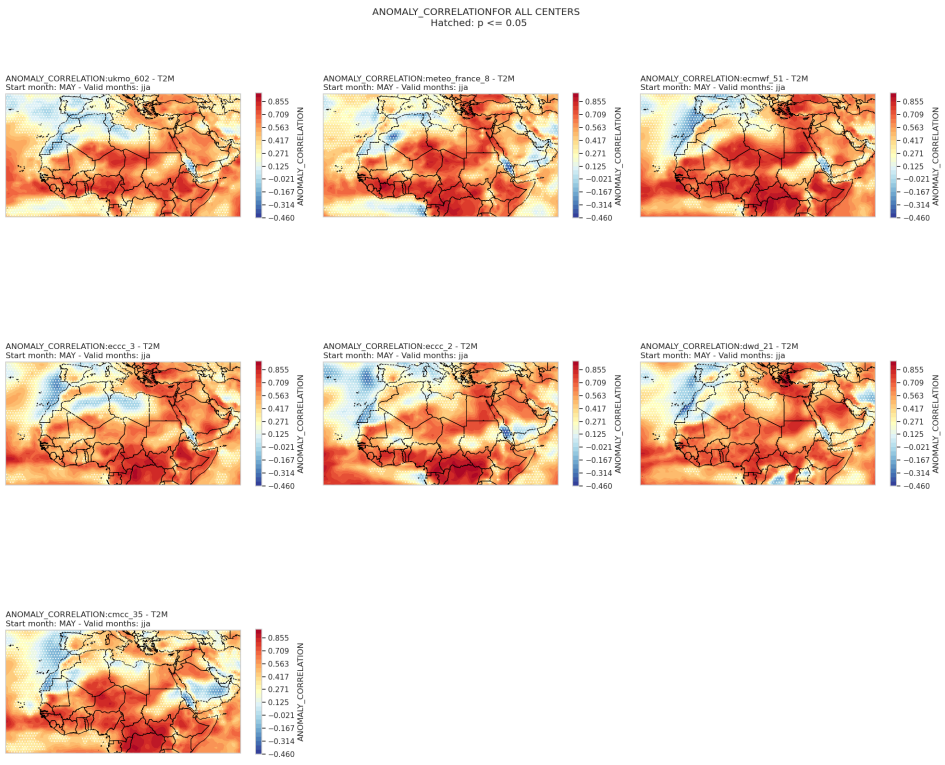

Figure 5.2: The 3 month rolling mean of JJA for ACC for the mena region ***(1 for perfect ACC)***

the ACC demonstrates strong performance across nearly the entire region, indicating that the model captures the spatial patterns of anomalies effectively. This suggests a high level of agreement between observed and predicted anomalies, reflecting the model's reliability in representing anomaly patterns. However, despite this strong performance in terms of anomaly correlation, the significance of these correlations is not consistent across all regions. Specifically, in the center of Africa and the Arabian Peninsula, the correlations are not statistically significant. This implies that while the model captures the general patterns of anomalies in these areas, the strength of the linear relationship between observed and predicted anomalies is weak or uncertain.

Focus on North Africa The heatmap below reveals that the ***ECMWF***, ***UKMO***, and ***ECCC-3*** models demonstrate relatively strong correlations over the North Africa region. This suggests that these models perform well in capturing the anomalies in this specific area.

Figure 5.3: ACC heatmap for the North Africa region across different periods.

Focus on the Arabian Peninsula The heatmap for the Arabian Peninsula indicates strong performance across all forecasting centers, with ***ECMWF***, ***UKMO***, and ***DWD*** exhibiting the highest correlation scores.

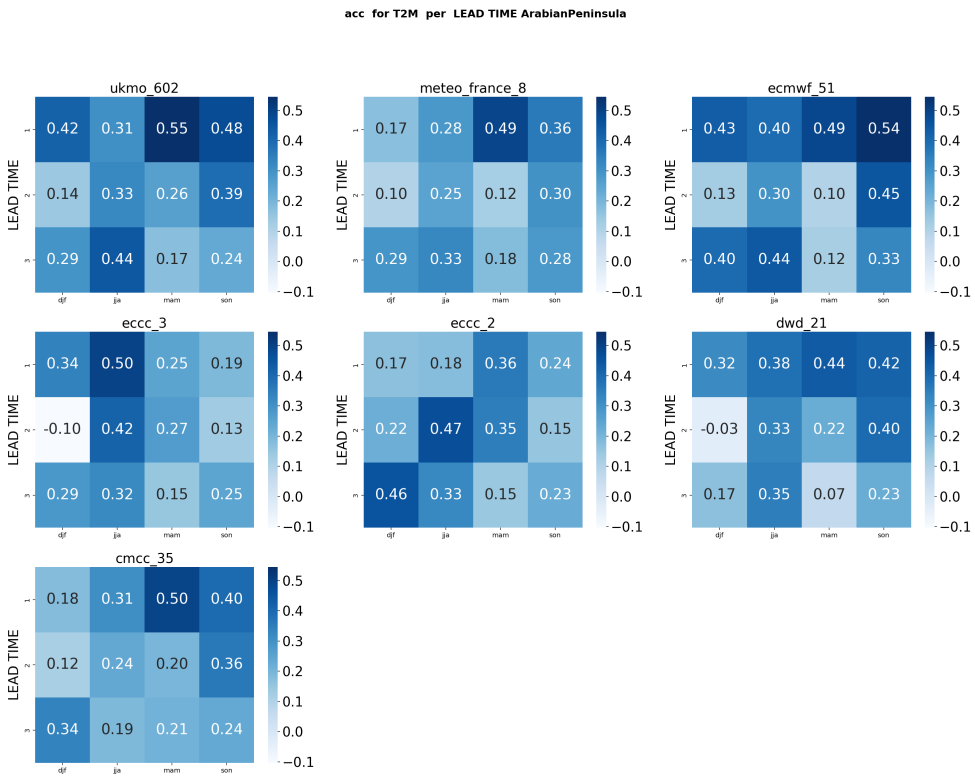

Figure 5.4: ACC heatmap for the Arabian Peninsula across different periods ***(1 indicates perfect correlation)***.

The analysis highlights that the Arabian Peninsula consistently achieves better ACC scores compared to the general MENA region. Notably, the ACC is particularly high for ***SON (September-October-November)*** at the third lead time.

Root Mean Square Error

The maps in this section show the RMSE (Root Mean Square Error) between observed and modeled surface temperatures across the MENA region for the four seasons: JJA, DJF, MAM, and SON. The RMSE, expressed in the same units as temperature, evaluates the accuracy of climate models, with lower values indicating better performance. These maps help identify the strengths and limitations of the models across seasons, contributing to the improvement of climate forecasts for the region.

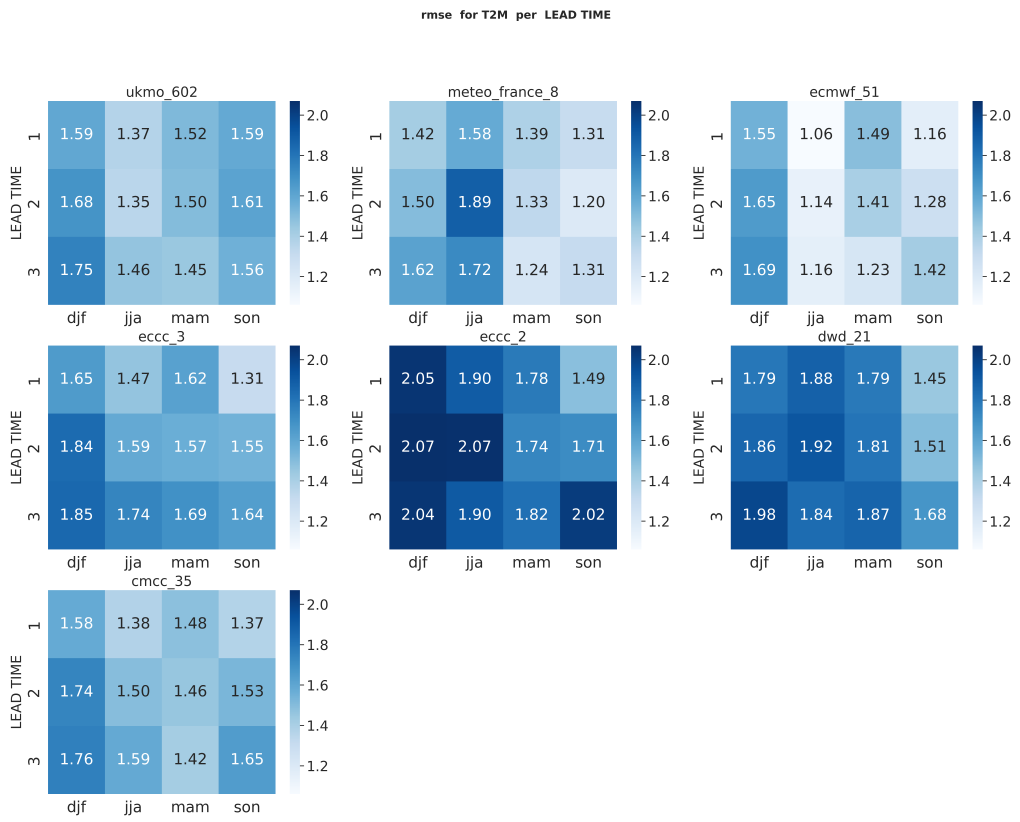

Figure 5.5: Temperature rmse heatmaps for all the sasons ***(0 for perfect RMSE)***

The heatmap for the seven models highlights distinct variations in their seasonal performance. The UKMO model shows moderate to good performance, particularly in JJA and MAM, as reflected by relatively low RMSE values ranging between 1.35°C and 1.51°C across the three lead times. This indicates that the UKMO model is reasonably effective in capturing surface temperature variations during these seasons, likely benefiting from its ability to simulate key atmospheric processes during these periods.

In contrast, Météo-France exhibits weaker performance in JJA, with higher RMSE values suggesting less accurate predictions of surface temperatures during this season. This could be attributed to the model's limitations in capturing summer-specific temperature drivers in the MENA region, such as heatwaves, desert-air interactions, or seasonal atmospheric circulation patterns.

The ECMWF model emerges as the best-performing model based on the heatmap, particularly during JJA, where it demonstrates high predictive accuracy and consistency. This is supported by the RMSE map, which shows significantly lower RMSE values across most of the MENA region. These low RMSE values confirm the ECMWF model's ability to capture regional temperature dynamics effectively, with reduced errors across diverse climatic zones.

Notably, the spatial distribution of RMSE on the map reinforces the ECMWF's strong performance, as it maintains relatively low error values in critical parts of the MENA region. This suggests that the ECMWF model is better equipped to account for the complex climatic interactions in the region, such as the influence of desert regions, coastal temperature gradients, and

seasonal weather patterns.

Overall, these findings emphasize the importance of selecting climate models based on their seasonal and spatial performance, as they play a critical role in improving the accuracy of temperature forecasts for the MENA region.

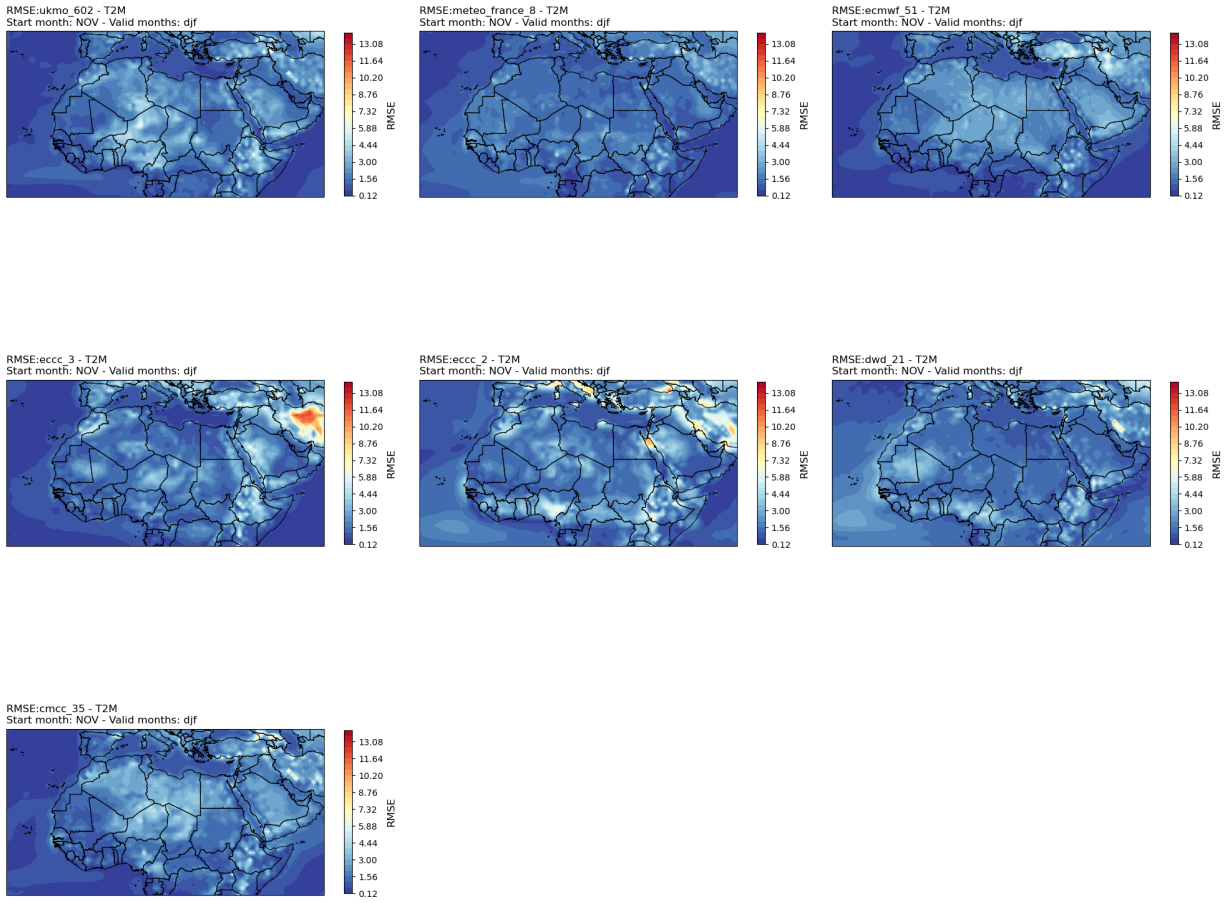

Figure 5.6: The 3 month rolling mean of Temperature rmse for DJF for all centers ***(0 for perfect RMSE)***

The analysis of the RMSE highlights several key patterns. Overall, most regions exhibit low RMSE values, indicating a good agreement between the hindcast and observations and suggesting strong model performance across a large portion of the domain. Central and western Africa consistently show low RMSE across all models, reflecting robust forecasting accuracy in these regions. However, variations are evident in northern Africa, particularly in the Sahara, where some areas experience moderate discrepancies depending on the model. The Arabian Peninsula also stands out with localized regions of higher RMSE, especially in its northern parts, suggesting challenges

in these areas. Among the models, the UKMO, Météo-France, and ECMWF generally perform well, with ECMWF demonstrating consistent low RMSE across the entire domain. In contrast, the ECCC-3 and ECCC-2 models exhibit higher RMSE in specific areas, such as northern Africa and the Arabian Peninsula, while the DWD model performs comparably to the best models in most regions.

focus on North Africa :

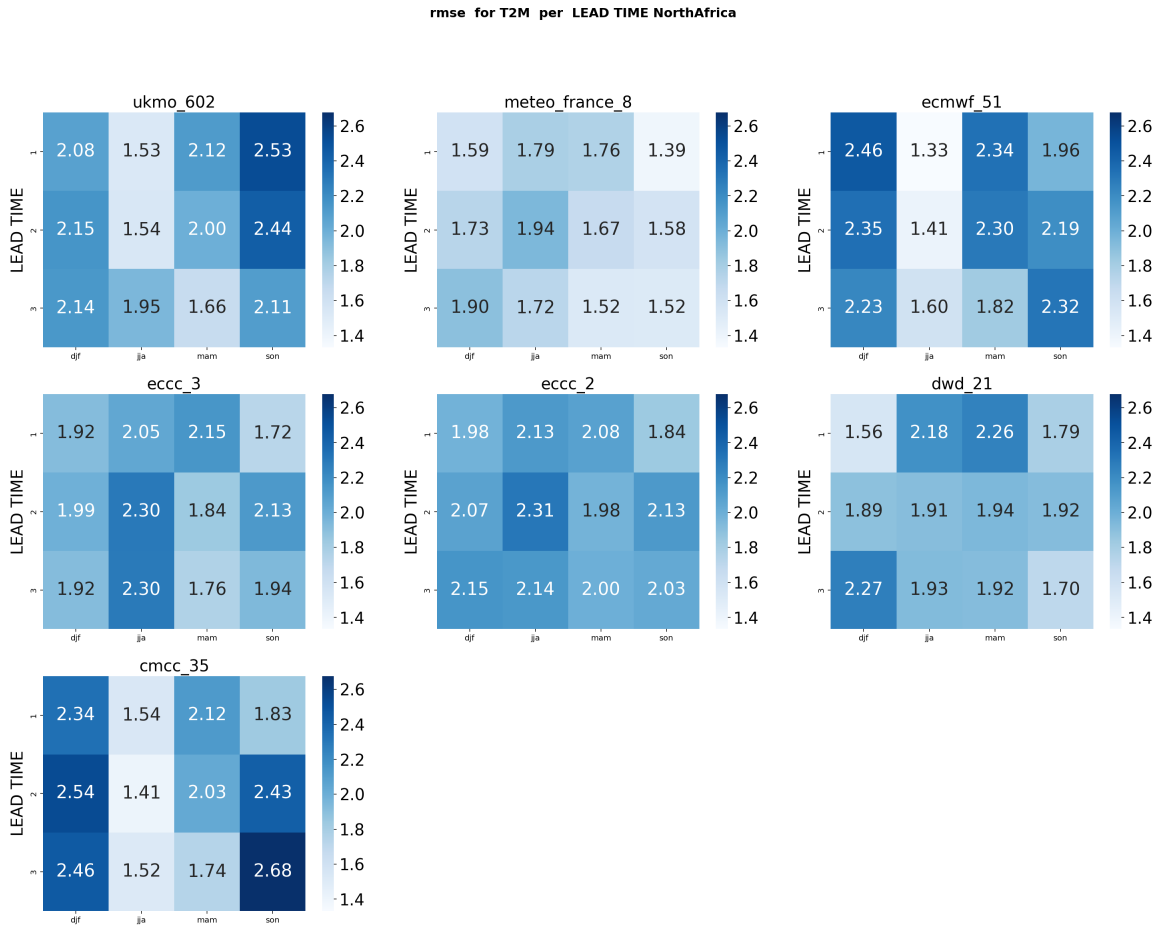

Figure 5.7: heatmap of RMSE For T2M (North Africa)

The North African climate poses challenges for modeling extreme temperatures and spatial variability. Heatmap analysis shows that ***ECMWF*** excels in JJA with RMSE values of 1.34°C–1.58°C but performs lower in other seasons. ***Météo-France*** delivers consistent accuracy, especially in SON, making it more reliable for multi-seasonal forecasting in this region.

focus on Arabian Peninsula :

Figure 5.8: heatmap of RMSE For T2M (Arabian Peninsula)

In the same way as North Africa, the RMSE for the Arabian Peninsula is significantly lower for ***Météo-France***, indicating superior performance.

coefficient of determination

The maps in this section show the Rsquared between observed and modeled surface temperatures across the MENA region for the four seasons: JJA, DJF, MAM, and SON. R-squared is a statistical measure that indicates how well the model explains the variability in observed data, with values closer to 1 signifying better performance. These maps provide valuable insights into the predictive skill of the climate models, highlighting their ability to capture seasonal temperature patterns.

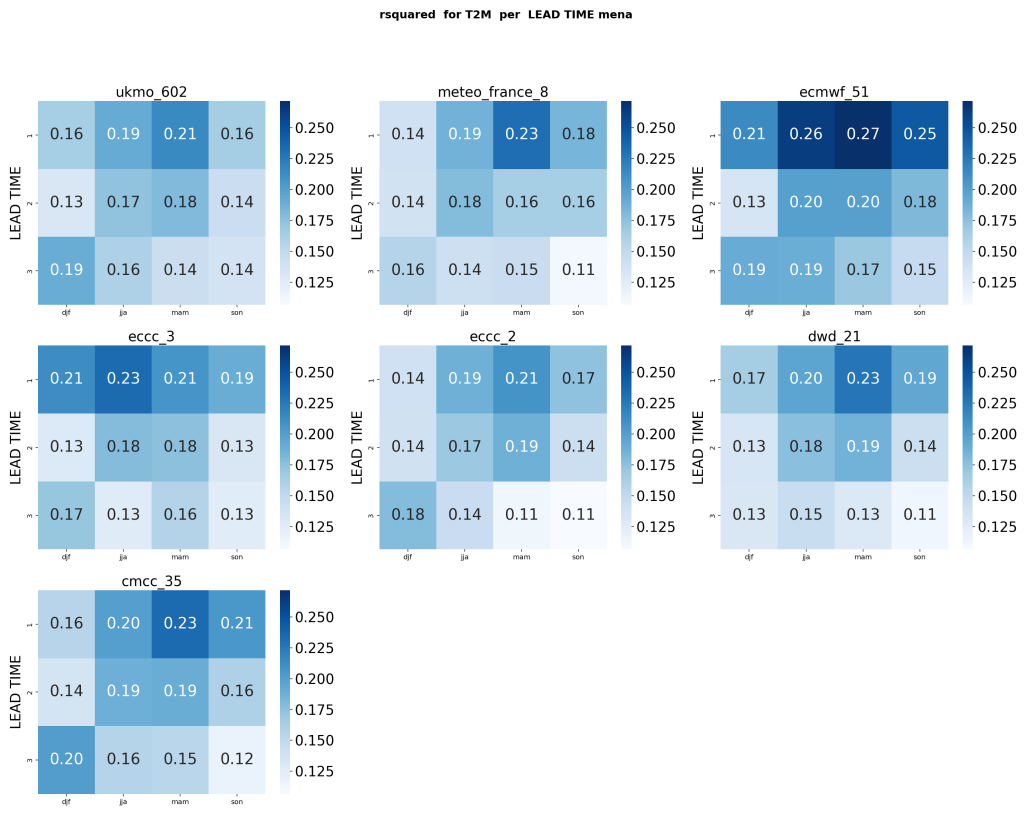

Figure 5.9: Temperature rsquared heatmaps for all the seasons ***(1 for perfect R-SQUARED)***

Based on this deterministic metric (R-squared), the ECMWF model demonstrates superior performance for lead time 1 across all four seasons, particularly during MAM. In general, the portion of variance explained by the model decreases as the lead time increases. This indicates that while the model is highly effective at capturing seasonal variability of surface temperatures in the short term, its predictive skill diminishes over longer time horizons.

The strong performance during MAM highlights the ECMWF model's ability to capture the complexities of spring, a season marked by transitional weather patterns in the MENA region. The high R-squared values during this period suggest that the model accurately reflects observed temperature variability by effectively simulating key drivers such as the gradual warming trend, atmospheric circulation changes, and the interaction between desert and coastal dynamics.

Such precision underscores the ECMWF model's reliability for short-term seasonal forecasting, particularly during periods of heightened climatic variability like MAM. However, the decreasing performance with increasing lead times suggests the need for careful interpretation of forecasts beyond lead time 1, as uncertainty increases with longer projections.

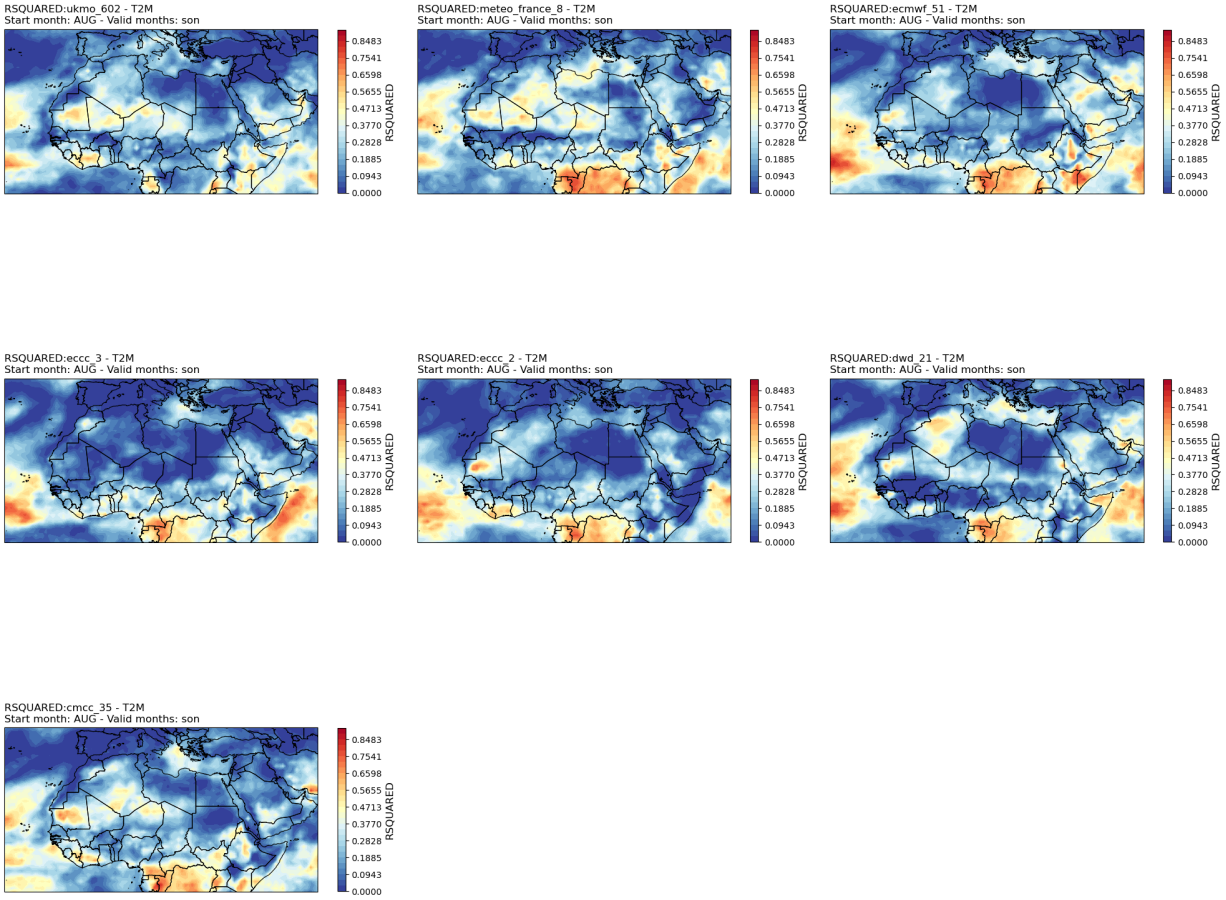

Figure 5.10: 3-months rolling mean of 2-meter temperature of SON R-SQUARED for all centers ***(1 for perfect R-SQUARED)***

From the figure above, the R-SQUARED is excellent at the equator, reaching a maximum value of 0.78. All the centers exhibit good R-SQUARED performance in this region; nevertheless, the ECMWF model demonstrates slightly better performance overall. Across the northern parts of Africa, the R-SQUARED values decrease, indicating a reduced agreement between the hindcast and observations. However, some models, such as Météo-France and ECCC-3, maintain moderate performance in certain localized areas. The Arabian Peninsula exhibits generally low R-SQUARED values across all models, signifying challenges in capturing seasonal variability in this region. The southern part of the domain, particularly over regions near Angola and Zambia, shows moderate R-SQUARED values in some models, such as UKMO and CMCC.

