Sensitivity of precipitation forecasts in tropical Africa to available observations – an idealized study using the TEEMLEAP testbed

Master's Thesis in
Meteorology and Climate Physics
by

Cédric Froidevaux

December 2025

INSTITUTE OF METEOROLOGY AND CLIMATE RESEARCH
KARLSRUHE INSTITUTE OF TECHNOLOGY (KIT)

Supervisor:	Prof. Dr. Peter Knippertz
Co-supervisor:	Prof. Dr. Andreas Fink

Abstract

Reliable weather forecasting remains a major challenge in tropical Africa due to sparse observational networks and limited understanding of key atmospheric processes. This thesis investigates how the availability and spatial density of observations influence the skill of numerical weather prediction (NWP), with a particular focus on precipitation. To quantify these effects, a series of controlled observing-system experiments is conducted using the TEEM-LEAP testbed, which simulates the full forecasting chain based on radiosonde-like pseudo-observations (PSOs) derived from ERA5 and forecasts generated using the ICOsahedral Non-hydrostatic (ICON) model. Six idealized global and Africa-focused observational networks are tested for September 2022. Forecast performance was evaluated over tropical Africa and, for comparison, Europe.

For standard meteorological variables (mean sea level pressure, temperature at 850 hPa, total column water vapour, and winds at multiple levels), increasing the number of PSOs systematically improved forecast quality. Reductions in root mean square error (RMSE) reach up to 30 % at short lead times, with gains decreasing as lead time increases — approximately exponentially in tropical Africa and more linearly in Europe. In tropical Africa, the additional benefit of a very dense PSO network (spacing of roughly 480 km) over a medium-dense configuration becomes marginal after about four days, whereas in Europe substantial differences between these experiments persist throughout the entire seven-day forecast horizon.

For precipitation, comparisons between IMERG and ERA5 data reveal substantial differences, particularly in mountainous and coastal regions, underscoring uncertainty in the available verification data. ICON forecasts exhibit a dry bias at short lead times, which increases with PSO density, potentially reflecting inconsistencies introduced by pushing the ICON background toward the ERA5 state. Despite these limitations, verification with the Stable Equitable Error in Probability Space (SEEPS) and the Fractions Skill Score (FSS) indicates modest but robust improvements in tropical Africa (typically 1–3 % overall, up to 30 % for heavy rainfall), and significantly larger improvements in Europe. Regionally focused PSO experiments further show that, while dense local observations over tropical Africa enhance 1–2 day forecasts, accurate information from the broader tropical belt becomes essential thereafter.

Overall, the results demonstrate that enhanced observational coverage improves forecast skill, but the magnitude and persistence of these improvements differ markedly between tropical and extratropical regions. In tropical Africa, forecast improvements generally saturate more quickly than in Europe. The results indicate that a spacing between PSOs on the order of 700 km is sufficient to constrain large-scale tropical waves that dominate medium-range predictability, while additional observational density provides limited benefit because smaller-scale errors are governed by model and parametrization uncertainties. These findings show that enhanced observational coverage improves forecast skill in both regions but that in tropical Africa further progress will require not only denser observations but also advances in model physics and the representation of convection.

Zusammenfassung

Zuverlässige Wettervorhersagen bleiben in tropischen Regionen Afrikas eine große Herausforderung, was vor allem auf die geringe räumliche Dichte von Beobachtungsstationen und ein unzureichendes Verständnis wichtiger atmosphärischer Prozesse zurückzuführen ist. Diese Arbeit untersucht, wie die Verfügbarkeit und räumliche Dichte von Beobachtungsdaten die Vorhersagegüte numerischer Wettermodelle beeinflussen, mit besonderem Fokus auf Niederschlag. Hierzu werden eine Reihe Experimente im TEEMLEAP-Testbed durchgeführt, das die gesamte Kette eines operationellen Vorhersagesystems simuliert und radiosondenähnliche Pseudo-Beobachtungen (PSOs) aus ERA5 sowie Vorhersagen mit dem ICOsahedral Nonhydrostatic (ICON) Modell nutzt. Insgesamt werden sechs idealisierte globale und Afrikaspezifische Beobachtungsnetzwerke für September 2022 getestet. Die Auswertung erfolgte für das tropische Afrika sowie zum Vergleich für Europa.

Für grundlegende meteorologische Variablen (Bodendruck, Temperatur auf 850 hPa, integrierter Wasserdampf und Windfelder in mehreren Höhen) zeigt sich, dass eine höhere PSO-Dichte systematisch zu verbesserten Vorhersagen führt. Die Reduktion des root mean square error (RMSE) erreicht bis zu 30 % für kurze Vorhersagezeiträume und nimmt mit zunehmender Vorhersagedauer ab — im tropischen Afrika etwa exponentiell, in Europa eher linear. Im tropischen Afrika ist der zusätzliche Nutzen eines sehr dichten PSO-Netzwerks (mit einem Abstand von etwa 480 km) gegenüber einer mitteldichten Konfiguration nach etwa vier Tagen gering. In Europa hingegen halten deutliche Unterschiede zwischen diesen zwei Experimenten über den gesamten siebentägigen Vorhersagezeitraum hinweg an.