Thus, while the equatorial region consistently shows excellent performance across all models, northern Africa and the Arabian Peninsula represent areas with lower R-SQUARED values, highlighting potential limitations in the seasonal forecasting system. Among the models, ECMWF

appears to have a slight advantage in terms of consistency and accuracy.

focus on North Africa Focusing on North Africa, ***ECMWF*** maintains its position as the most reliable center, consistent with its performance across the broader MENA region.

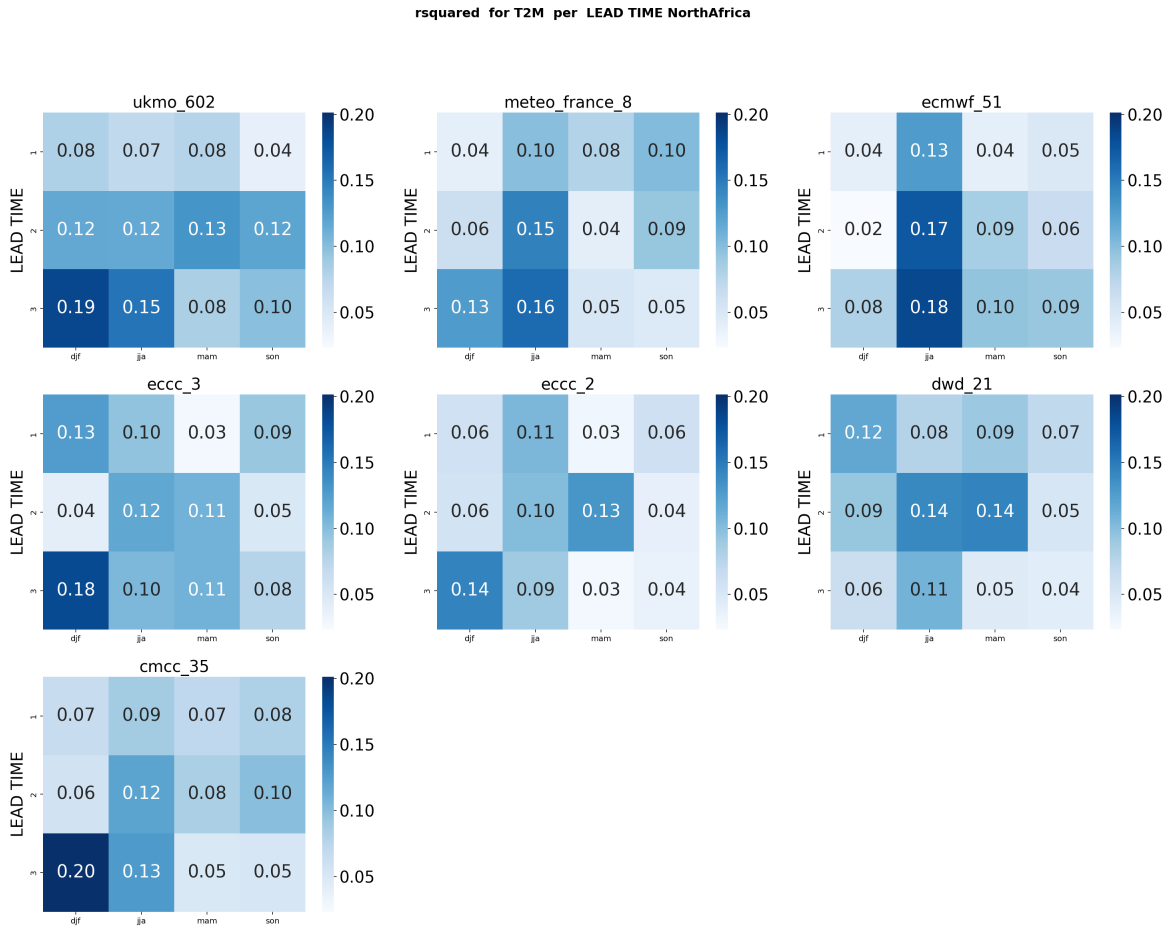

Figure 5.11: Heatmap of T2M RSQUARED in North Africa Region for all centers

focus on Arabian Peninsula :

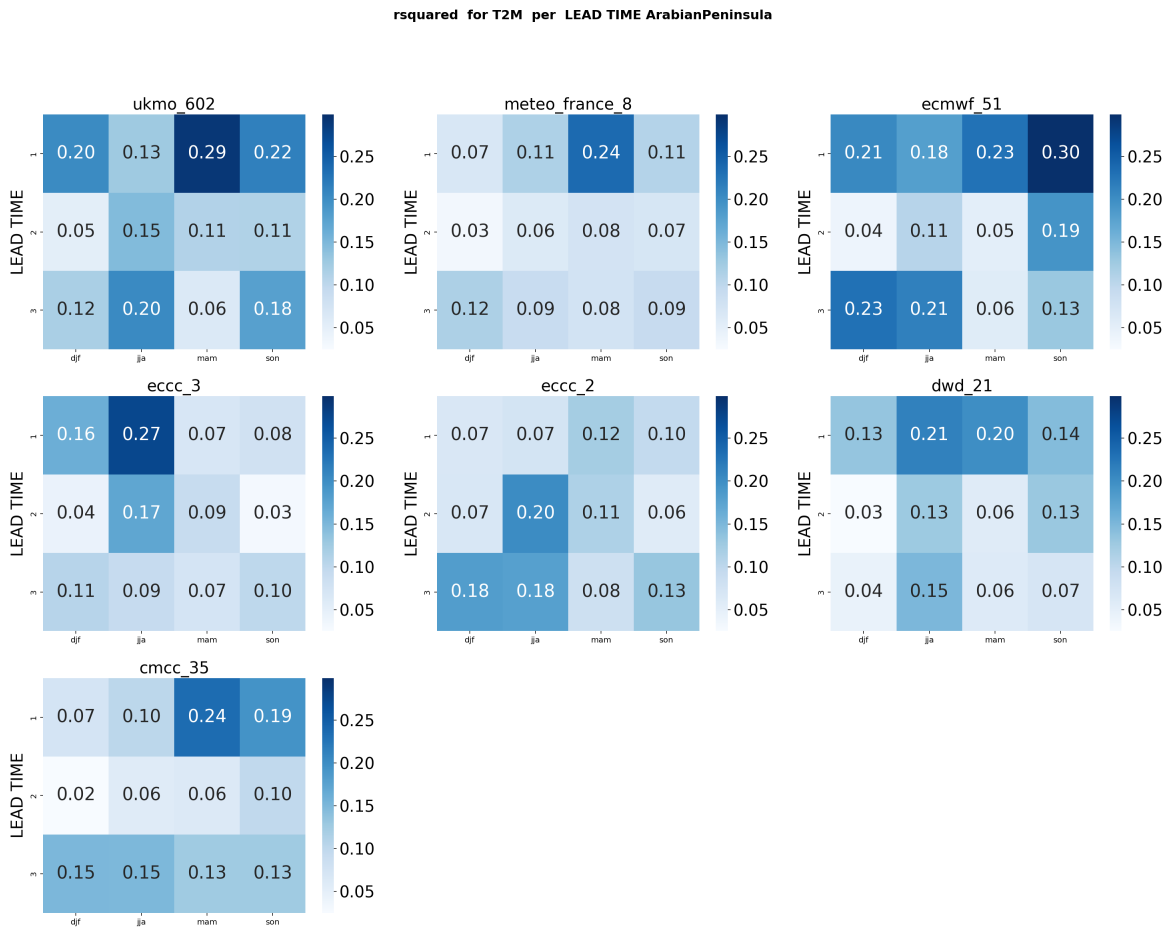

Figure 5.12: Heatmap of T2M RSQUARED in MENA Region for all centers Arabian Peninsula

the R-SQUARED for the Arabian Peninsula shows a little improvement.

5.1.2 Probabilistic evaluation results

To complement the deterministic evaluation of model performance, probabilistic evaluation metrics are employed to assess the reliability and skill of climate models in predicting the likelihood of specific outcomes. Unlike deterministic metrics, which focus on the accuracy of single-point predictions, probabilistic metrics evaluate the quality of the models' forecast distributions, accounting for uncertainty and variability in predictions. These metrics are essential for understanding how well models represent the range of possible outcomes, particularly in regions like MENA, where climatic variability and extremes are prominent. By incorporating probabilistic metrics, this analysis provides a more comprehensive evaluation of the models' predictive capabilities and their usefulness in decision-making under uncertainty. The figures illustrate two main approaches to probabilistic assessment metrics, including the Brier Score (BS) and others. The first approach averages across lead times and grid points, while preserving categories, where the final figure contains the value of

the metric for each season across all four seasons, for each category (mean, lower, upper) defined by the 1/3 quartiles. This method provides insight into the predictive ability of models under different seasonal conditions and forecast probability categories, particularly how well models capture temperature variations in the middle, lower, and upper quartiles of the predicted probability distribution. The diversity of metric values across these categories helps highlight the sensitivity of models to different levels of forecast confidence. It indicates their ability to differentiate between forecast uncertainty and actual observed outcomes, providing a nuanced understanding of how accurately models predict different temperature ranges.

The second approach averages all grid points in the MENA region while retaining all lead times and seasons. This aggregated view provides an overall assessment of the measure for each season, considering all lead times and forecast categories. This approach focuses on how models perform across different forecast scenarios and how well they produce accurate and reliable temperature forecasts, regardless of forecast probability. By retaining lead times and seasons in the analysis, this method provides a comprehensive picture of model performance over time and under different climate conditions. It reveals how well models generalize across different forecast scenarios, helping to identify which models are most effective at producing consistent and reliable forecasts.

Brier score

The figure at the bottom illustrates that most models demonstrate relatively high performance, as indicated by a small Brier Score (BS) around 0.2, meaning that the predicted probabilities are close to the observed ones. This reflects accurate forecast probabilities for T2M. Notably, the middle category presents lower performance relative to the other two categories (lower and upper). This indicates that while some models capture temperature variability well in extreme conditions (upper category), their skill may diminishes when forecasting moderate changes (middle category). This discrepancy highlights the challenges models face in translating predicted probabilities into reliable forecasts, particularly for temperature variations that are neither extreme nor outlier events.

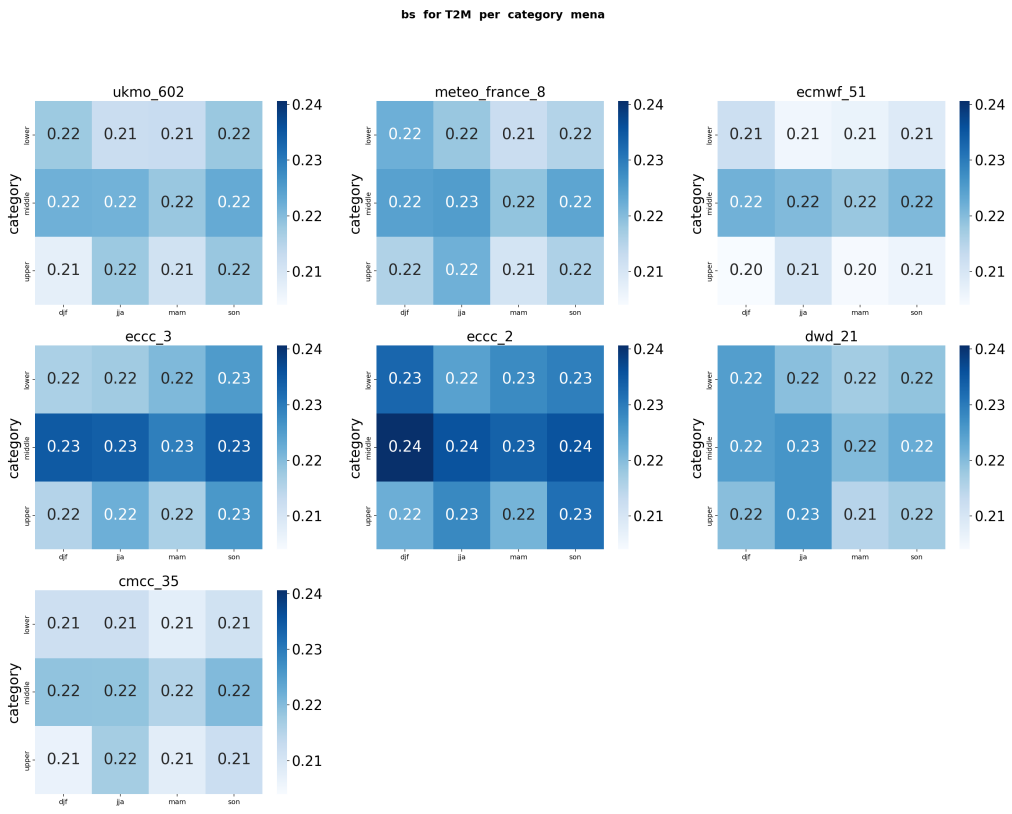

Figure 5.13: Temperature Brier score heatmaps for all the seasons per categories ***(0 represents perfect BS)***

An analysis by lead time revealed that the models ***Meteo France, ECMWF,UKMO and CMCC-35*** exhibit superior performance, as indicated by lower Brier Scores (BS) between 0.21 and 0.23. This suggests that these models provide more accurate probabilistic forecasts for T2M compared to others. Moreover, the differences in BS values between successive lead times are minimal, indicating that the predictive skill of these models remains relatively consistent as the forecast horizon increases.

This stability in performance across lead times is particularly noteworthy, as it reflects the robustness of these models in maintaining their ability to produce reliable forecasts over time. The lower BS values also suggest that these models effectively capture the relationship between predicted probabilities and observed outcomes, ensuring high confidence in their probabilistic predictions. Such consistent performance across lead times is crucial for operational forecasting, as it highlights these models' reliability for both short- and medium-term forecasts in the MENA region.

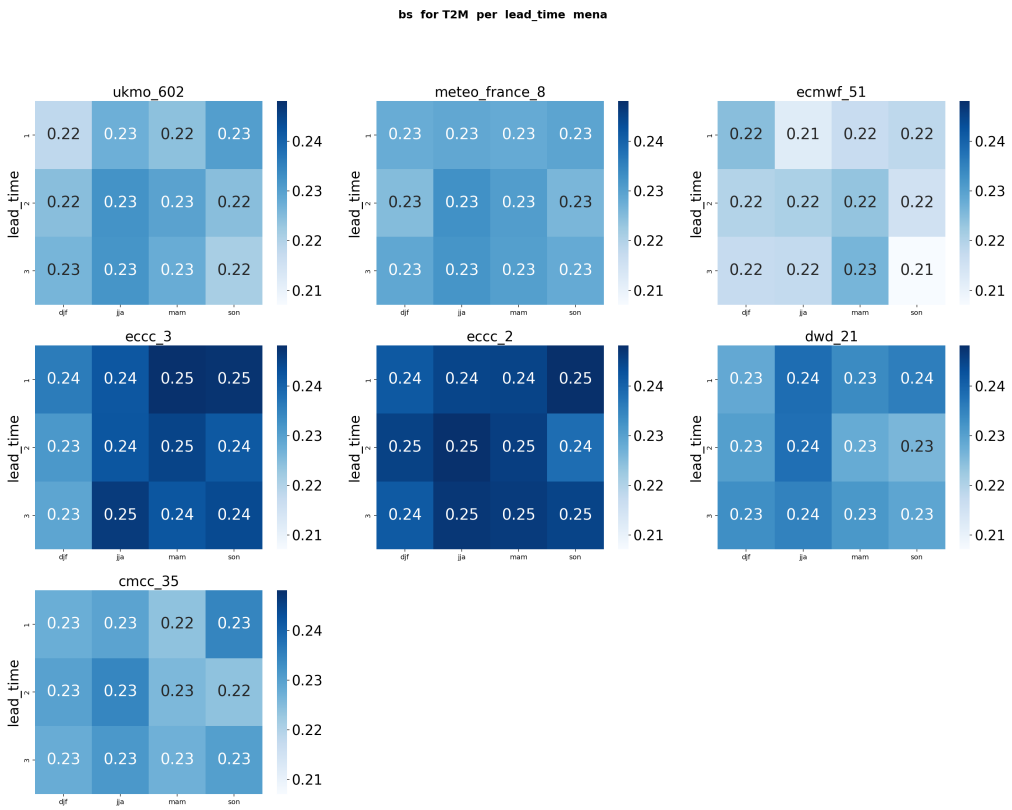

Figure 5.14: Temperature brier score heatmaps for all the seasons per lead time ***(0 represents perfect BS)***

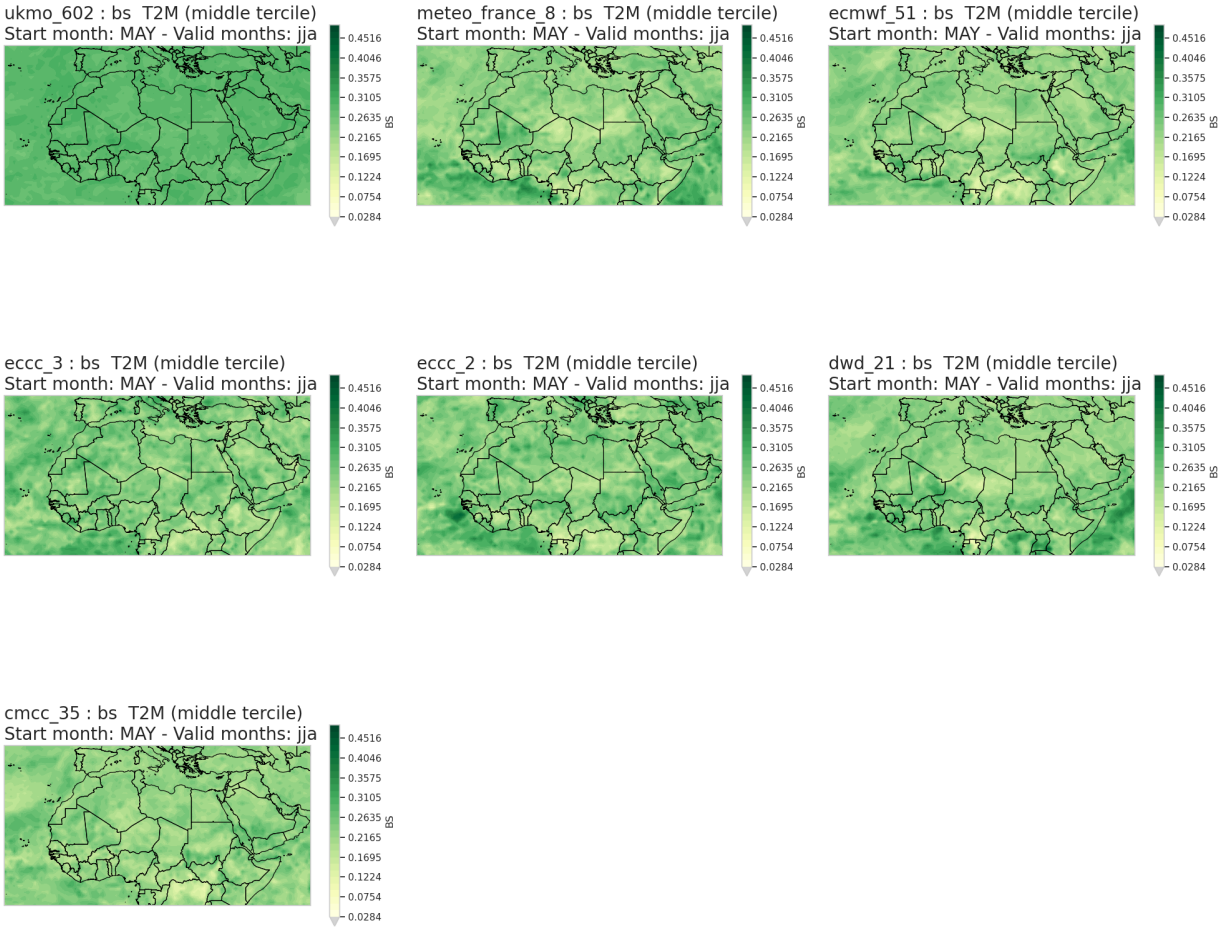

Figure 5.15: 3 months rolling mean for 2-meter-temperature of brier score in the middle tercile JJA. ***(0 represents perfect BS)***

For the ***ECMWF***, the Brier Score reaches an impressive value of 0.02 at the equator, signifying an exceptionally high level of forecast reliability and accuracy in this region. This score underscores the center's strong capability to generate precise probabilistic forecasts, particularly in equatorial zones where weather patterns can be highly dynamic.

While the other centers exhibit a slight decrease in performance compared to the ***ECMWF***, their results still fall within the range of noteworthy accuracy. The relatively modest decline does not overshadow the overall competency demonstrated by the centers, as their scores remain indicative of robust forecasting.

This comprehensive evaluation of Brier Scores across different centers reflects an impressive level of **Resolution**, which refers to the ability of a forecast to distinguish between different outcomes effectively.

focus on North Africa :

To evaluate model performance in North Africa, Brier Score analysis confirms that ***ECMWF, CMCC, and UKMO*** maintain low scores, reflecting consistent reliability across lead times and seasons. These findings indicate that North Africa's unique climatic variability does not significantly impact the predictive skill of these models, affirming their adaptability within the broader MENA region. Minimal score variations across lead times further emphasize their robustness for accurate probabilistic forecasting over varying temporal ranges.

Figure 5.16: Heatmap of T2M brier score for all centers in North Africa region

focus on Arabian Peninsula : there is no big difference in Arabian Peninsula.

Reliability

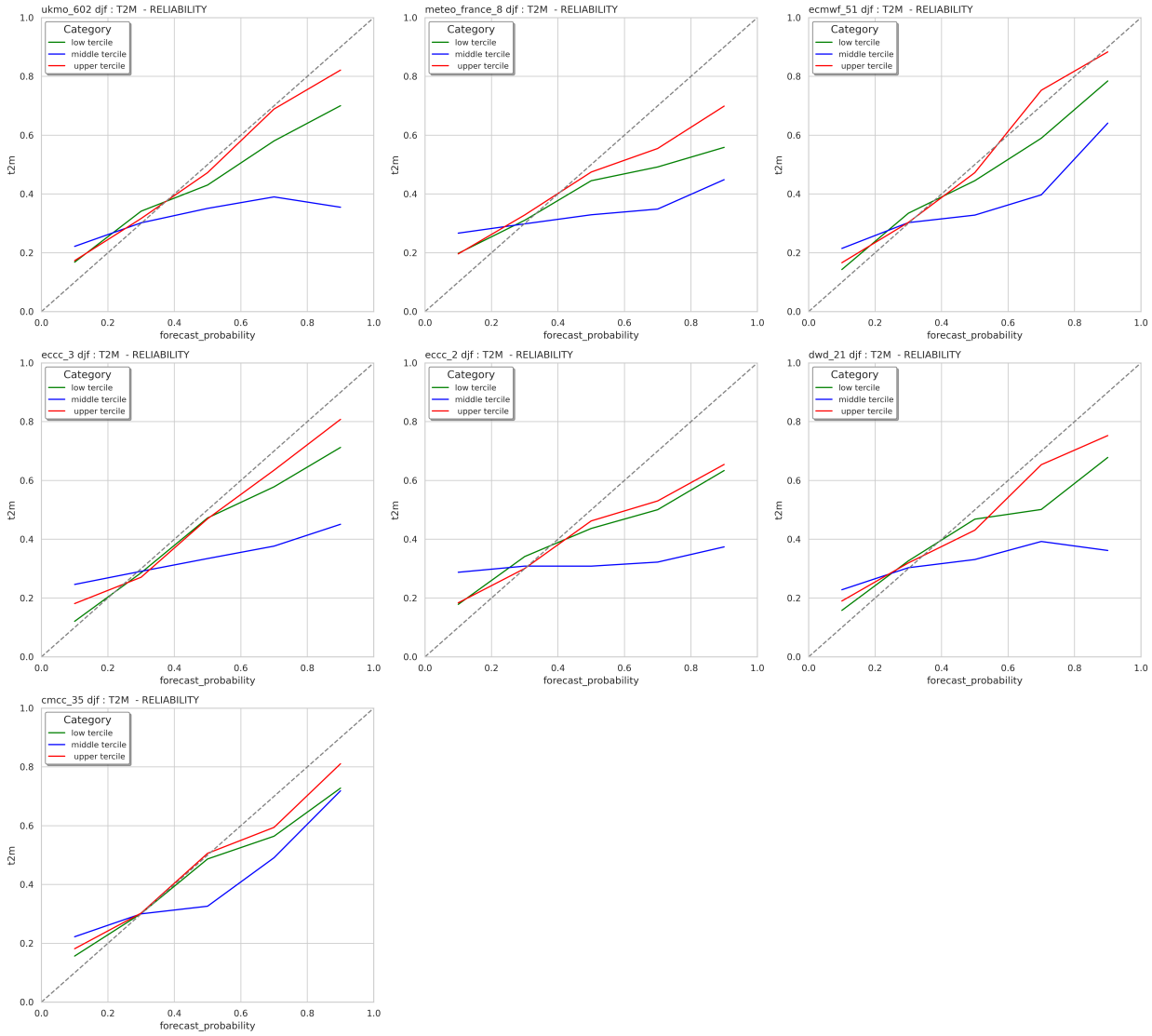

Figure 5.17: temperature reliability diagram DJF ***(45 degree means perfect reliability)***

The reliability diagram reveals strong performance in predicting the upper and lower terciles across all centers, with particularly notable accuracy observed in ***UKMO, ECMWF, and ECCC-3***. However, the middle tercile presents some notable challenges, with predictions being less reliable for all centers. Among them, the ***CMCC-35*** demonstrates comparatively better performance in this middle range, indicating potential advantages in handling predictions for moderate conditions.

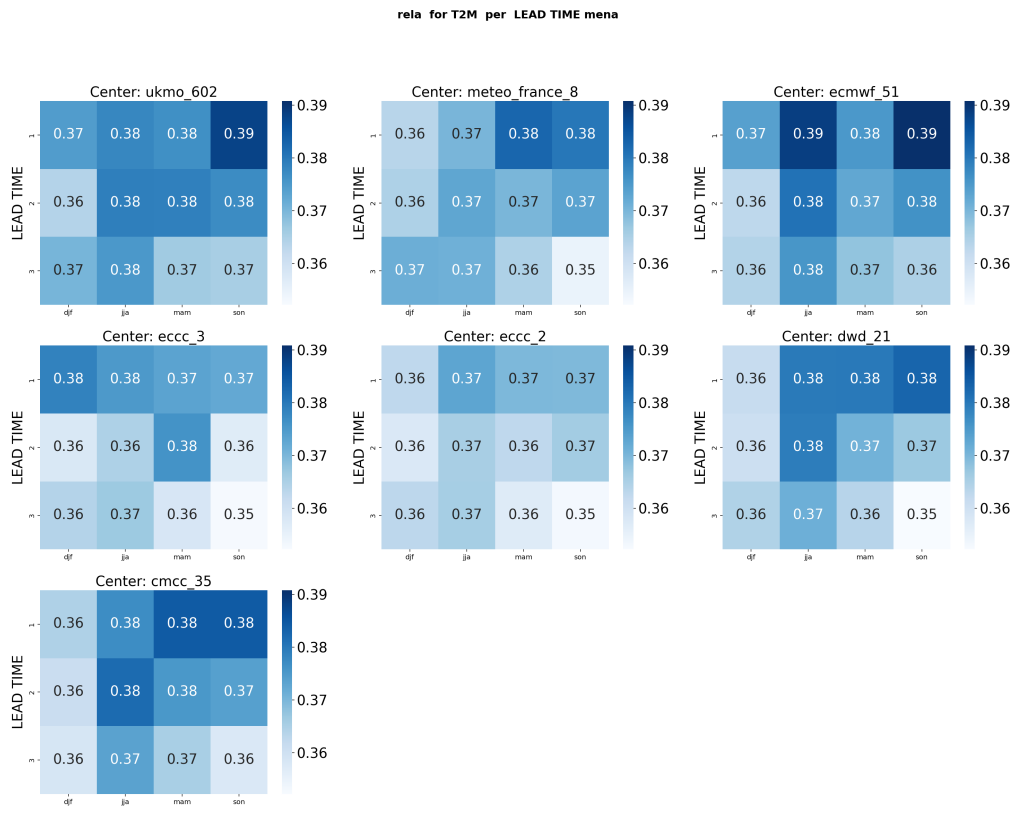

Figure 5.18: temperature reliability heatmap ***(0 means perfect Reliability)***

The heatmap highlights similar trends observed across the models and seasons. Thus, there is no big difference between centers, the score is moderate and the variability along lead-time is very week.

focus on North Africa :

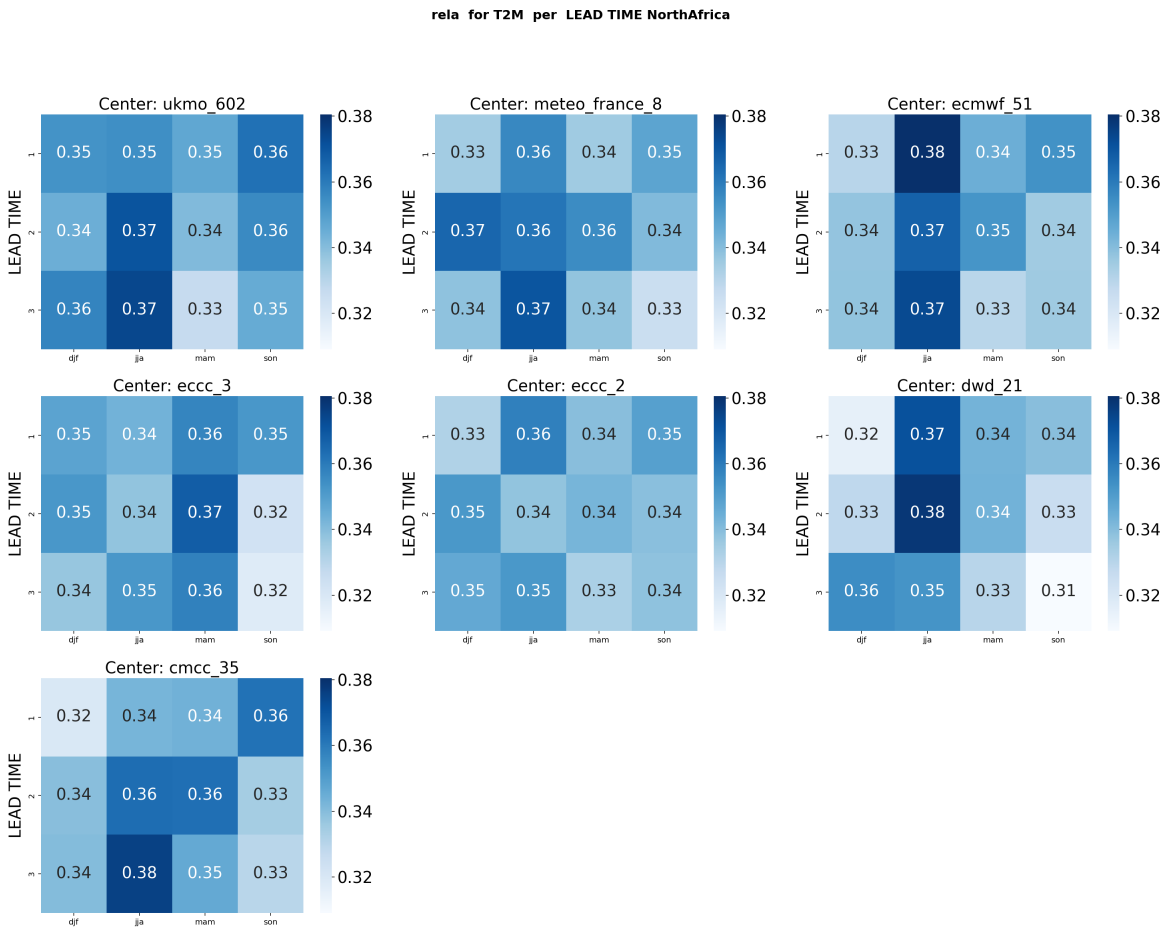

Figure 5.19: Heatmap of T2M reliability for all centers in North Africa

A more focused analysis on North Africa has not significantly altered the overall conclusions derived from the broader MENA region. The consistent performance patterns observed across the different models and seasons remain largely unchanged when examining the North African context.

focus on Arabian Peninsula :

there is no big difference in Arabian Peninsula.

The ranked probability score

The Ranked Probability Score (RPS) provides a valuable measure of forecast performance by evaluating the accuracy of probabilistic predictions across different categories. It combines both the skill in predicting the occurrence of events and the sharpness of the forecast distribution. By comparing the forecasted probability distribution against the observed outcomes, the RPS quantifies the deviation between the predicted and actual probabilities. A lower RPS value indicates better forecast accuracy, reflecting both how well the forecast aligns with observed frequencies and how well it discriminates between different probability categories. This metric helps to identify which models

offer the most reliable probabilistic predictions, particularly in terms of capturing the likelihood of various temperature outcomes within a given forecast period.

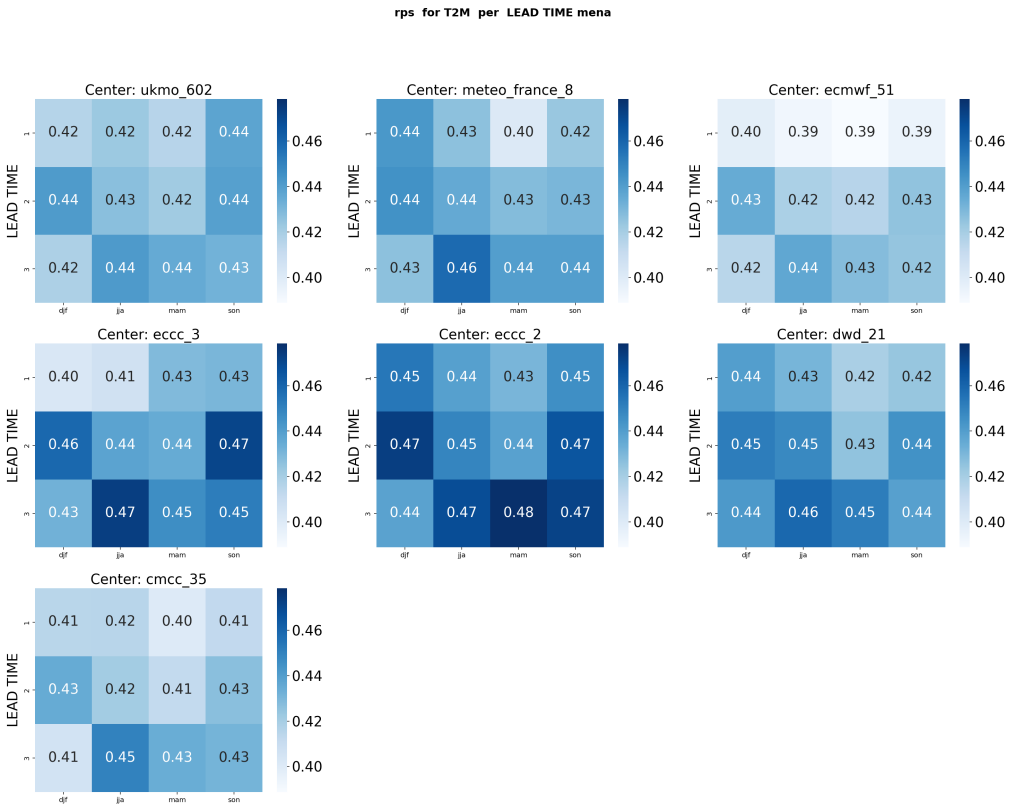

Figure 5.20: Temperature RPS heatmaps for all the seasons per categories ***(0 means perfect RPS)***

The figure displaying the Ranked Probability Score (RPS) for different climate models and seasonal periods provides a detailed view of model performance across various start months (DJF, JJA, MAM, SON). Each cell in the matrix represents the RPS value for a specific model and season combination, with the color intensity indicating how well the forecast probabilities match the observed data.

From this figure, it is evident that ***ECMWF, CMCC-35, METEO-FRANCE and UKMO*** consistently show lower RPS values, indicating better predictive accuracy across different seasons. This suggests that this forecasts are more closely aligned with observed temperature variability. The relatively higher RPS values for ***ECCC*** model underscore their challenges in accurately capturing temperature variations.

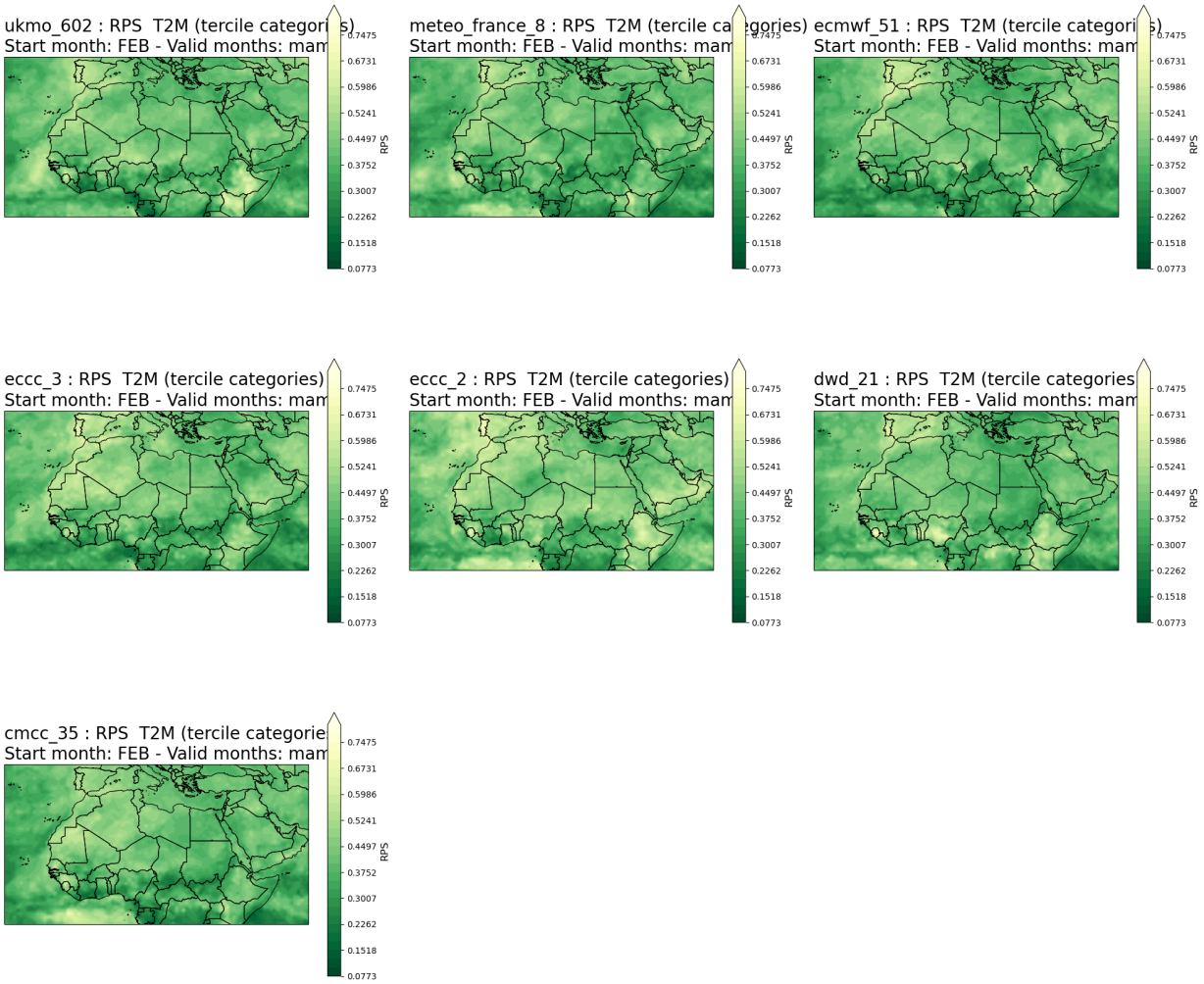

Figure 5.21: Temperature RPS heatmaps for all the seasons per categories ***(0 means perfect RPS)***

The figure above illustrates an excellent Ranked Probability Score (RPS), with values approaching 0, which signifies a high degree of accuracy and reliability in the probabilistic forecasts. This indicates that the forecasts effectively capture the uncertainty of outcomes and closely align with observed events.

Moreover, the spatial distribution of the RPS highlights consistent performance across the entire MENA region. The near-uniformity in low RPS values across this extensive and diverse area underscores the robustness of the forecasting system in managing the variability of climatic and weather conditions characteristic of the region.

Overall, the analysis confirms strong forecasting performance in the MENA region, with some variability in accuracy between forecasting centers. The results reinforce the importance of ongoing evaluation and optimization of forecasting systems to ensure consistent and high-quality predictions across all regions.

focus on north africa: For the North African region, the results mirror those observed in the broader MENA region.

This suggests that despite the localized focus on North Africa, the model performance differences remain significant, particularly for ***ECCC***.

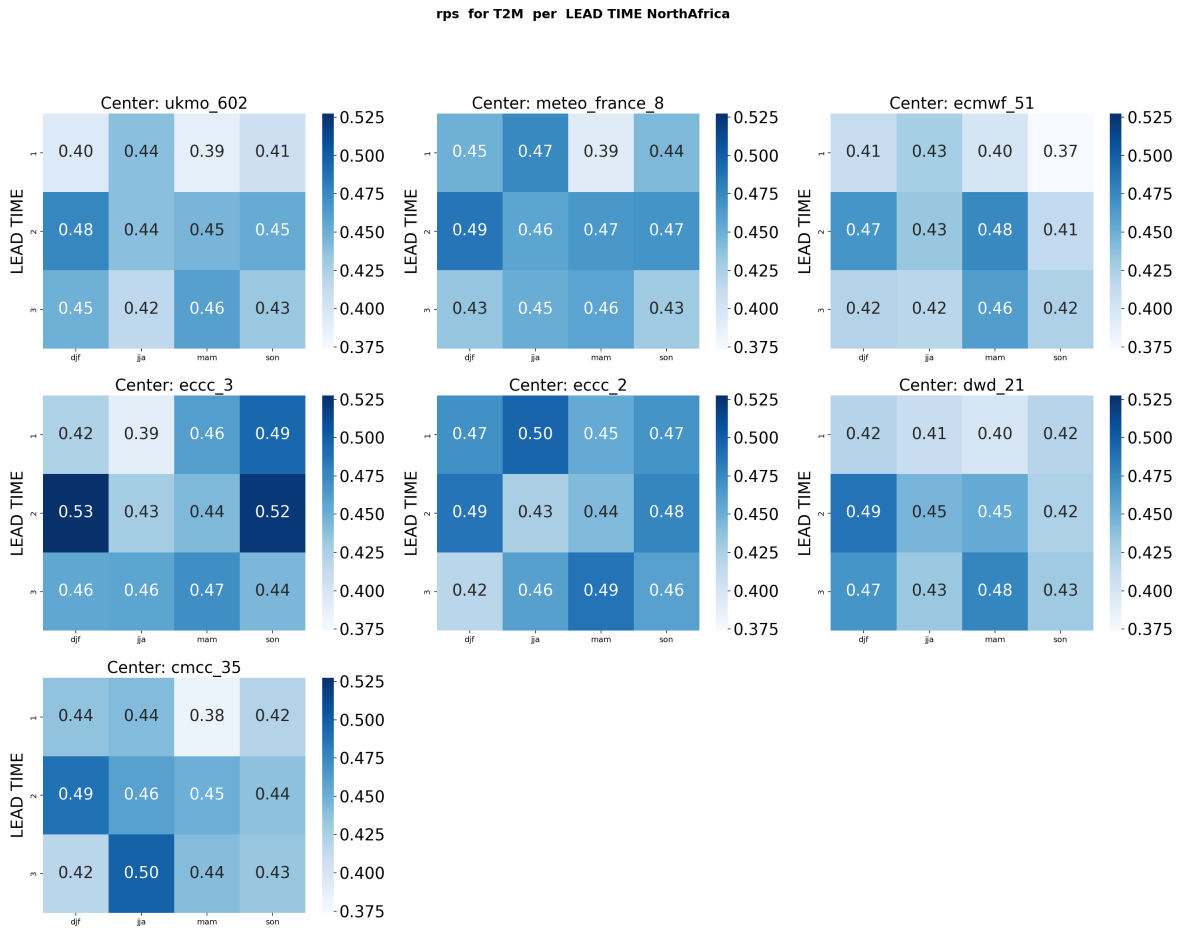

Figure 5.22: Heatmap of T2M rps for all centers in North Africa regions

focus on Arabian Peninsula : All centers exhibit nearly identical performance, with the exception of *ECCC*, which demonstrates lower skill compared to the others.

Receiver Operating Characteristic

The ROC (Receiver Operating Characteristic) curve is an important tool for evaluating the performance of predictive models, particularly in the context of probabilistic forecasts. It provides a graphical representation of the trade-off between the true positive rate and the false positive rate across various threshold levels.

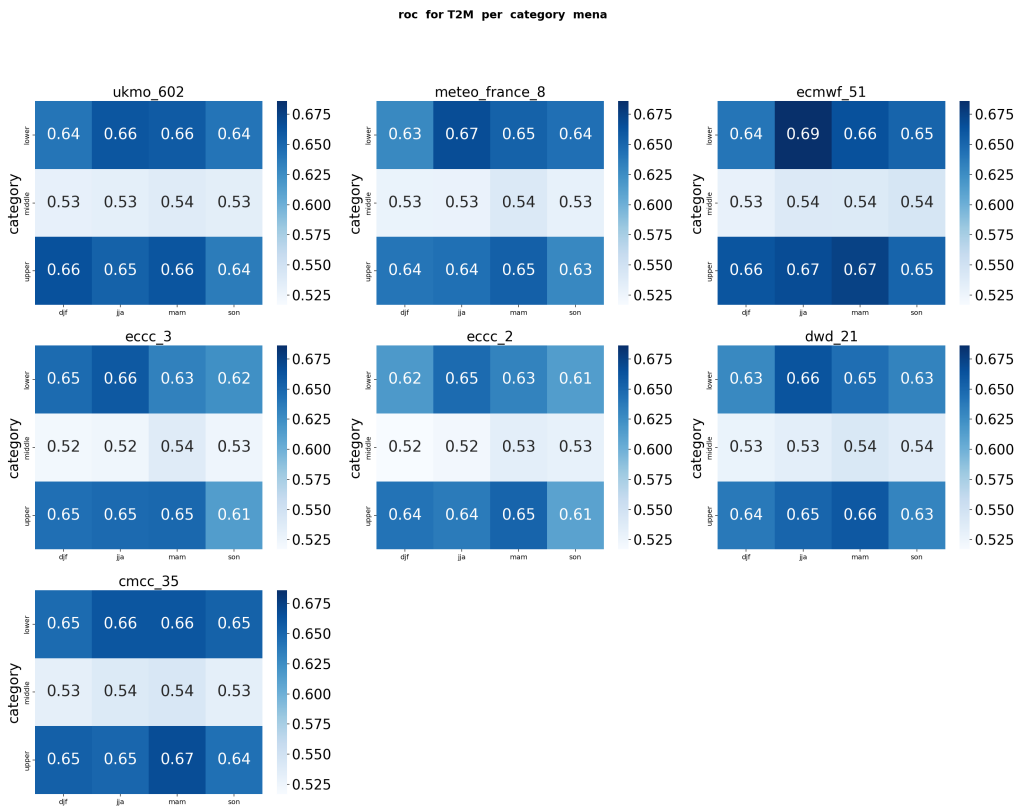

Figure 5.23: Temperature AUC heatmaps ***(1 means perfect ROC)***

Models generally exhibit similar performance, as indicated by the high Area Under the ROC Curve (AUC) values, which reflect their ability to effectively discriminate between predicted probabilities and observed outcomes. All centers perform relatively well in terms of the AUC, demonstrating good skill in distinguishing between forecasted events and non-events. Similar to the findings with the Brier Score (BS), the "middle" probability category tends to show weaker performance compared to the "lower" and "upper" categories. This highlights the models' greater sensitivity in accurately predicting events with extreme probabilities (high or low), but reduced skill for moderate probability scenarios. This consistency across metrics underscores the need to address forecast performance specifically in the middle category to further improve model predictions.

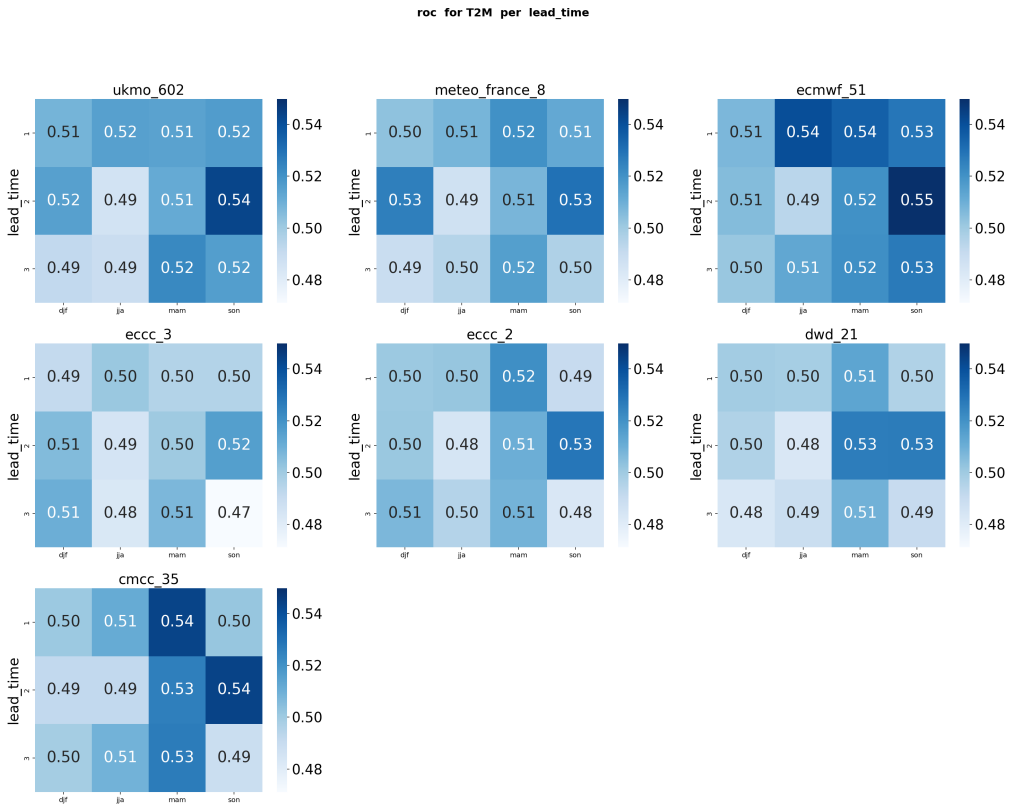

Figure 5.24: Temperature AUC heatmap per lead-time. ***(1 means perfect ROC)***

The figure above, shows the roc score along lead-time, a noteworthy good performance for SON is observed for the second lead-time, for all centers with values reaching 0.55. As for the other seasons, the performance stay in general stable along lead-time for all centers with scores around 0.5.

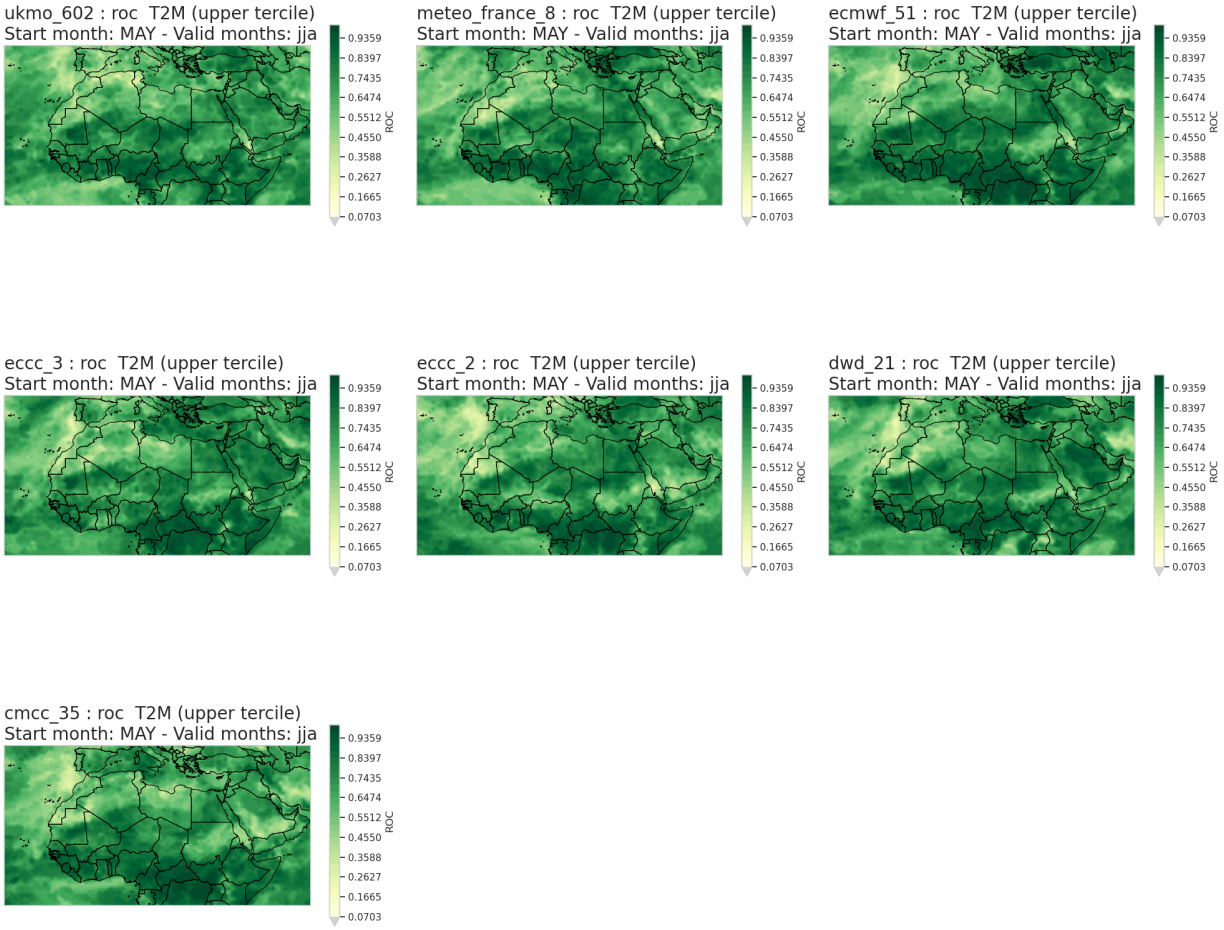

Figure 5.25: 2-meter Temperature ROC JJA Upper tercile ***(1 means perfect ROC)***

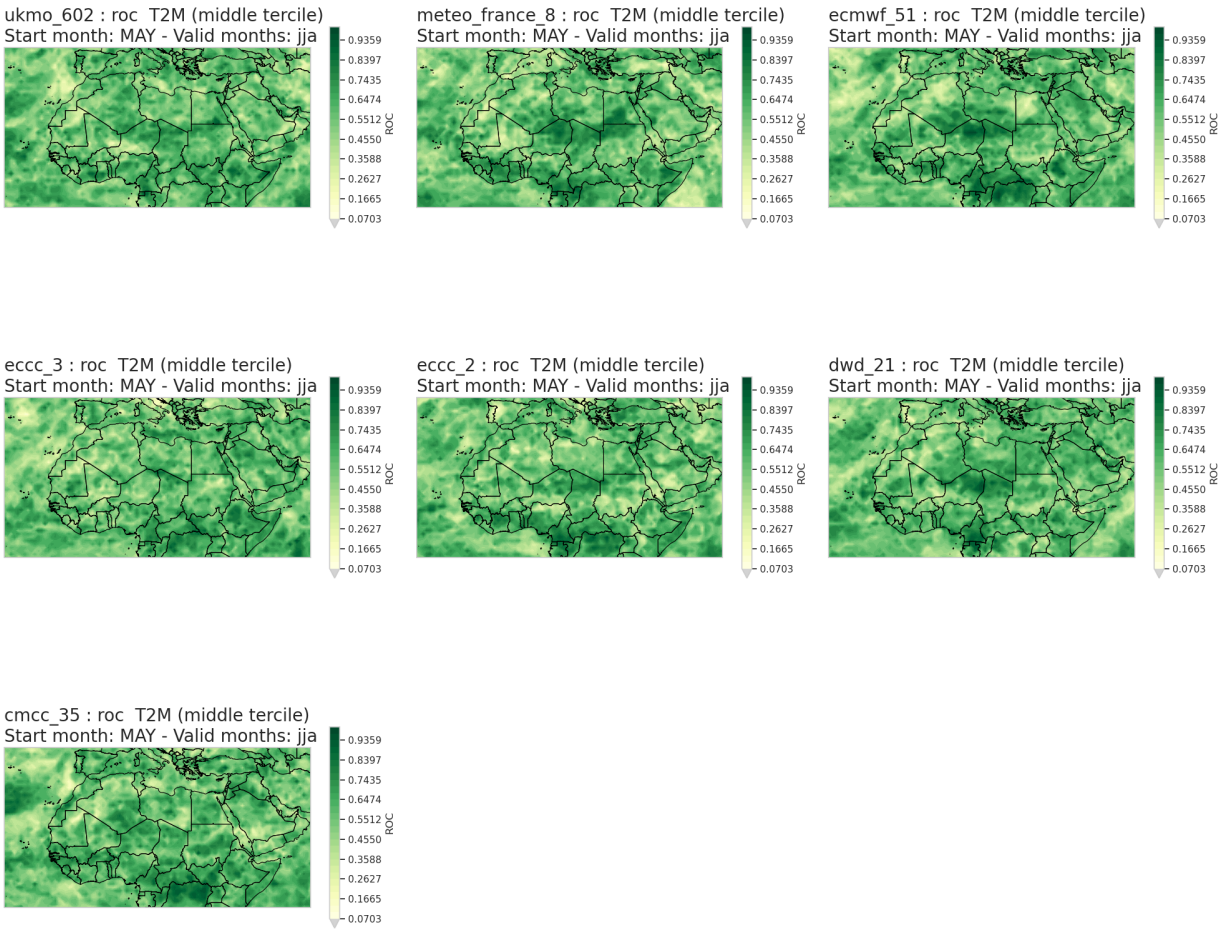

Figure 5.26: 2-meter Temperature ROC JJA Middle tercile ***(1 means perfect ROC)***

For the Upper tercile, the analysis of the spacial variation of ROC score (AUC), shows excellent scores for all scores especially ***ECMWF***. The spacial distribution shows a quite variation in North Africa. Thus, the ***DWD*** exhibit the best performance for Arabian peninsula. Hence, the discrimination (Ability to distinguish events from non-events) is good for the upper tercile as well as for the lower tercile.

As for the Middle tercile, the performance is much lower. The spacial variability is very high and there is no clear pattern of the score, reflecting a week discrimination for the middle tercile. Thus, the discrimination of extreme events (lower and above normal) is much better than the normal situations.

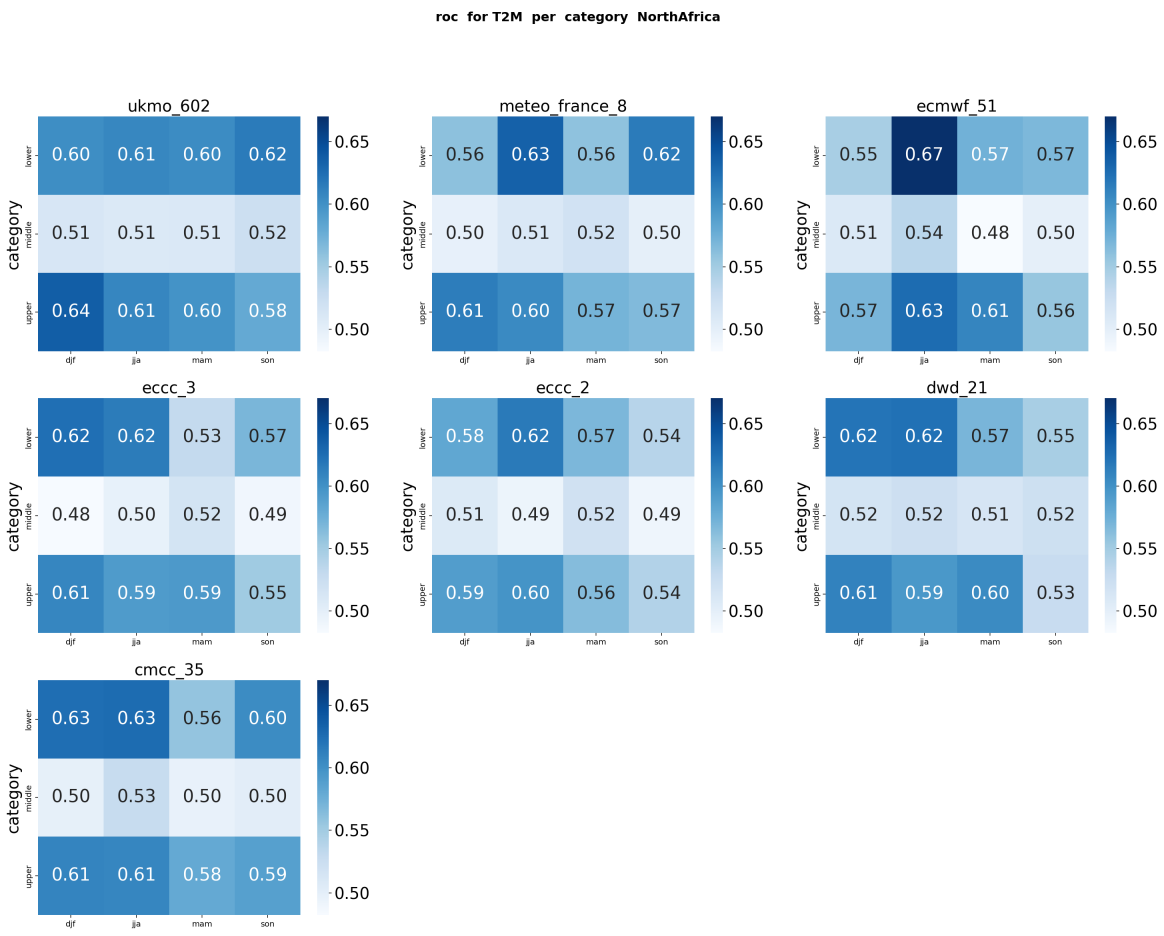

Figure 5.27: Temperature AUC heatmaps for north africa

focus on north africa: The figure above confirms the same conclusions for the North Africa region. The models generally maintain similar performance, with high AUC values reflecting strong discrimination skill across all categories. The UKMO continues to show robust results in terms of ROC. As observed previously, the "middle" probability category remains the least performant compared to the "lower" and "upper" categories, indicating the models' reduced ability to predict moderate probability events. This consistency in findings suggests that the regional focus on North Africa does not significantly alter the overall assessment of model performance.

focus on Arabian Peninsula : There are no significant differences in performance across the centers, indicating comparable skill levels.

Relative operating characteristics Skill Score

ROCSS provides an assessment of a model's ability to discriminate between observed and forecasted events relative to a reference model, often a climatological or random forecast. A higher ROCSS

indicates that the model has skill in distinguishing between occurrences and non-occurrences of an event, while a score close to zero suggests no significant improvement over the reference.

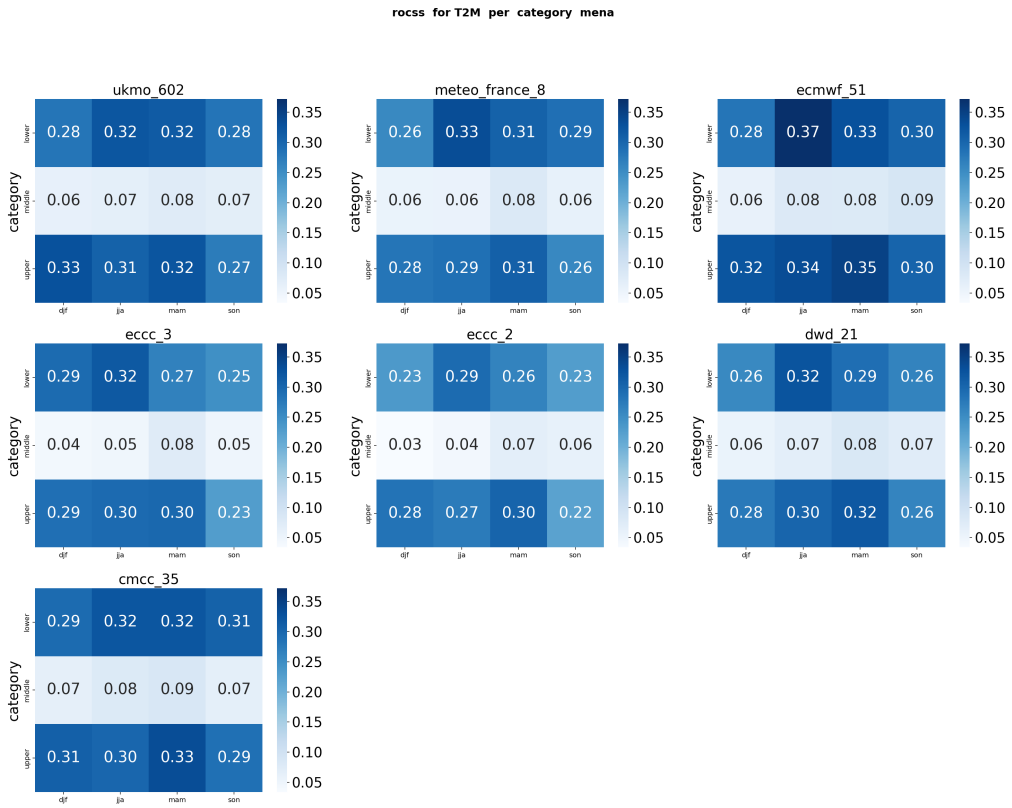

Figure 5.28: Temperature ROCSS heatmaps for MENA region per category ***(1 means perfect ROCSS)***

The models generally demonstrate consistent and positive skill, highlighting their ability to discriminate between observed and forecasted events. UKMO, which showed good performance in reliability metrics, continues to perform well in terms of ROCSS, confirming its relative robustness in event discrimination. Additionally, as observed with the ROC scores, the "middle" category exhibits lower performance compared to the "lower" and "upper" categories. This indicates that while the models excel at predicting extreme events with high or low probabilities, their ability to capture moderate probability events remains limited.

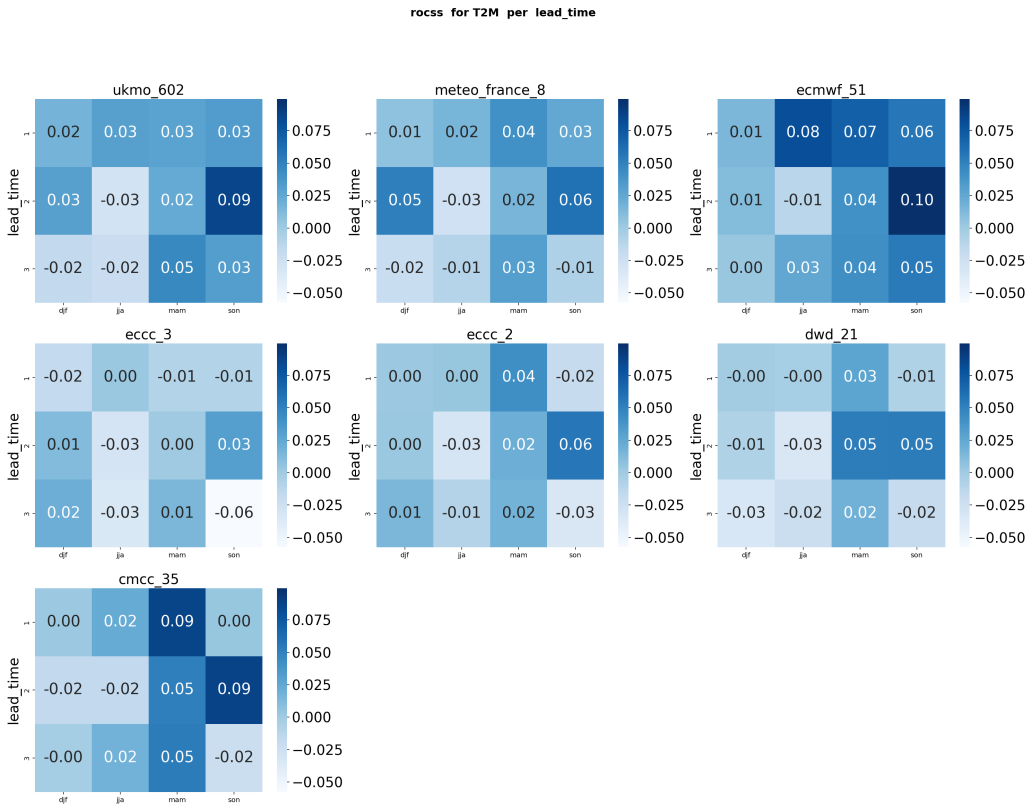

Figure 5.29: Temperature ROCSS heatmaps for MENA region per lead-time. ***(1 means perfect ROCSS)***

The SON maintain its good performance for the second lead-time with 0.10 for the ***ECMWF*** and 0.09 for the ***UKMO and CMCC-35*** , in general the performance is low for all centers. The DJF and JJA exhibit the lowest performance for all centers, although, the ***ECMWF*** shows relatively good performance for the 1st lead-time of JJA (0.08).

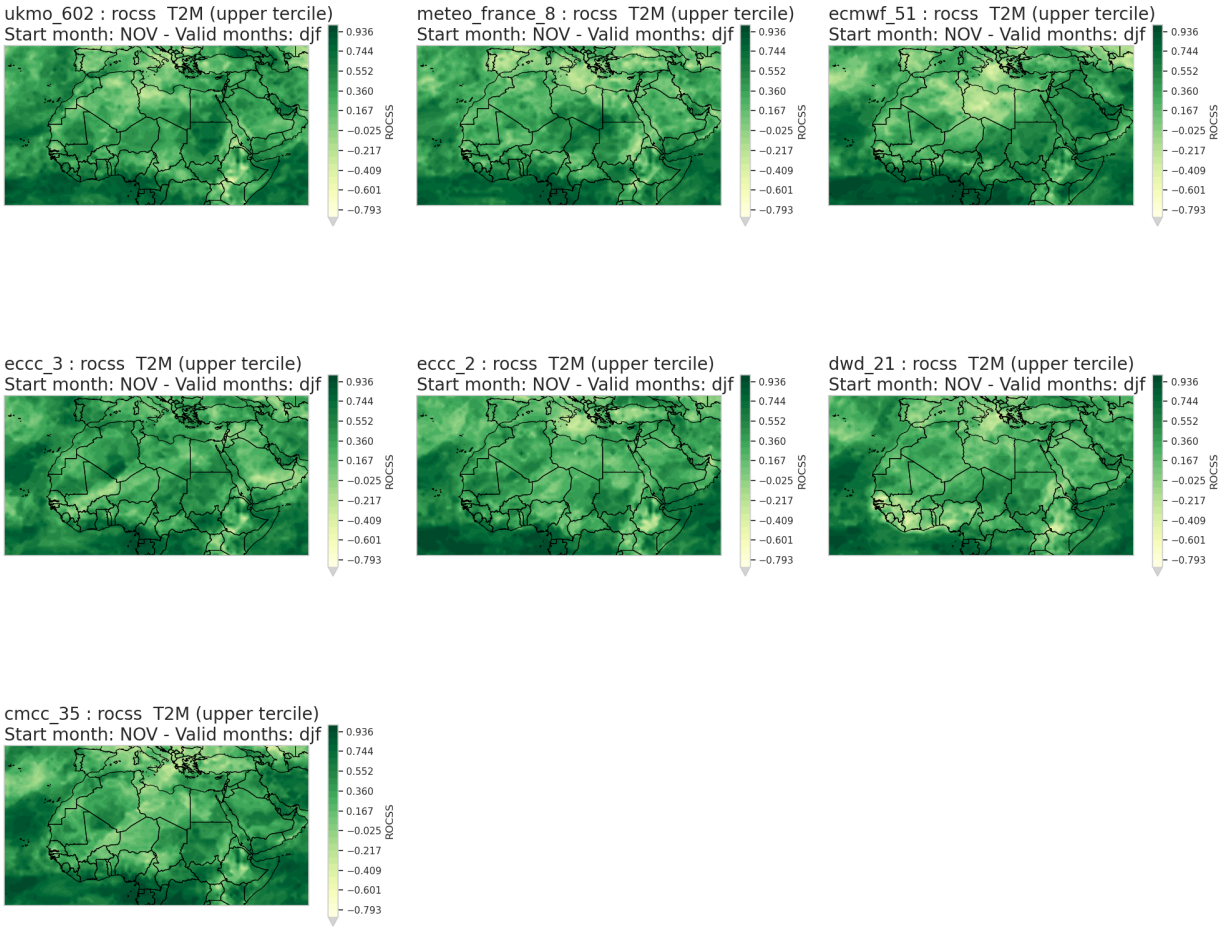

Figure 5.30: 2-meter Temperature ROCSS DJF Upper tercile ***(1 means perfect ROCSS)***

For the upper tercile of DJF, the performance is good in the equator and the Arabian Peninsula with rocss around 0.9. Nevertheless, the North Africa shows negative values which means that the forecast performs worse than random guessing, indicating a forecast that is worse than a no-skill forecast.

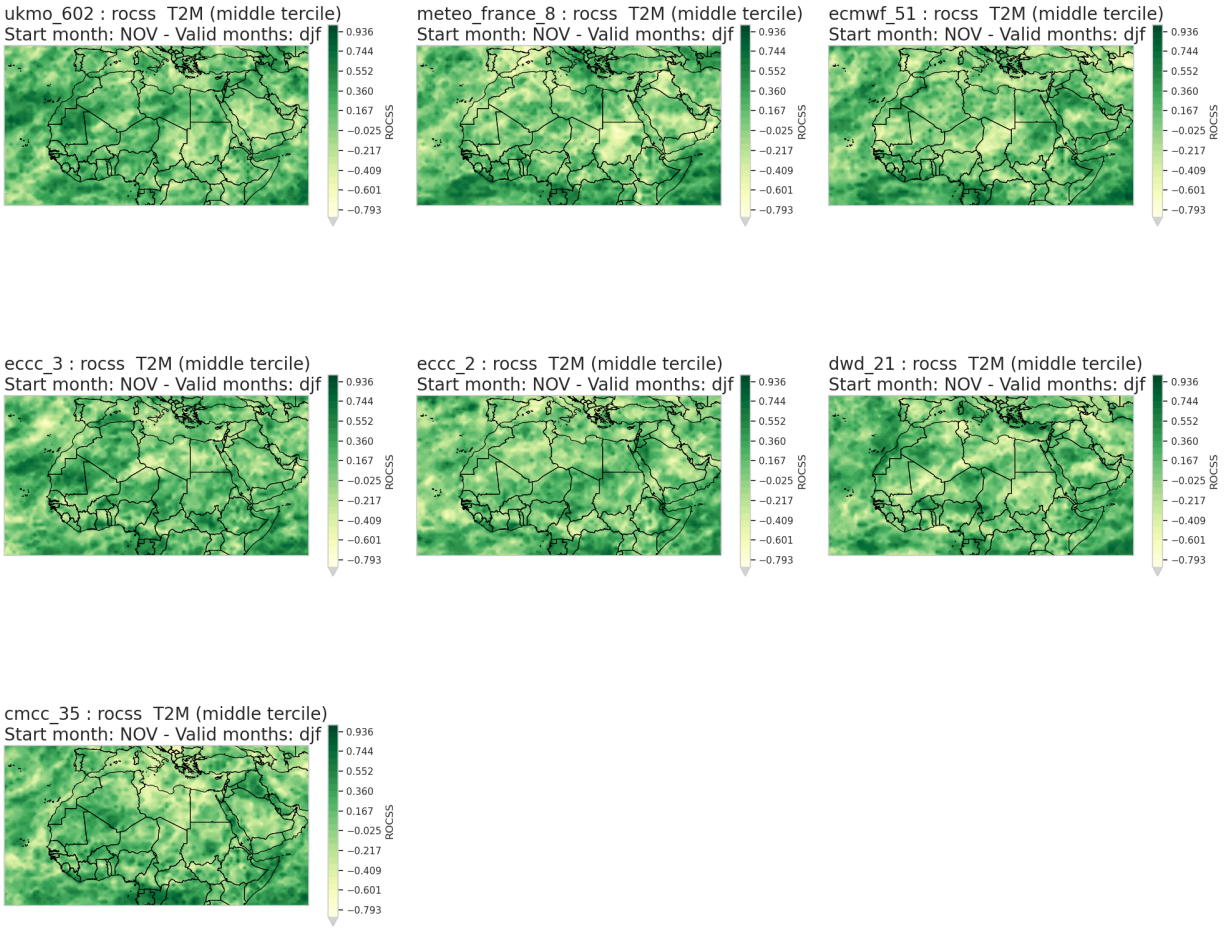

Figure 5.31: 2-meter Temperature ROCSS DJF Middle tercile ***(1 means perfect ROCSS)***

The Middle Tercile exhibits weak and inconsistent performance, with no clear spatial variability across the region. The scores are generally low and even negative in many areas, with values dropping as low as -0.79, particularly in North Africa. Despite this overall weak performance, there are a few isolated regions where the scores are notably higher, reaching up to 0.9 in certain areas of the Arabian Peninsula and along the equator. These localized high scores contrast with the broader pattern of poor performance, suggesting some regional variability in the model's predictive ability.

focus on north africa: The ROC Skill Score (ROCSS) analysis reveals similar conclusions to the ROC results for Mena region and north Africa.

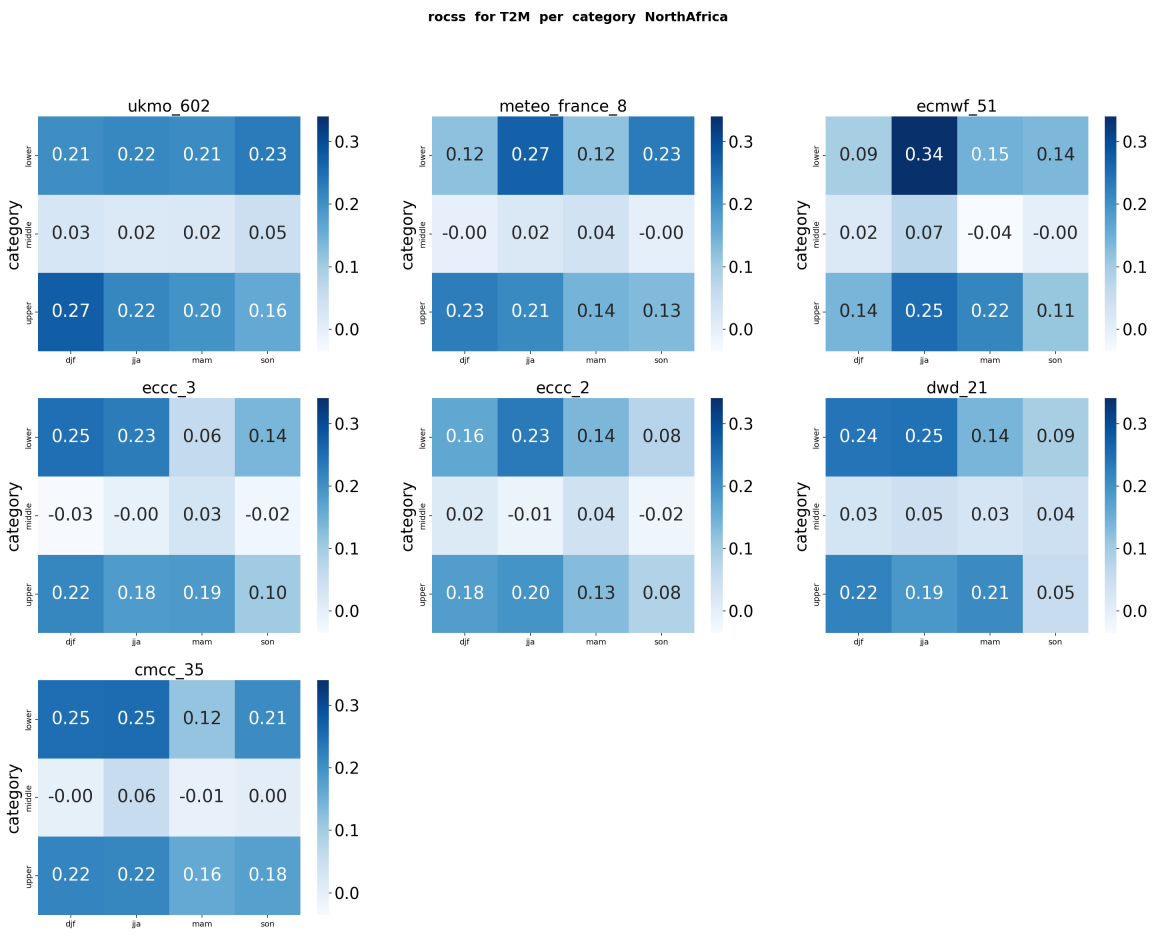

Figure 5.32: Temperature ROCSS heatmaps for north africa

focus on Arabian Peninsula : There are no significant differences

5.2 PRECIPITATIONS

IN general, the forecast of precipitations is more complicated than temperature, thus the scores are a little less good for this part especially the deterministic ones.

5.2.1 Deterministic Evaluation Metrics

Anomaly Correlation Coefficient

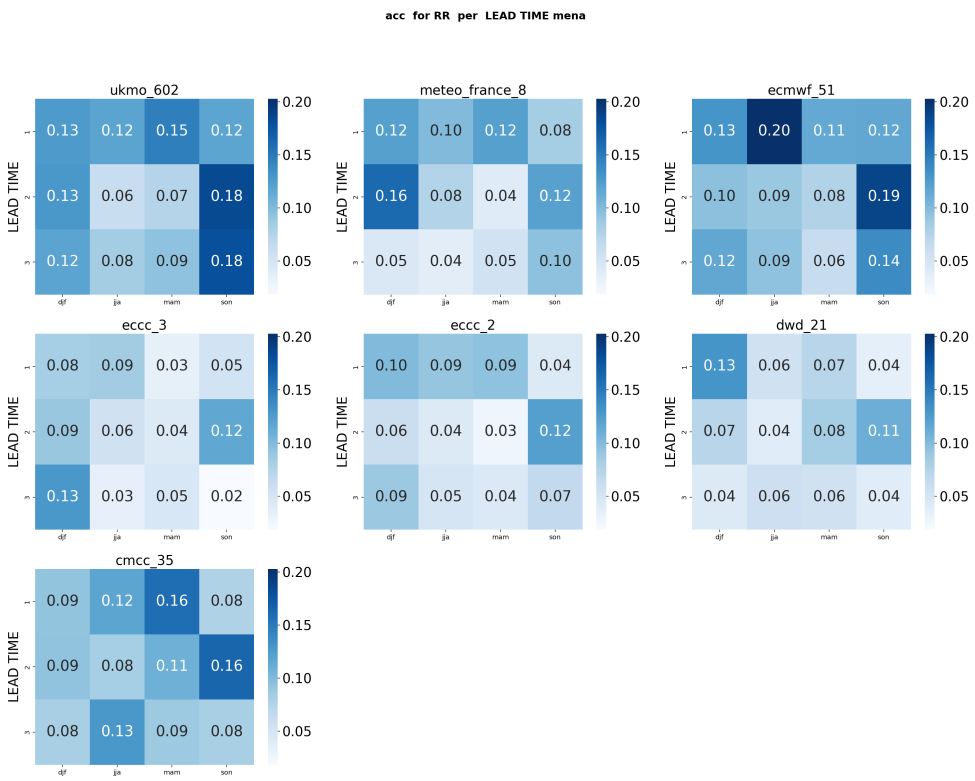

Figure 5.33: The Heatmap of acc for the mena region for every period ***(1 for perfect ACC)***

The acc is moderate for all centers; however, the best models are ***ECMWF, UKMO, and CMCC-35***. There is no clear variability in performance along lead-time. For SON, the performance is good at lead-time 2 for all centers. As for the other seasons, the performance is generally strong at the 1st lead-time but decreases with increasing lead-time. Hence, Meteo-France also shows good performance, but it decreases significantly over time.

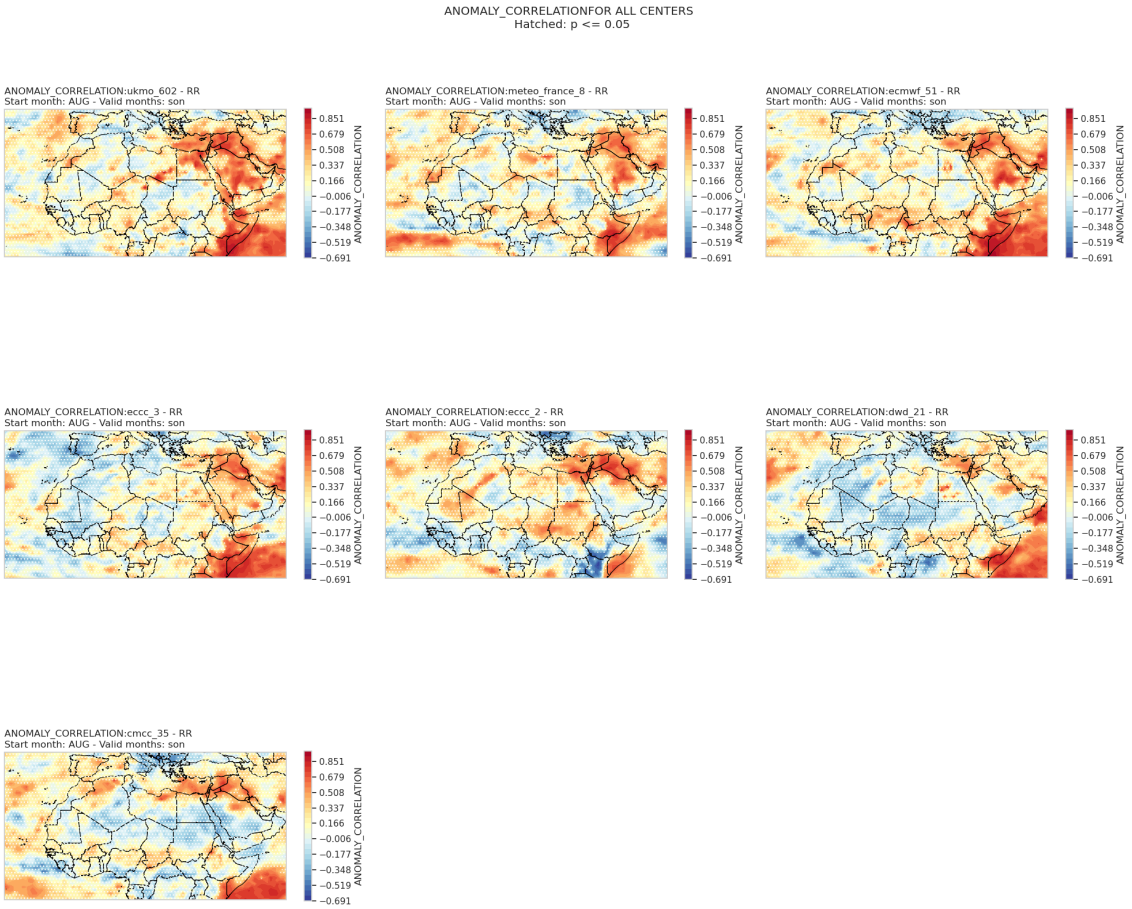

Figure 5.34: 3-months Rolling mean of Anomaly Correlation in MENA Region for all centers SON

For temperature, the models demonstrate the best performance in the tropical regions. However, for precipitation, the situation is different. Hence the results show good performance during SON, where the Arabian Peninsula and East Africa exhibit the highest performance.

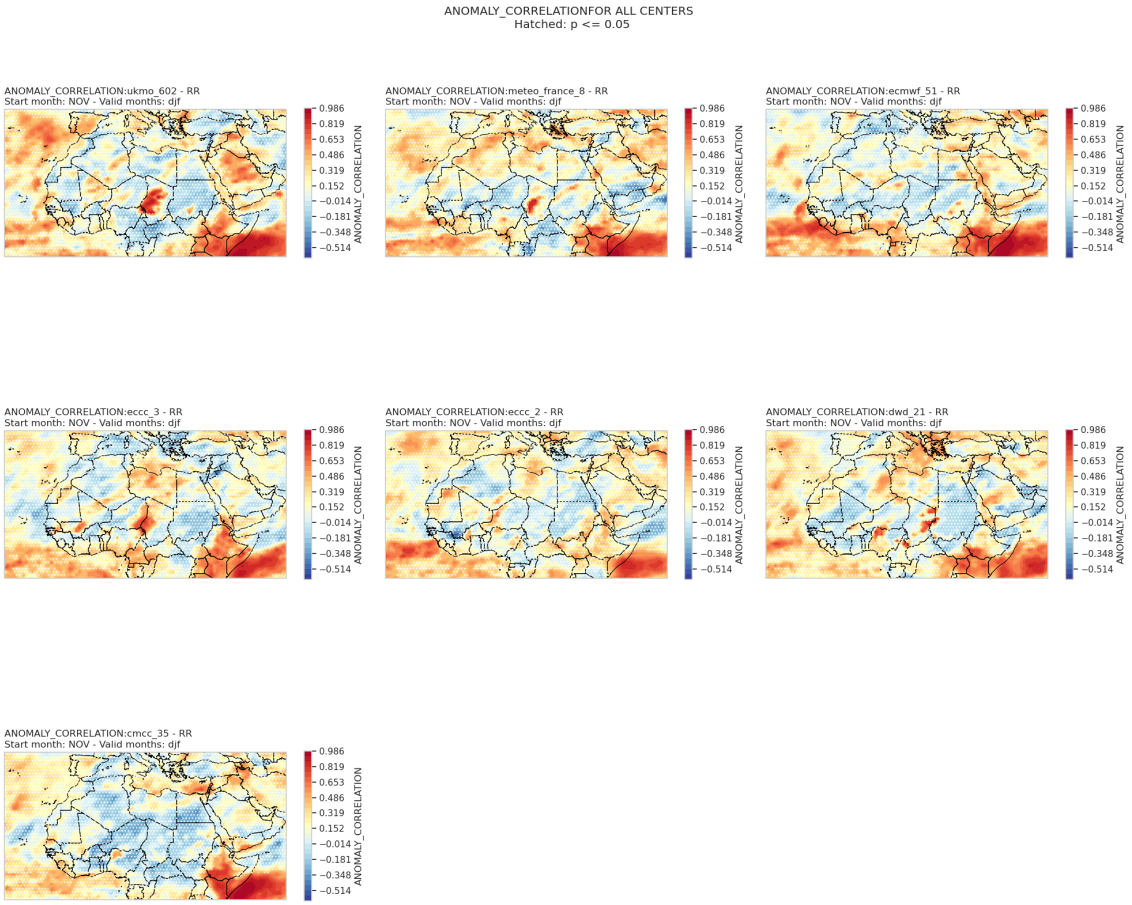

Figure 5.35: 3-months Rolling mean of Anomaly Correlation in MENA Region for all centers DJF

The 3-month rolling mean for DJF acc shows that the best models are ***ECMWF, UKMO, and Meteo-France***. The acc is significant across most of the MENA region, except in the east of Africa. Thus, for all centers the East of Africa have the highest score. However, the ***ecmwf and ukmo*** show good performance for

For SON, the situation is generally better than for DJF in the Arabian peninsula. In general, ***ECMWF and Meteo-France*** are the best. Nevertheless, the ACC isn't significant in the east of africa and the north of the Arabian peninsula despite of the high acc.

focus on north africa: according to the heatmap below, the correlation shows no big difference for the first lead-time, but for the second and third lead-times, it became lower. Thus, the ***ecmwf,ukmo and meteo-france*** maintain relatively good correlation.

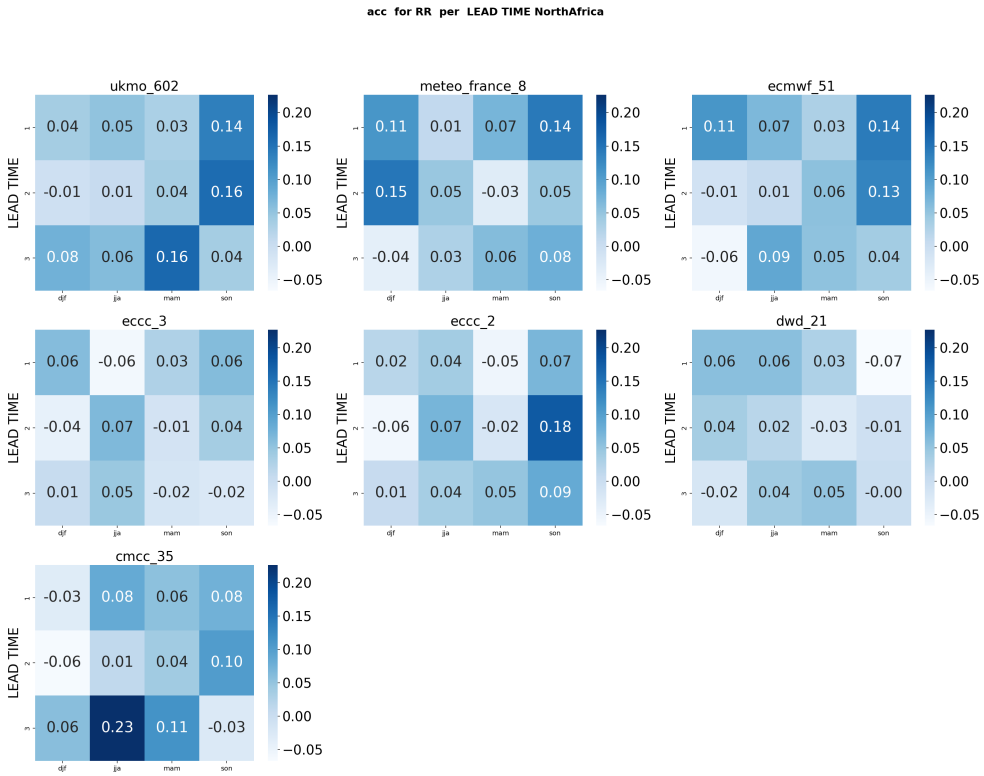

Figure 5.36: The Heatmap of ACC for the North Africa region for every period ***(1 for perfect Correlation)***

focus on Arabian Peninsula :

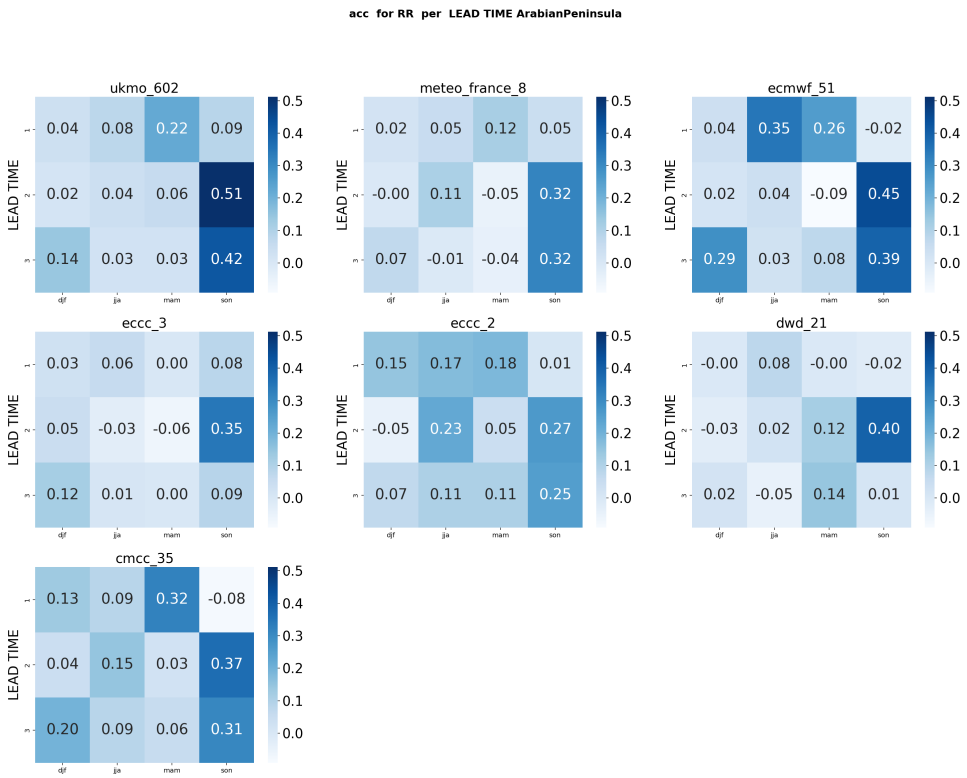

Figure 5.37: The Heatmap of acc for the Arabian Peninsula region for every period ***(1 for perfect ACC)***

The analysis of the Arabian Peninsula shows that the score is much better than the general situation of the mena region. The acc is good especially for SON for the second and third lead-times.

Root Mean Square Error

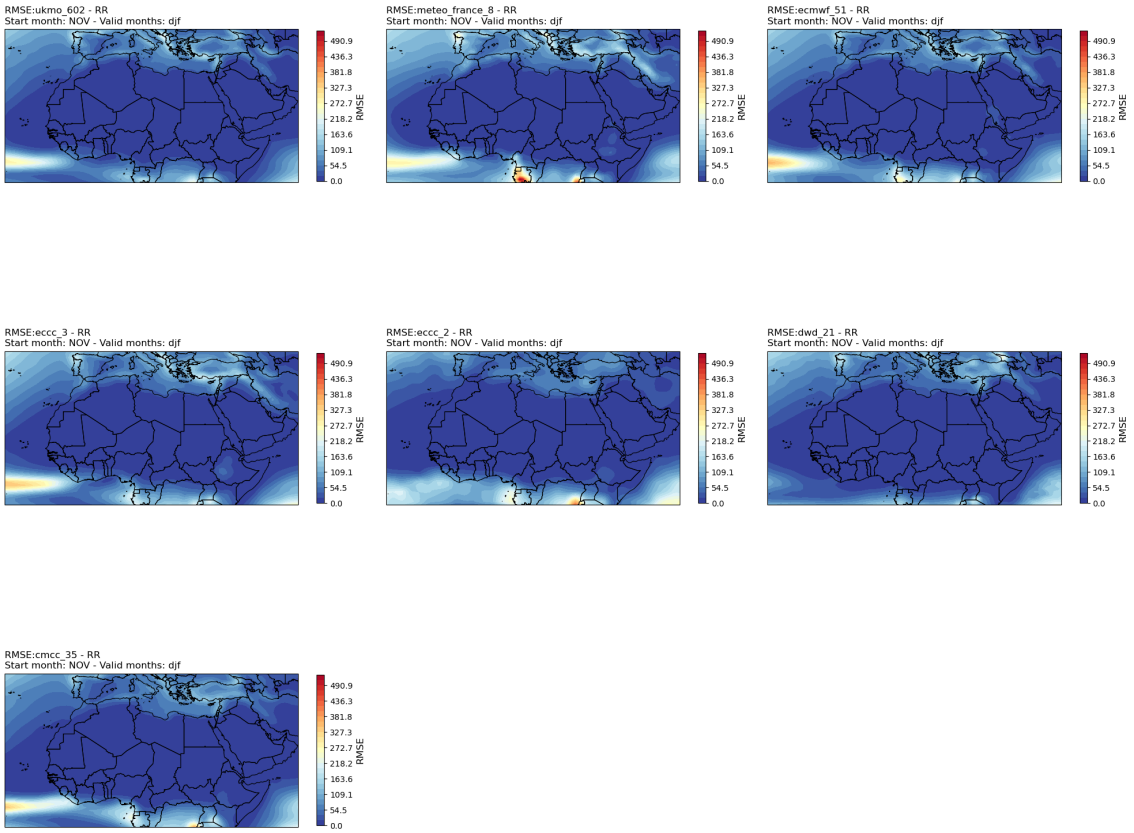

Figure 5.38: 3-months Rolling mean of RMSE in MENA Region for all centers DJF in mm

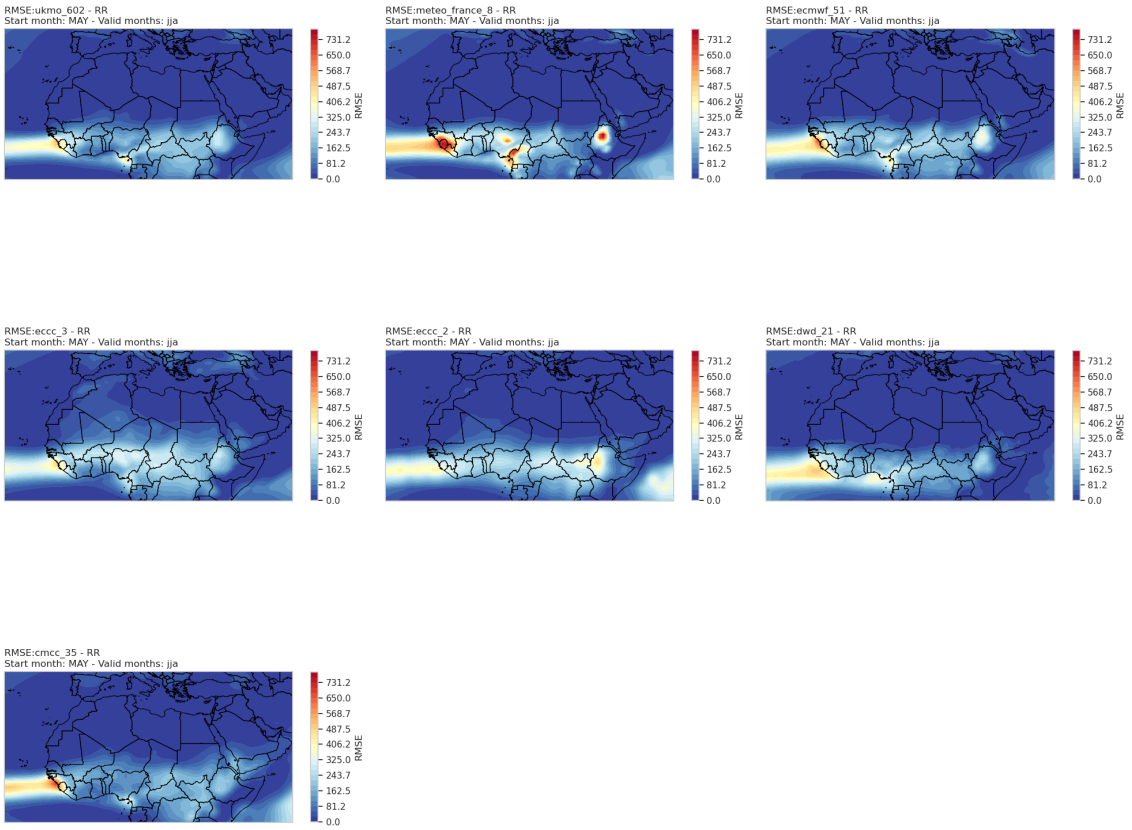

Figure 5.39: 3-months Rolling mean of RMSE in MENA Region for all centers JJA in mm

also for the spacial dimension, the RMSE stay stable and exhibit moderate performance for all centers. Thus, all models have almost the same skill and they are consistent with each other. We can see some isolated areas in the equator where the RMSE is very high reaching 731 mm for all centers especially ***ECMWF and METEO-FRANCE*** for both DJF and JJA especially in JJA.

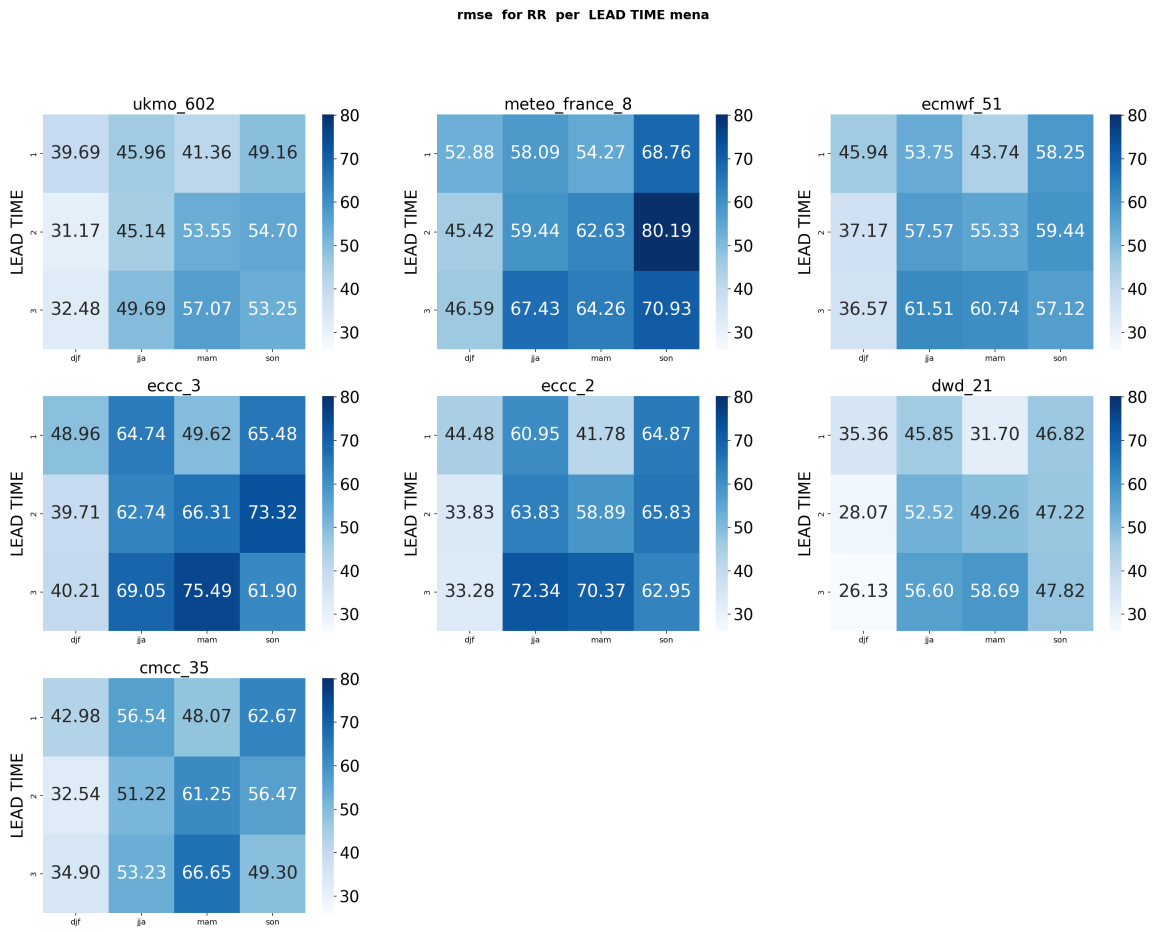

Figure 5.40: heatmap of RMSE For RR in mm

for the Root Mean Squared Error, the best models shown in the heatmap above are ***DWD, ECMWF and UKMO*** .The RMSE score demonstrate a moderate performance for all models especially ***DWD, ECMWF and UKMO***. The performance is stable over lead-times and it is much better for djf in all centers with values between 26 and 35 for ***DWD*** . The SON exhibit the biggest RMSE for all centers, with values around 70 in ***METEO-FRANCE***.

focus on North Africa :

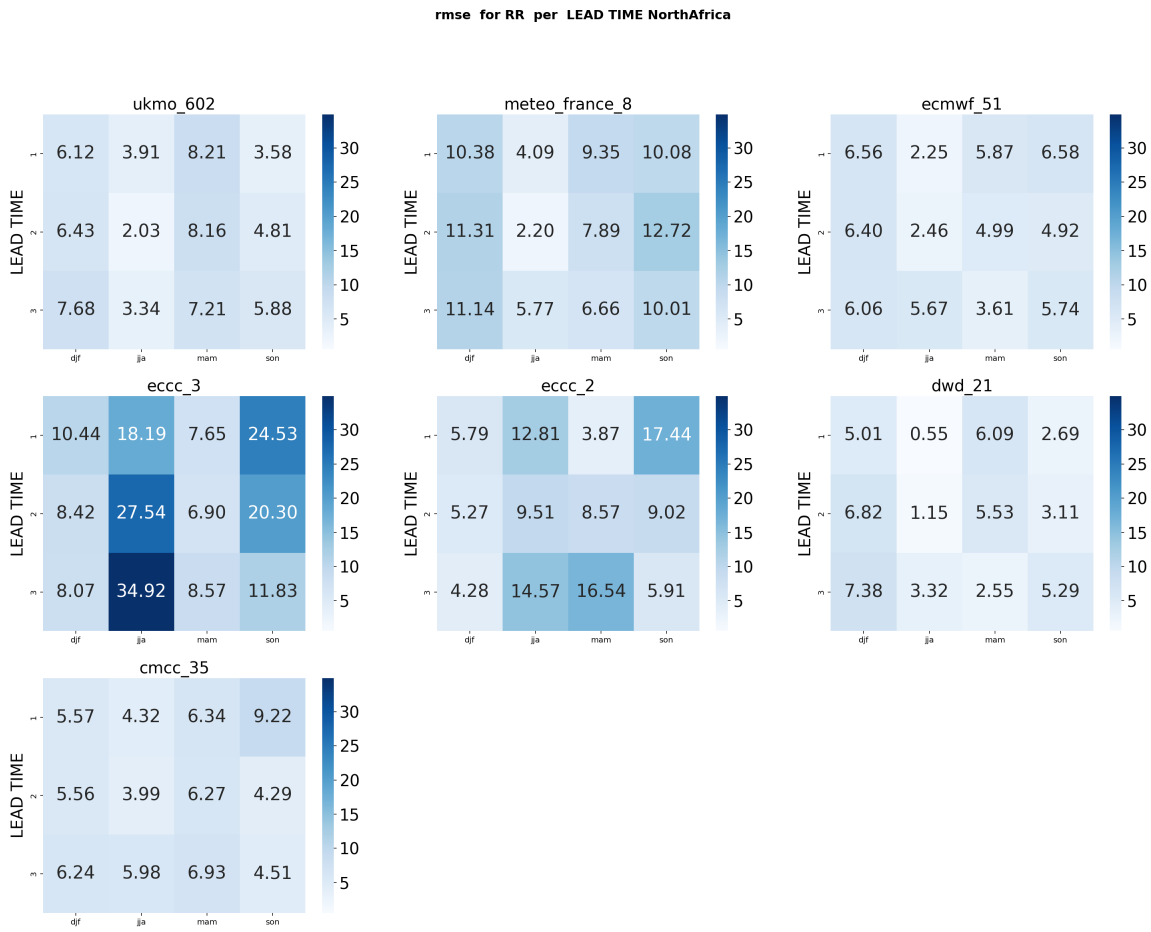

Figure 5.41: heatmap of RMSE For RR in mm (North Africa)

the RMSE is much better for North africa, the score is good over all lead-times and seasons. The centers, ***ecmwf, ukmo and dwd*** show very good performance.

focus on Arabian Peninsula :

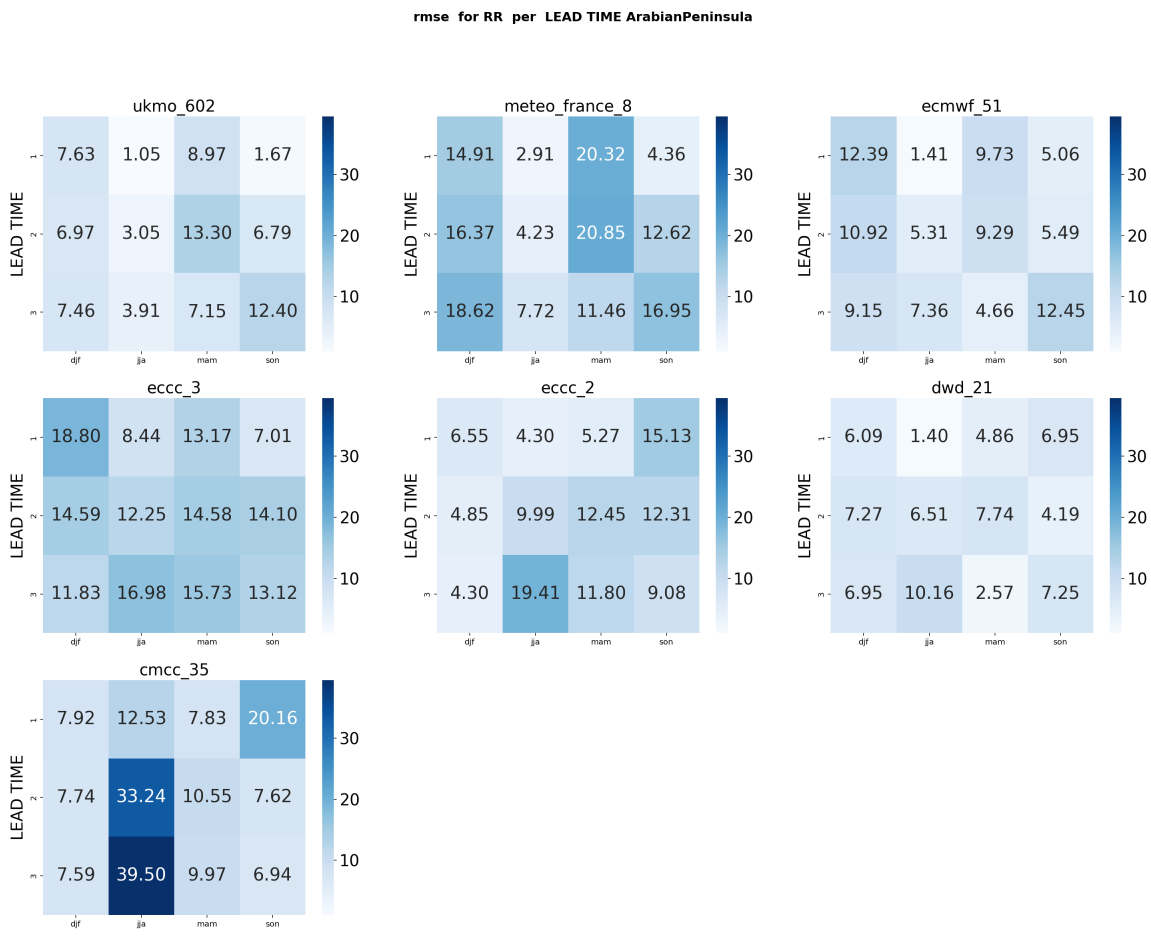

Figure 5.42: heatmap of RMSE For RR in mm (Arabian Peninsula)

In the same way as North Africa, the RMSE for the Arabian Peninsula is much better than mena. The centers, ***ecmwf, ukmo and dwd*** show very good performance.

Coefficient of Determination (R^2)

for precipitation, the R-SQUARED is very low, the maximum value is less than 0.1. However, the ecmwf is the best in term of R-SQUARED. for DJF,JJA and MAM the highest performance is in the first Lead-time, and it decrease along time, But for SON the best score is in the second Lead-time for all centers.

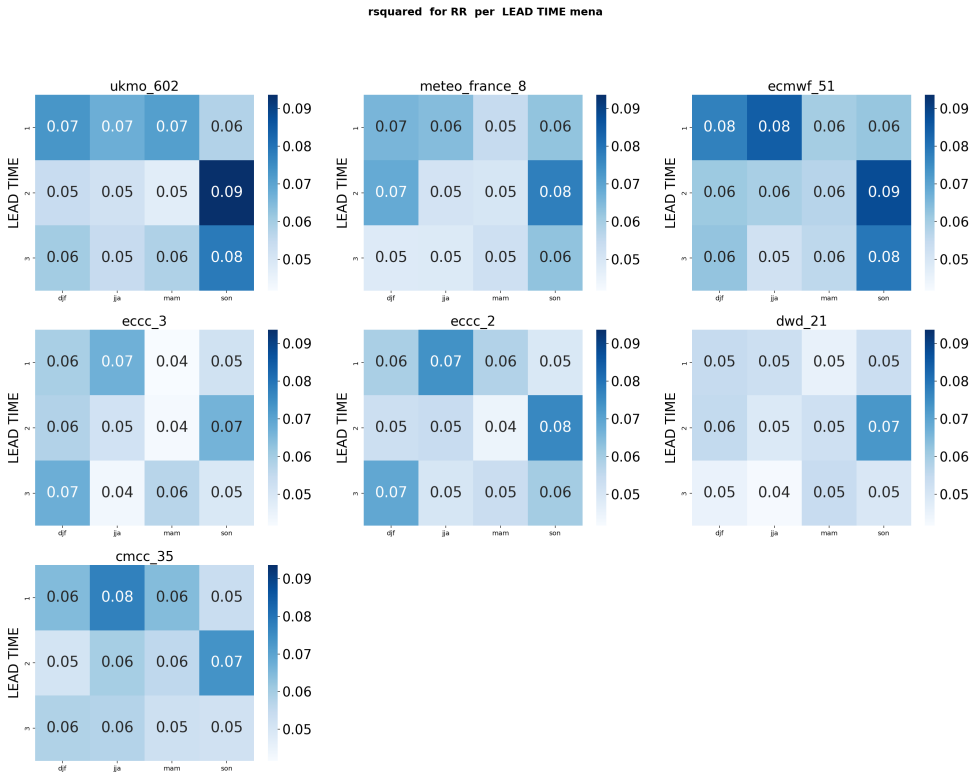

Figure 5.43: The Heatmap of rsquared for Precipitations in the mena region for every period ***(1 for perfect RSQUARED)***

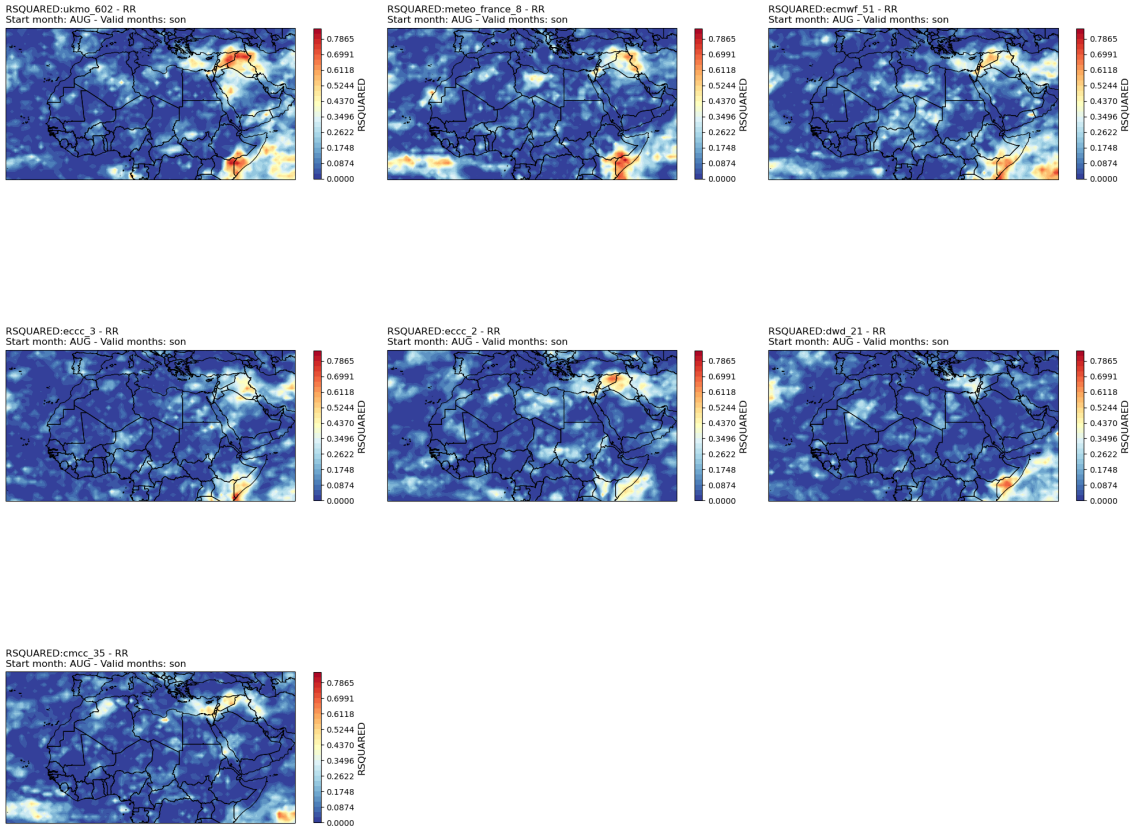

Figure 5.44: 3-months Rolling mean of RSQUARED in MENA Region for all centers SON

there is some isolated zones where the r-squared is good especially in Syria, Irak, Jordan ,Palestine and East Africa, this high performance is observed in all centers. For the rest of the MENA region the performance is very bad with score near to 0. Hence, there is no constant pattern for the R-SQUARED, the spacial variation is very high for all centers.

focus on North Africa there is no big difference in North Africa.

focus on Arabian Peninsula :

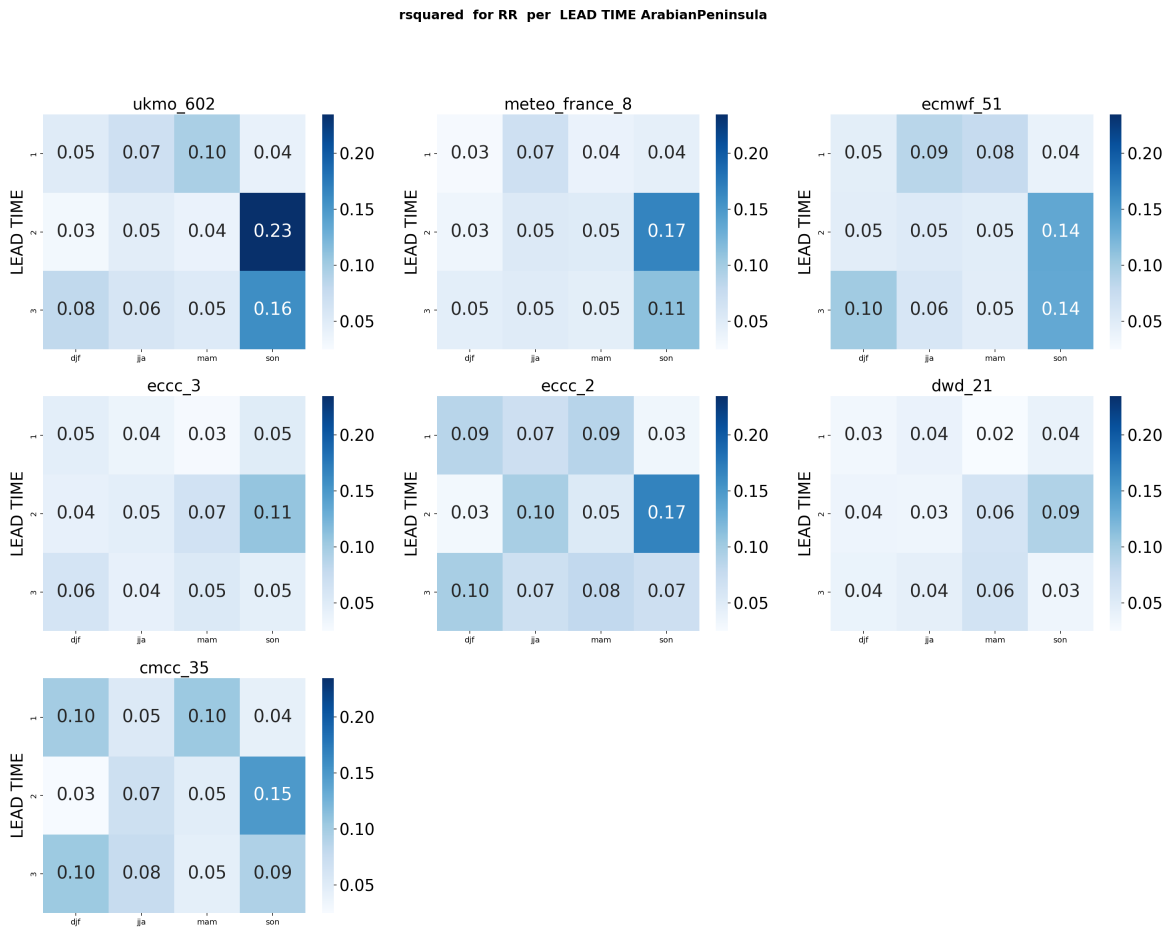

Figure 5.45: Heatmap of RR RSQUARED in MENA Region for all centers Arabian Peninsula

the R-SQUARED for the Arabian Peninsula shows a little improvement.

5.2.2 Probabilistic Evaluation Metrics

The Brier Score (BS)

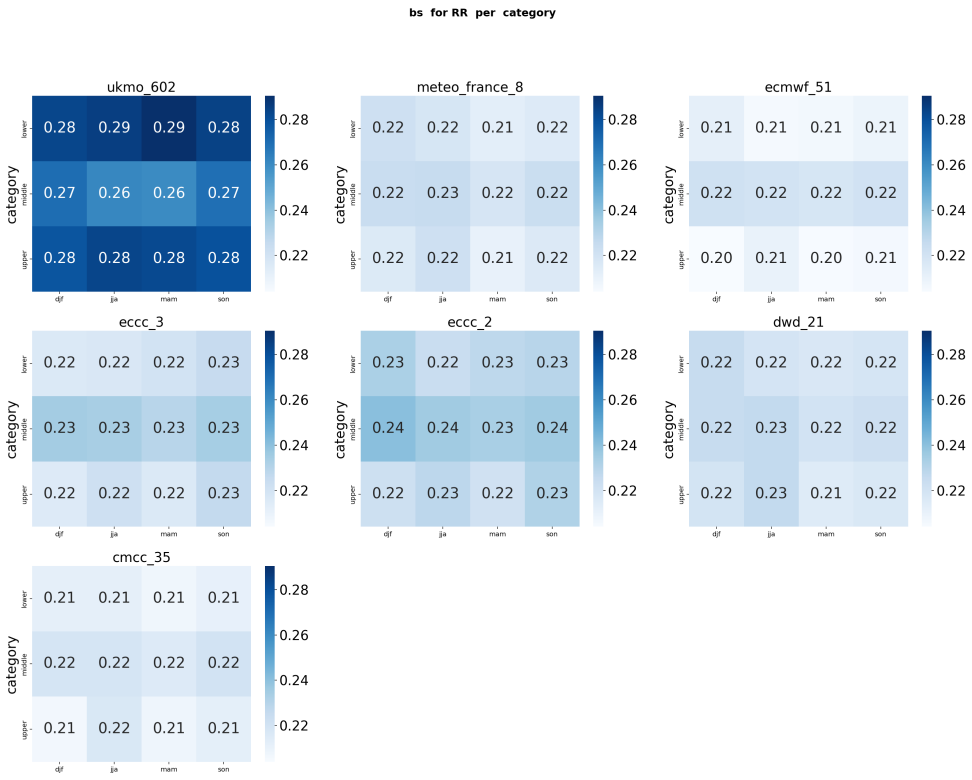

Figure 5.46: The Heatmap of Brier Score for each category . ***(0 represents perfect BS)***

for the analysis per category, we can see in the figure above that all centers exhibit good performance in term of Brier Score. Overall, the middle tercile shows lower performance (higher Brier Score) for all centers. the figure below shows the analysis per lead-time. the same result is found, but the ***ECMWF,METEO-FRANCE and CMCC-35*** are the best models in Brier Score for lead-time analysis.The performance stay stable along time which is a reliable signal. Despite the UKMO have the lower performance, it stays close to the other centers, the difference isn't so wide. In general, the performance stays stable over category, lead-time and space.

Figure 5.47: The Heatmap of Brier Score for lead-time. ***(0 represents perfect BS)***

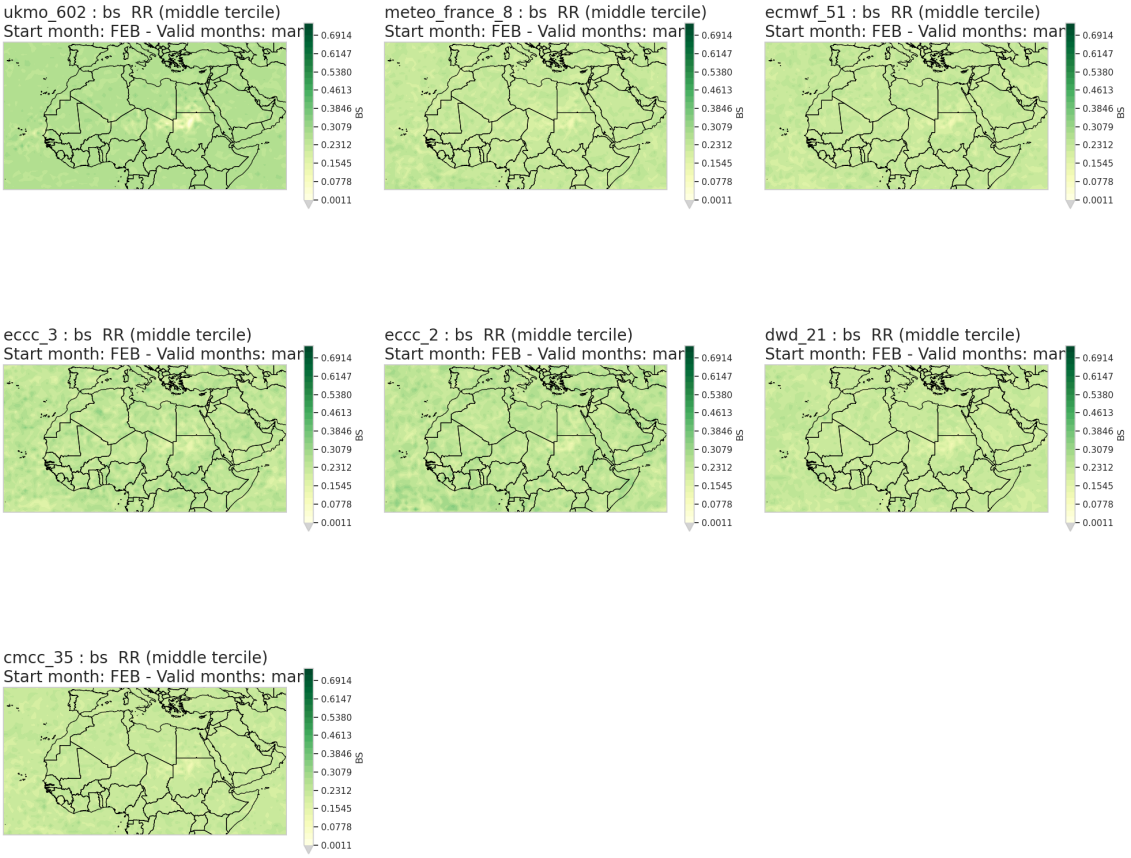

Figure 5.48: 3-months Rolling mean of Brier Score in MENA Region for all centers middle tercile MAM

the spacial distribution of the BS is homogeneous, the same performance across the MENA region, almost all centers perform well for all lead-times, for tercile there is a little lower performance for the middle tercile. Hint, for the other seasons the results are almost the same.

focus on North Africa :
there is no big difference in North Africa.

focus on Arabian Peninsula :
there is no big difference in Arabian Peninsula.

Reliability

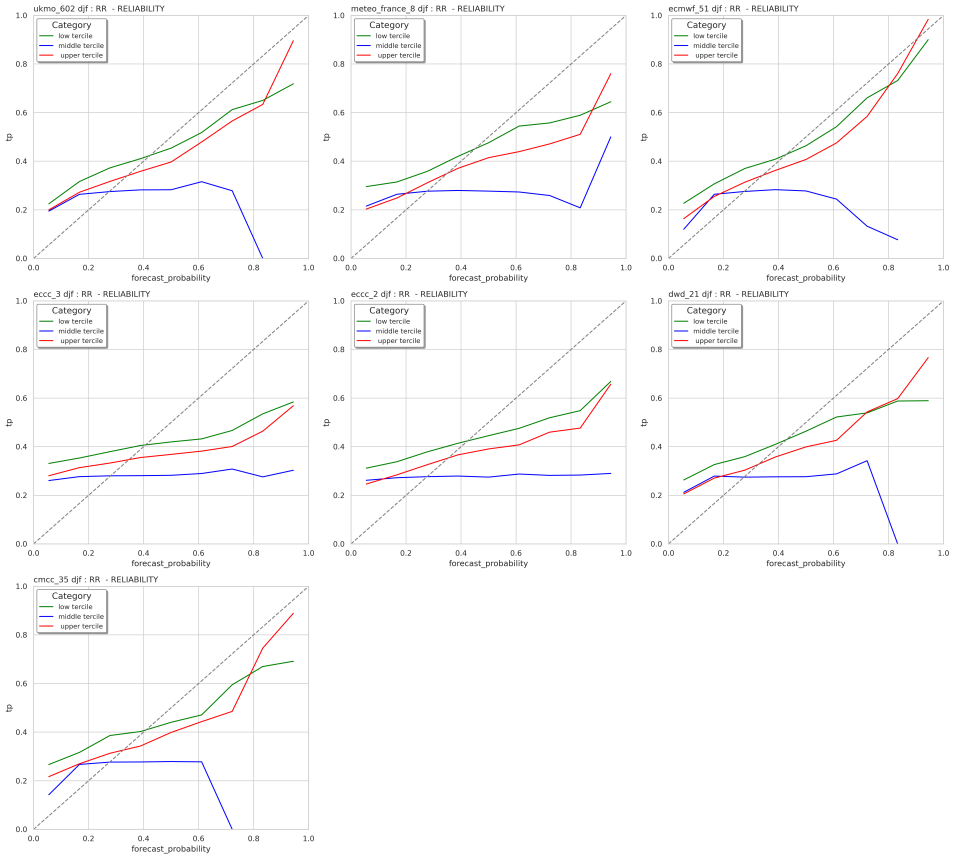

Figure 5.49: The Reliability diagram . ***(45 degree for perfect reliability)***

The reliability diagram for DJF, shows moderate performance, the ***UKMO, ECMWF and CMCC-35*** have good performance especially for upper and lower terciles, nevertheless, the performance for middle tercile is week for all centers.

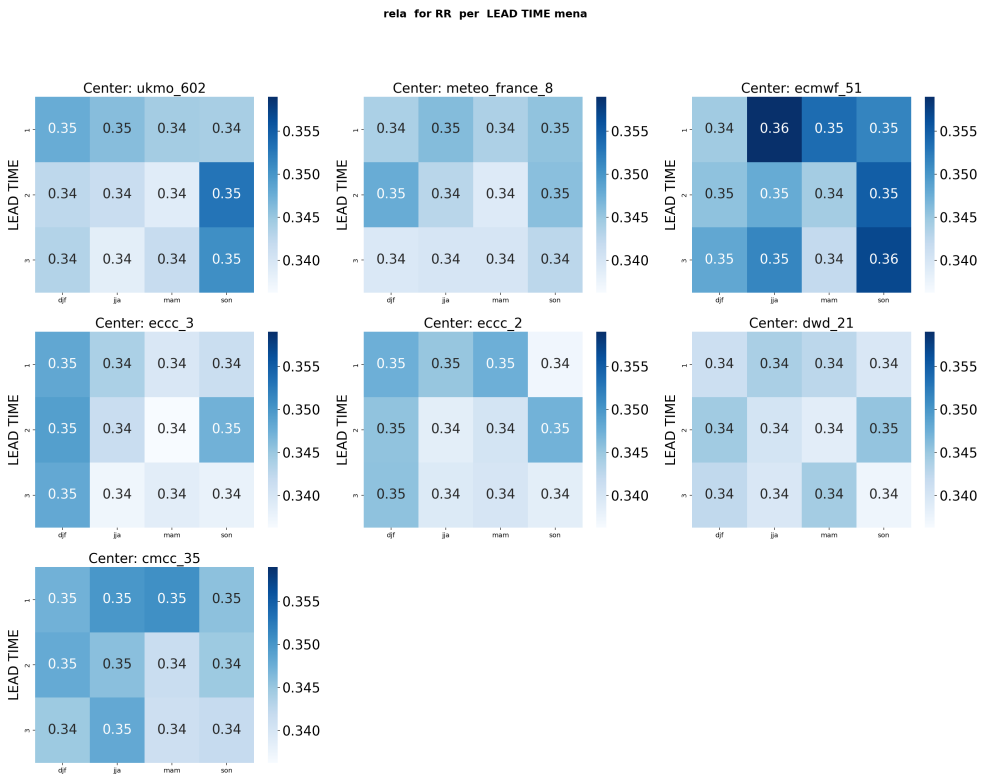

Figure 5.50: The Reliability Score . ***(0 means perfect Reliability)***

In the figure above, all centers demonstrate similar moderate performance in term of reliability. But deep analysis within the figure below, shows that all centers give similar description of the reliability, also the stability along lead-time is a good indicator despite of the moderate results (0.3), we can rely on all models because of the acceptable results and the stability along time.

focus on North Africa :

there is no big difference in North Africa.

focus on Arabian Peninsula :

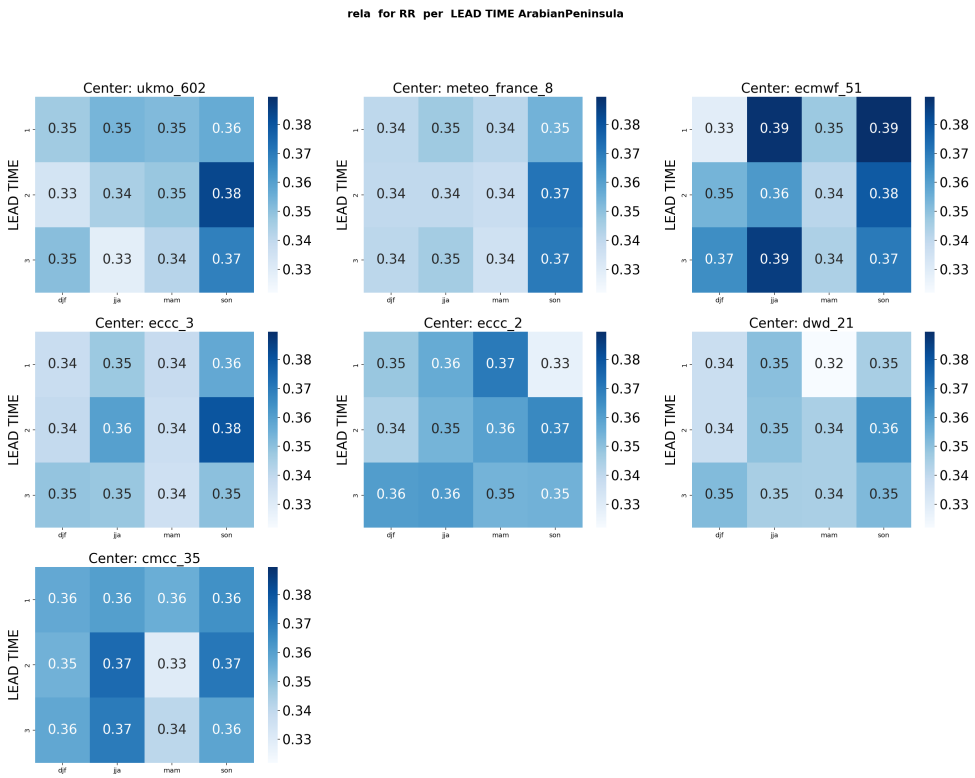

Figure 5.51: The Reliability Score Arabian Peninsula . ***(0 means perfect Reliability)***

There is a little Decline in the reliability of the Arabian Peninsula, especially f***ECMWF***

The ranked probability score (RPS)

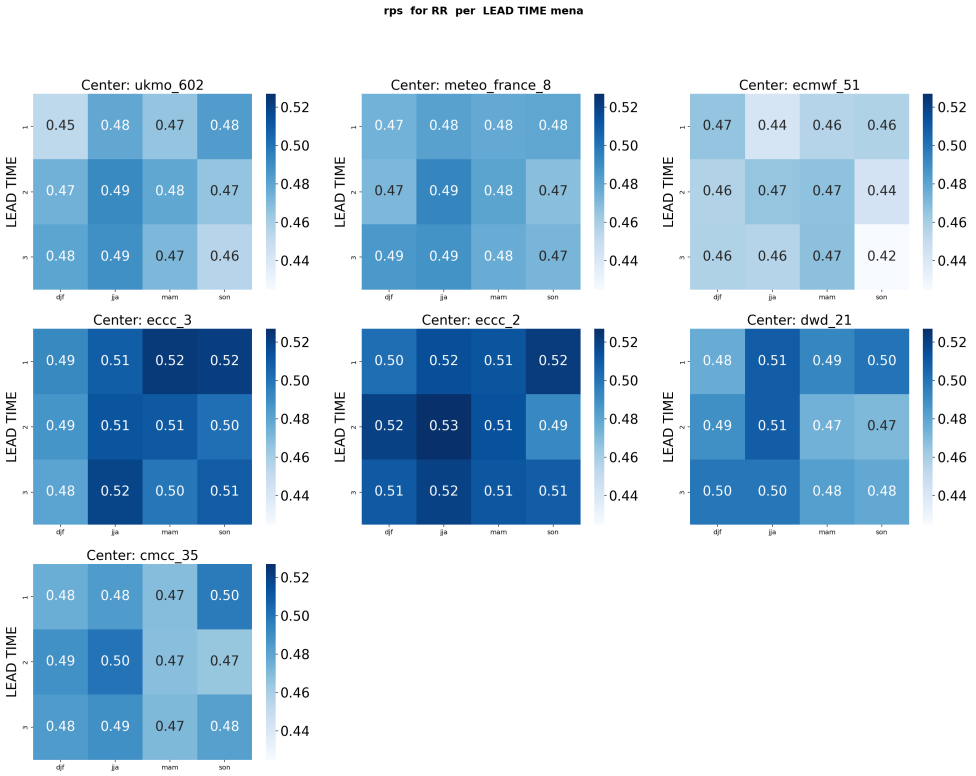

Figure 5.52: The Heatmap of RPS Score on MENA region for Precipitations . ***(0 means perfect RPS)***

In the figure above, all centers demonstrate moderate performance, except for ***ECCC***, which shows noticeably lower performance.

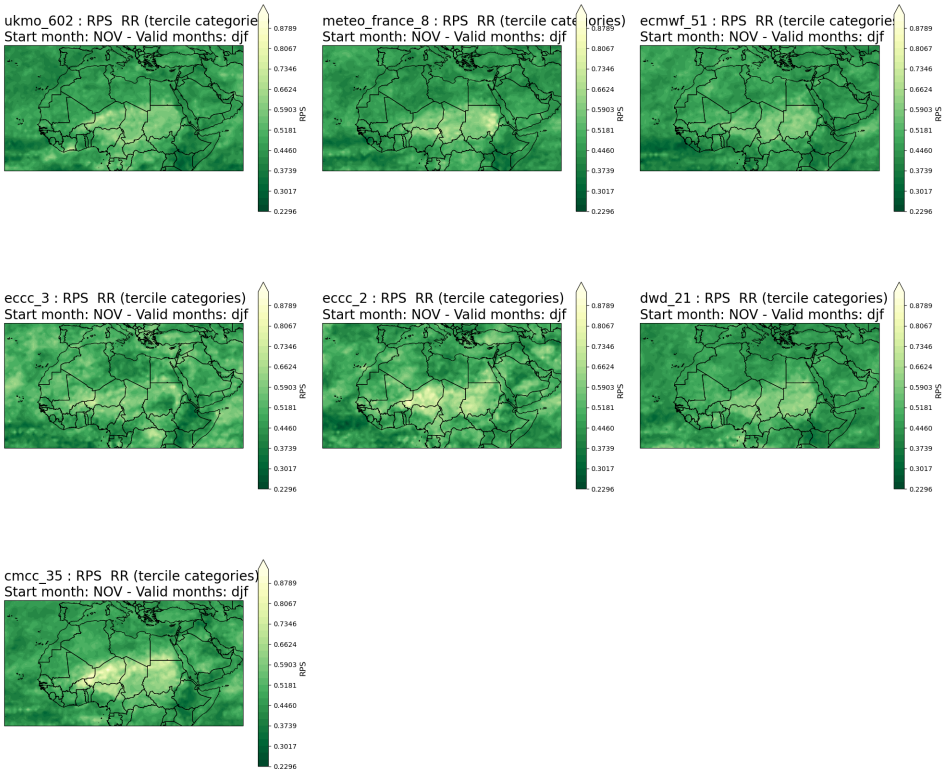

Figure 5.53: The RPS Score on MENA region for Precipitations DJF . ***(0 means perfect RPS)***

the spacial distribution of the RPS, is homogeneous, in all the region the score is good for all centers.

focus on north africa: there is no big difference in North Africa.

focus on Arabian Peninsula :
there is no big difference on Arabian Peninsula.

Relative operating characteristics

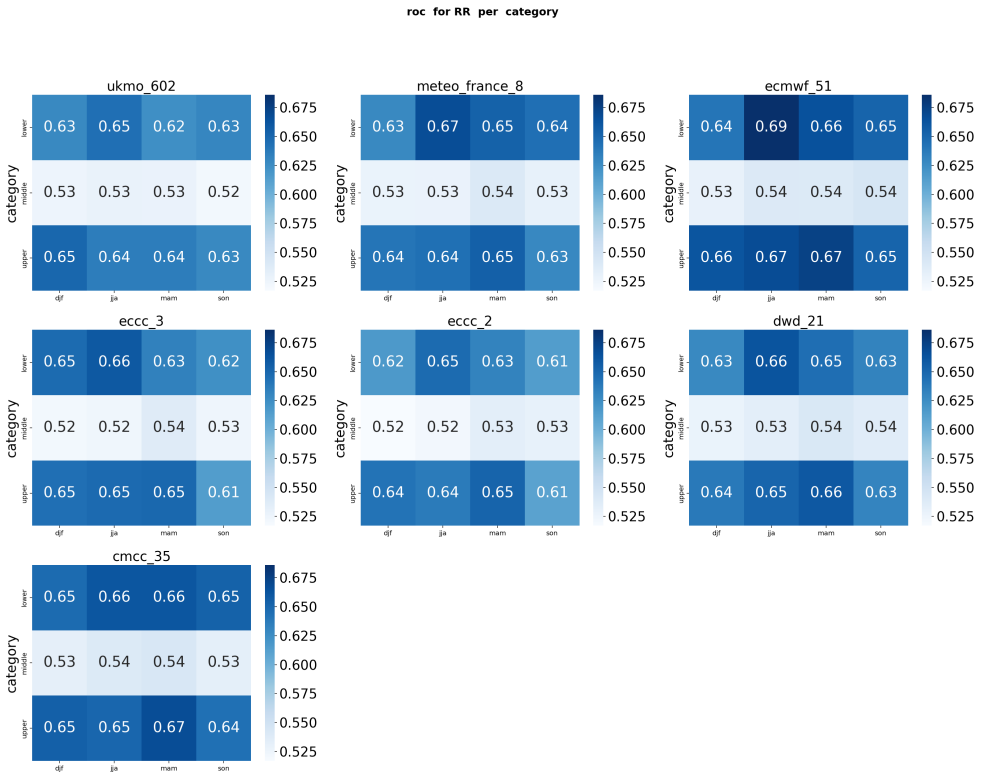

Figure 5.54: The Heatmap of ROC Score for each category . ***(1 means perfect ROC)***

In the figure above, it is clear that all forecasting centers show similar performance levels overall. However, the Middle Tercile consistently yields the lowest scores across all models. This finding is significant as it underscores the ability of all models to effectively predict extreme events, both above and below normal. In contrast, the models struggle more with predicting normal situations, which are inherently more difficult to forecast. The lower performance in the Middle Tercile highlights the challenge of accurately predicting conditions that are neither extreme nor easily distinguishable from typical climate variability. This suggests that while the models are proficient in forecasting extreme events, they face greater challenges when predicting more moderate, "normal" conditions.

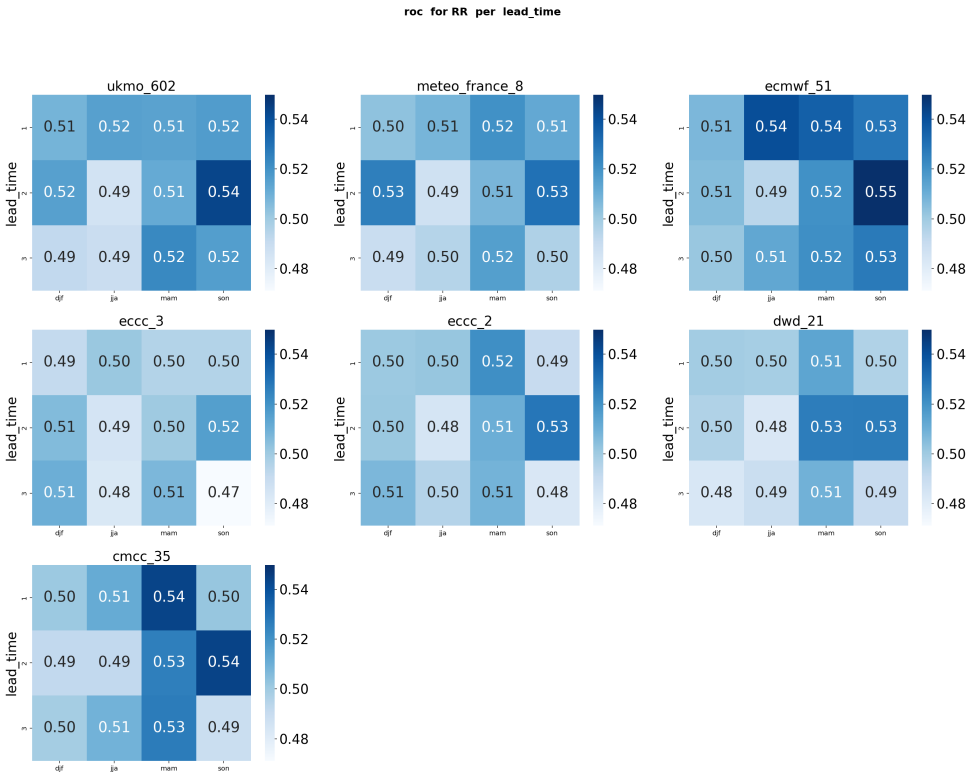

Figure 5.55: The Heatmap of ROC Score for lead-times. ***(1 means perfect ROC)***

For the analysis per lead-time , all forecasting centers demonstrate similar and generally good performance. In most cases, the highest scores are observed for the first lead-time. However, an exception occurs during the SON (September-October-November) season, where the best performance is seen at the second lead-time. This variation suggests that while the models are consistently strong in the first lead-time, certain seasonal conditions, like those in SON, may benefit from a slightly longer forecast horizon to capture the full dynamics of the climate

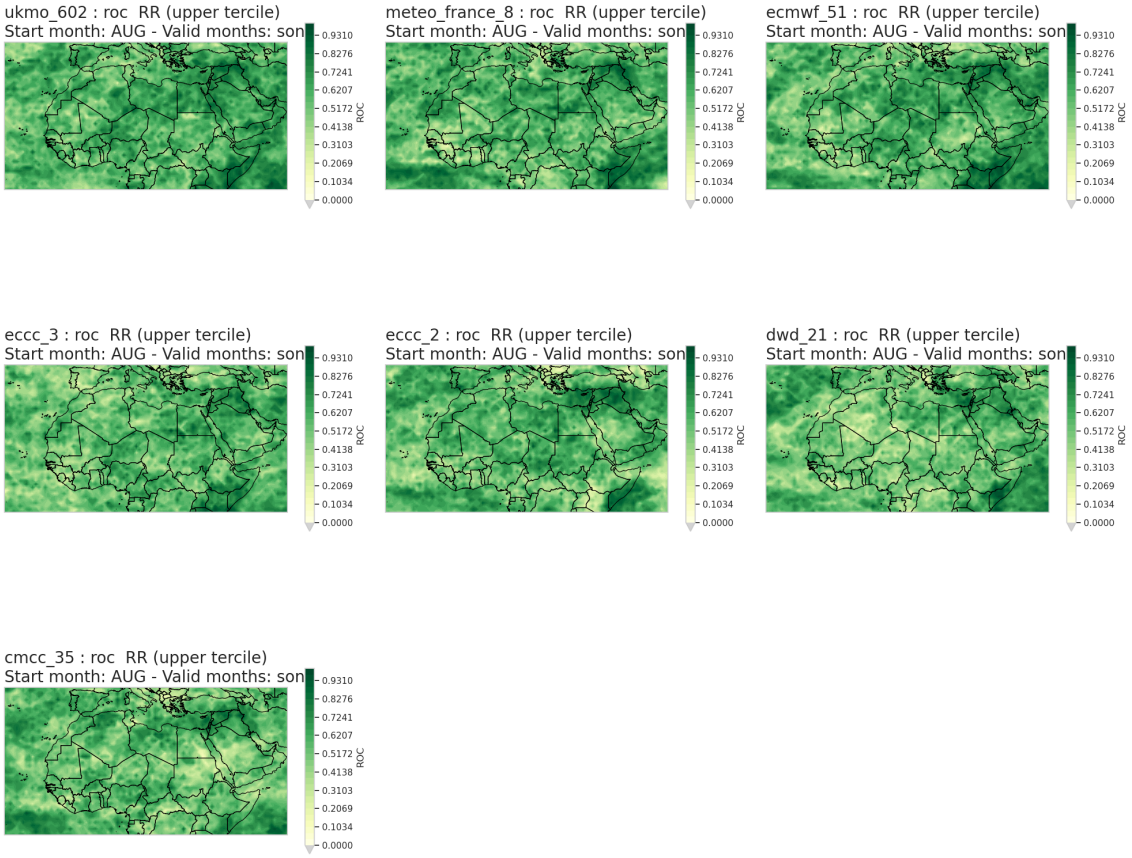

Figure 5.56: The ROC Score Upper tercile SON . ***(1 means perfect ROC)***

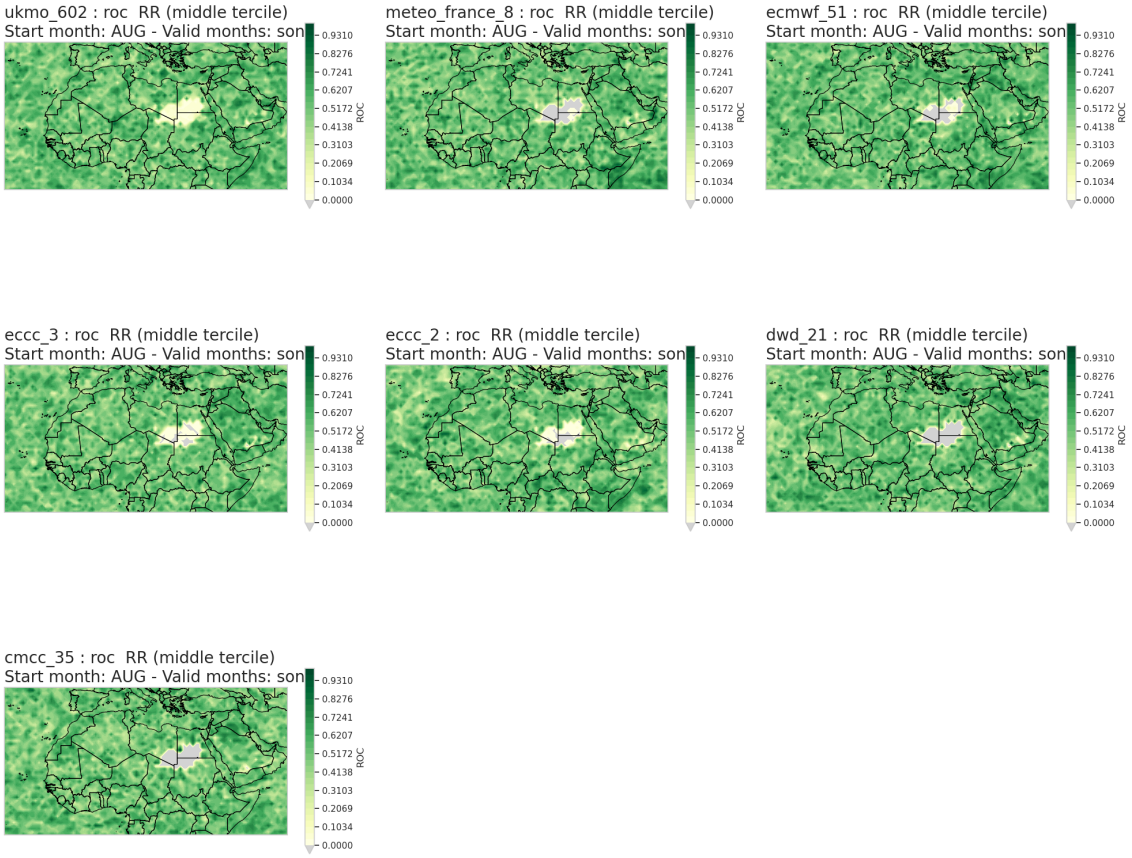

Figure 5.57: The ROC Score Middle tercile SON . ***(1 means perfect ROC)***

The spatial distribution of the ROC score highlights an important finding. For precipitation, all centers demonstrate better performance in predicting the upper tercile, with similar results observed for the lower tercile. However, the performance in the middle tercile is considerably weaker, with high variability and lower scores, particularly in regions such as Egypt and Libya, where values are notably low. In this tercile, no significant differences are observed between the models, indicating a consistent struggle across all forecasting systems in dealing with moderate or normal precipitation events.

In contrast, the upper tercile displays some variability in the spatial distribution of scores, with the best performance seen in the Arabian Peninsula and parts of East Africa, particularly for the ***ECMWF and UKMO*** models. This result underlines the models' relative strength in forecasting extreme precipitation events while emphasizing the challenges they face in predicting more moderate conditions.

focus on north africa: there is no big difference in North Africa.

focus on Arabian Peninsula :

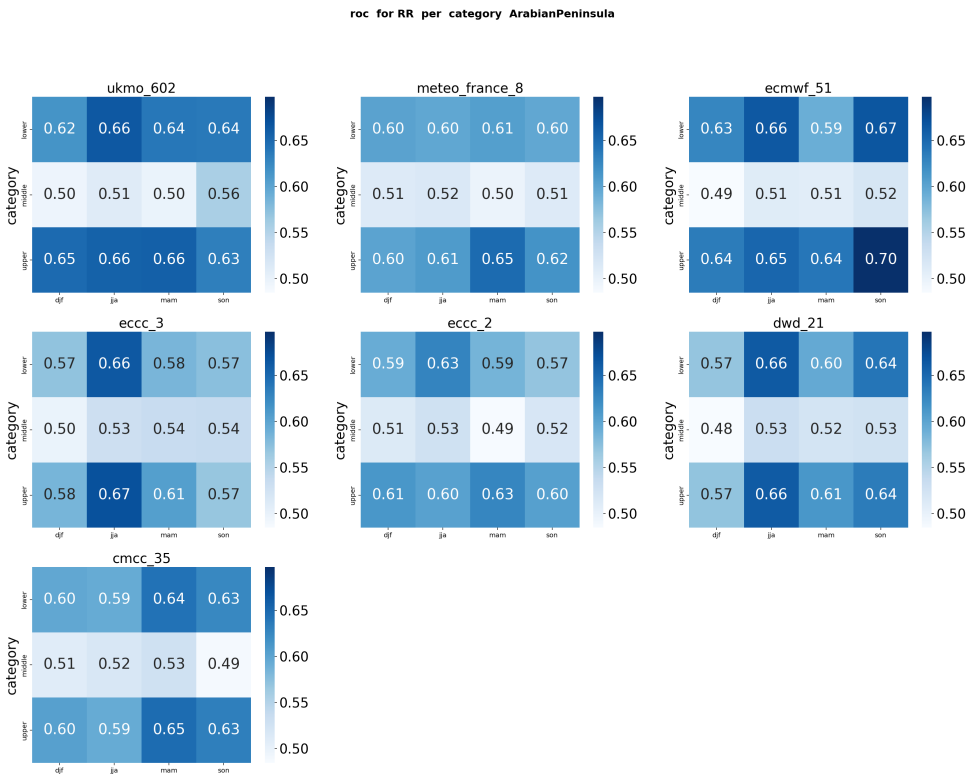

Figure 5.58: The ROC Score for each category Arabian Peninsula . ***(1 means perfect ROC)***

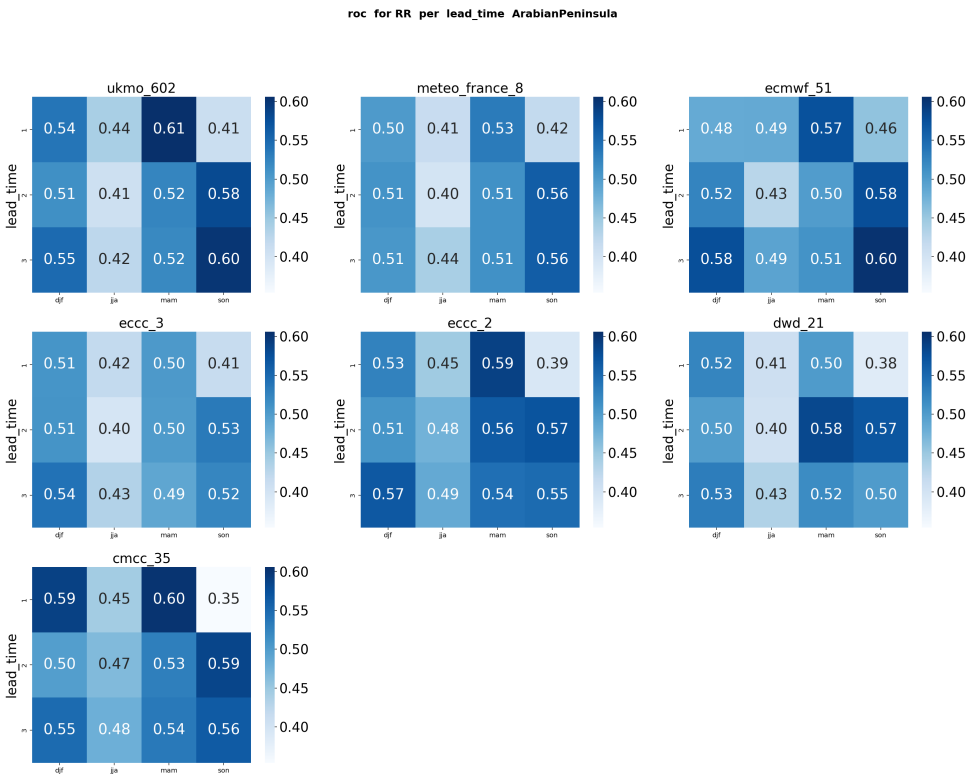

Figure 5.59: The average of ROC Score on all categories for Arabian Peninsula . ***(1 means perfect ROC)***

For the roc score, the focus on Arabian Peninsula in category, show no big difference, as for the analysis along lead-time, there is a few improvement for DJF, MAM and SON, instead of JJA that shows low values for all centers.

Relative operating characteristics Skill Score

In the figure above, the ECMWF exhibit the best performance for all terciles and periods. However, we should notice that the performance is very low for the middle tercile in all centers. For the analysis along time, the performance is so low, the best performance is in the first lead-time, except the SON which shows the best performance for the second lead-time. Above all, the ***ecmwf*** shows the best performance.

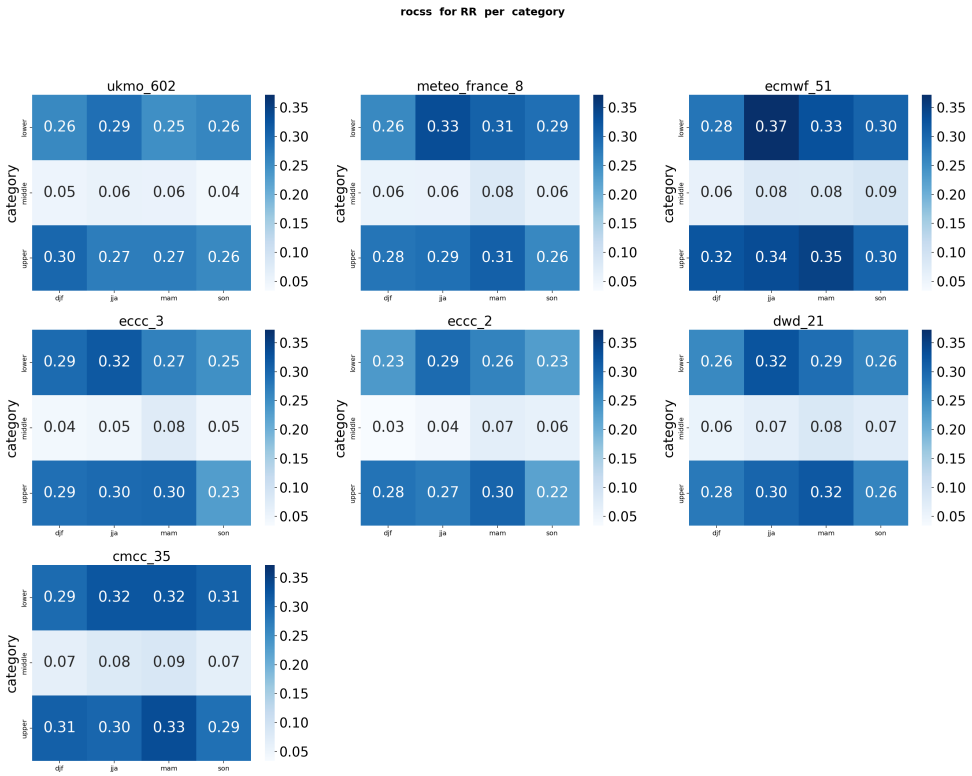

Figure 5.60: The ROCSS Score for each category . ***(1 means perfect ROCSS)***

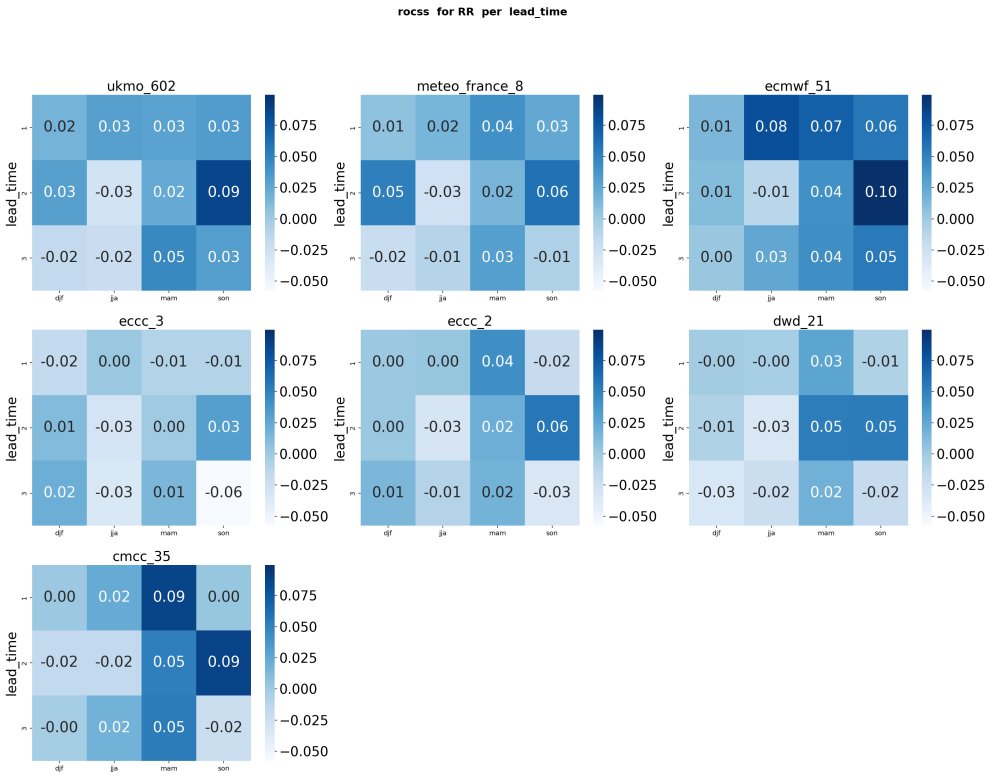

Figure 5.61: The average of ROCSS Score on all categories . ***(1 means perfect ROCSS)***

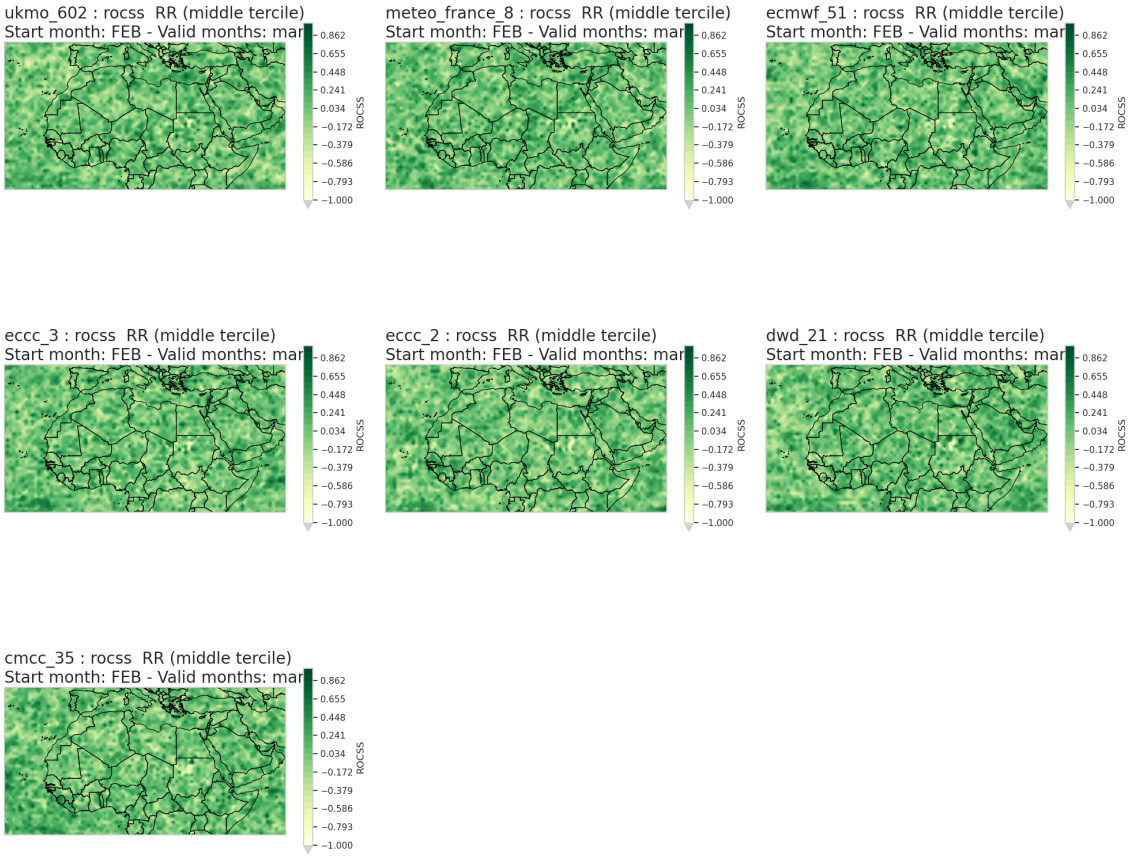

Figure 5.62: The ROC Skill Score Middle tercile MAM . ***(1 means perfect ROCSS)***

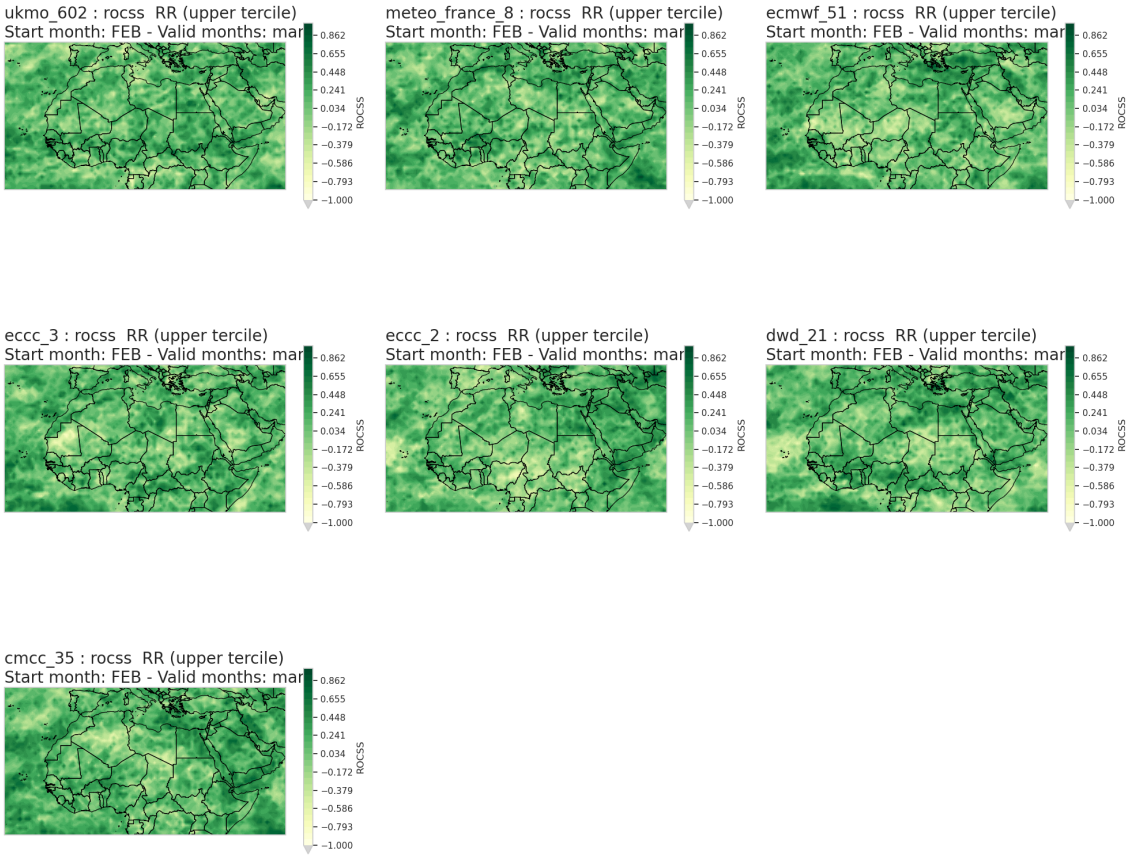

Figure 5.63: The ROC Skill Score Upper tercile MAM . ***(1 means perfect ROCSS)***

The spatial distribution of the ROCSS indicates consistent performance across all forecasting centers for this score. However, the spatial patterns are not well-defined for both the upper and middle terciles. Notably, the middle tercile exhibits high spatial variability, with generally lower values compared to the other terciles. These findings align with the previously observed results for the ROC score, further confirming that forecasting performance is weaker and more inconsistent for the middle tercile. This underscores the ongoing challenge of accurately predicting moderate conditions, while extreme events are handled with greater reliability.

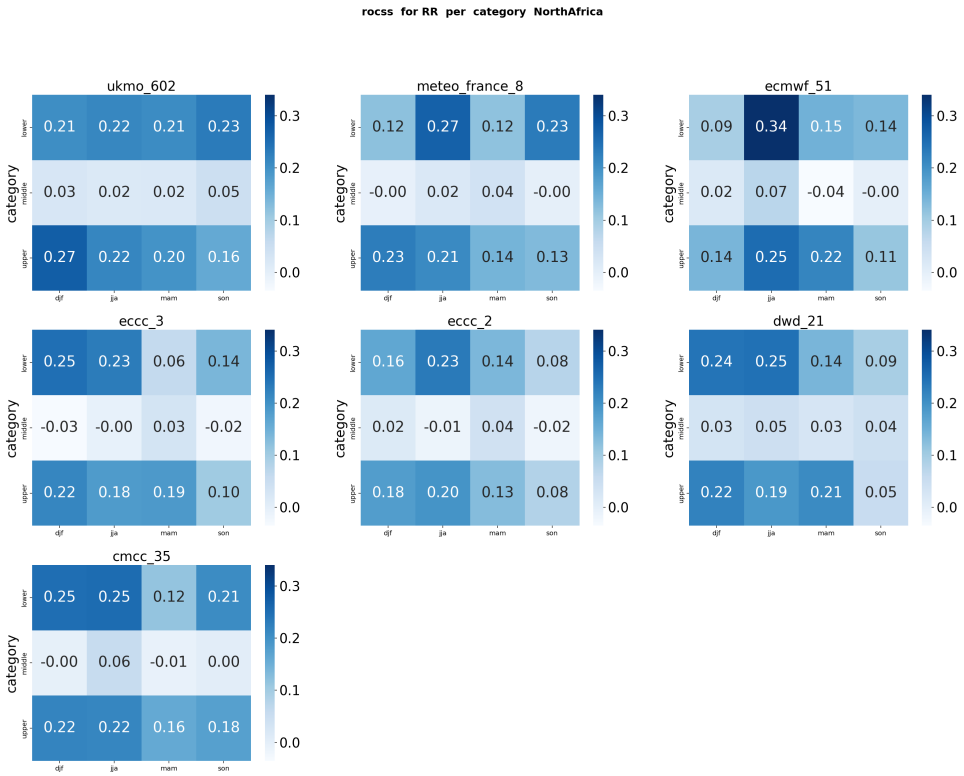

Figure 5.64: The ROCSS Score for each category North Africa . ***(1 means perfect ROCSS)***

Figure 5.65: The average of ROCSS Score on all categories North Africa . ***(1 means perfect ROCSS)***

focus on north africa: the rocss for North Africa is in general lower, thus the performance is less accurate.

focus on Arabian Peninsula :

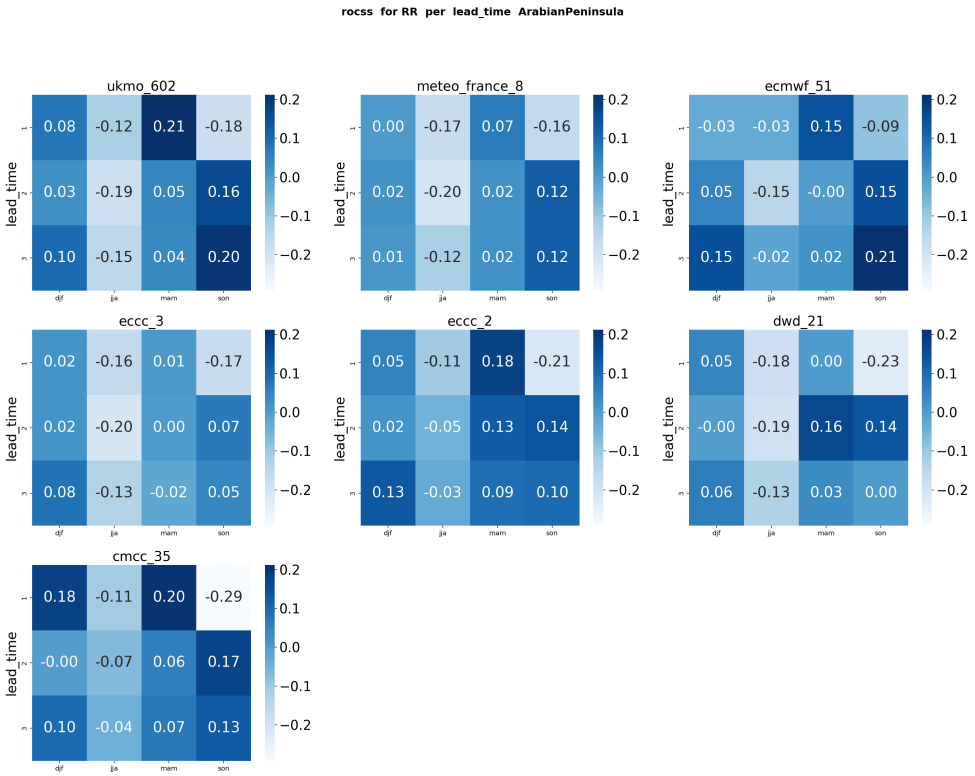

Figure 5.66: The average of ROCSS Score on all categories for Arabian Peninsula . ***(1 means perfect ROCSS)***

the rocss for Arabian Peninsula is in general lower, thus the performance is less accurate.

summary

Metric	Focus	What it Measures	Dependent on Observed Outcomes?	Visualization/Tools
Reliability	Probabilities match observed frequencies	Calibration of probabilities	Yes	Reliability diagram
Discrimination	Differentiating between outcomes	Ability to distinguish events from non-events	Yes	ROC curve, AUC
Sharpness	Boldness of probabilities (away from average)	Confidence of the forecast	No	Histogram of forecast probabilities
Resolution	Informativeness and variability of forecast	Ability to provide specific, useful info	Yes	Brier Score decomposition

Table 5.1: Key differences between reliability, discrimination, sharpness, and resolution in seasonal forecasting.

CHAPTER 6

CONCLUSION

This study has thoroughly evaluated the performance of leading climate models in predicting temperature and precipitation across the MENA region, North Africa, and the Arabian Peninsula. By assessing key metrics such as Accuracy (ACC), Root Mean Square Error (RMSE), Coefficient of Determination (R^2), Brier Score (BS), ROC score, and others we have identified the strengths and weaknesses of each model, while highlighting regional variations in their predictive capabilities.

For temperature, all climate models demonstrated generally similar performance in the MENA region, with ECMWF and UKMO consistently emerging as the most reliable models, achieving top performance in both deterministic and probabilistic metrics across all regions.

For precipitation, all models performed less effectively compared to their temperature predictions, particularly in the MENA region. ECMWF and UKMO once again demonstrated the highest performance across all metrics and regions, showcasing its robustness in handling diverse climatic conditions. METEO-FRANCE, CMCC and DWD exhibited similar performance in deterministic measures such as ACC and RMSE. However, in probabilistic metrics, most models showed approximately similar performance, particularly in the Arabian Peninsula, where challenges related to the region's hyper-arid climate and limited observational data are evident.

It is worth noting that the models perform better for temperature predictions than for precipitation in the MENA region. This contrasts with tropical regions, where models tend to excel in both precipitation and temperature forecasting due to the more predictable influence of large-scale climate drivers such as ENSO. The limited observational data and complex climate dynamics in arid and semi-arid regions of MENA further challenge precipitation predictions, reducing model accuracy compared to temperature forecasts.

In conclusion, climate models generally exhibit similar performance across the four classical seasons, with ECMWF and UKMO consistently demonstrating higher performance compared to the other models. Each model provides valuable complementary strengths depending on the season and region. These findings highlight the importance of using multi-model ensembles to leverage the strengths of each model, improve seasonal forecast accuracy, and correct the influence of prominent modes of climate variability, such as ENSO and NAO, in the climate models using state-of-the-art statistical techniques.

Metric	**MENA**	**North Africa**	**Arabian Peninsula**
ACC	ECMWF, METEO-FRANCE, ECCC-3	ECMWF, UKMO, ECCC-3	ECMWF, UKMO, DWD
RMSE	UKMO, ECMWF	METEO-FRANCE	METEO-FRANCE
R^2	ECMWF	ECMWF	ECMWF
BS	METEO-FRANCE, ECMWF, CMCC-35, UKMO	ECMWF, UKMO, CMCC-35	ECMWF, UKMO, CMCC-35
RELA	ALL	ALL	ALL
RPS	ECMWF	ECMWF	ECMWF
ROC	ALL	ALL	ALL
ROCSS	ALL	ALL	ALL

Table 6.1: Comparison of Metrics across MENA, North Africa, and Arabian Peninsula for TEMPERATURE.

Metric	**MENA**	**North Africa**	**Arabian Peninsula**
ACC	ECMWF, CMCC-35, UKMO	ECMWF, UKMO and METEO-FRANCE	ECMWF, CMCC-35, UKMO
RMSE	DWD, ECMWF and UKMO	ECMWF, UKMO and DWD	ECMWF, UKMO and DWD
R^2	ECMWF	ECMWF	CMCC-35, ECCC2
BS	ECMWF, METEO-FRANCE and CMCC-35	ECMWF, METEO-FRANCE and CMCC-35	ECMWF, METEO-FRANCE and CMCC-35
RELA	ECMWF, CMCC and UKMO	ECMWF, CMCC-35 and UKMO	METEO-FRANCE, DWD
RPS	ALL	ALL	ALL
ROC	ALL	ALL	ALL
ROCSS	ECMWF	ECMWF	UKMO, CMCC-35

Table 6.2: Comparison of Metrics across MENA, North Africa, and Arabian Peninsula for PRECIPITATION

List of Figures

List of Tables