Der Vergleich zwischen IMERG- und ERA5-Niederschlagsdaten zeigt erhebliche Unterschiede, insbesondere in Gebirgsregionen und in Küstennähe, was die Unsicherheiten der verfügbaren Verifikationsdaten unterstreicht. ICON-Vorhersagen weisen für kurze Vorhersagezeiten einen Trockenbias auf, der mit zunehmender PSO-Dichte stärker ausgeprägt ist — vermutlich infolge von Inkonsistenzen, die durch das stärkere Heranführen des ICON-Hintergrunds an den ERA5-Zustand entstehen. Trotz dieser Einschränkungen zeigen sowohl der Stable Equitable Error in Probability Space (SEEPS) und der Fractions Skill Score (FSS) leichte, aber robuste Verbesserungen im tropischen Afrika (typisch 1–3 %, bis zu 30 % bei Starkniederschlag), während die Verbesserungen in Europa deutlich größer ausfallen. Regionale PSO-Konfigurationen zeigen zudem, dass dichte lokale Beobachtungsnetze über dem tropischen Afrika zwar kurzfristige 1–2-Tage-Vorhersagen verbessern, für längerfristige Zeiträume hingegen Informationen aus dem gesamten tropischen Gürtel entscheidend werden.

Insgesamt zeigen die Ergebnisse, dass eine verbesserte Beobachtungsabdeckung die Vorhersagegüte erhöht, der Nutzen jedoch regional sehr unterschiedlich ausfällt. In den Tropen treten Sättigungseffekte deutlich früher auf als in Europa. Die Resultate deuten darauf hin, dass ein PSO-Abstand in der Größenordnung von 700 km ausreicht, um die großskaligen tropischen Wellen zu erfassen, welche die mittelfristige Vorhersagbarkeit dominieren, während zusätzliche Beobachtungsdichte nur begrenzte Vorteile bietet, da Fehler auf kleineren Skalen stärker durch Modell- und Parametrisierungsunsicherheiten geprägt sind. Diese Ergebnisse verdeutlichen, dass zwar in beiden Regionen eine dichtere Beobachtungsabdeckung die Vorhersagequalität verbessert, im tropischen Afrika jedoch weitere Fortschritte nicht nur durch mehr Beobachtungen, sondern auch durch Fortschritte in der Modellphysik und der Darstellung konvektiver Prozesse erzielt werden müssen.

Contents

1 Introduction

The accuracy of numerical weather prediction (NWP) has improved remarkably over recent decades, bringing substantial benefits to society and the economy (Bauer et al. 2015). Yet these advances have not been distributed evenly across the globe. In tropical Africa, reliable weather forecasting remains a major challenge due to limited scientific understanding of local weather systems, sparse observational networks, limited data availability, poor model performance, and challenges in forecast communication (Lamptey et al. 2024; Parker et al. 2022).

The weather in the tropics differs substantially from that in mid-latitudes. While the extratropics are dominated by high and low pressure systems associated with baroclinic instability, tropical weather is primarily governed by convective processes and equatorial wave dynamics, including Kelvin, mixed Rossby–gravity, equatorial Rossby, and inertio-gravity waves (Sobel 2012; Kiladis et al. 2009; Knippertz et al. 2022). This fundamental difference in governing dynamics leads to distinct patterns of intrinsic predictability. As shown by Judt (2020) and Keane et al. (2025), error growth is strongly latitude-dependent: initial error growth is largest in the tropics, whereas error growth in the extratropics accelerates after the first few days, eventually leading to a faster loss of predictability outside the tropics. The resulting picture — longer intrinsic predictability in the tropics — stands in contrast to practical NWP experience and highlights a discrepancy between intrinsic and practical predictability. An incomplete understanding of tropical weather systems continues to limit their representation in NWP models, contributing to poor forecast performance, particularly for precipitation (Lamptey et al. 2024).

Despite long-standing recognition of the need for improved observational coverage, progress in Africa has been modest. Many countries still lack a dense enough network of regularly reporting weather stations, and access to existing data is often restricted by political, financial, or logistical barriers (IPCC 2023; Lamptey et al. 2024). Observations are vital for data assimilation, model verification, postprocessing, and advancing process understanding. Figure 1.1 illustrates the global distribution of radiosonde observations assimilated by the European Centre for Medium-Range Weather Forecasts (ECMWF) on a representative day. The sparse coverage across Africa highlights an important constraint on forecast quality in this region. In addition, emerging data-driven forecast approaches hold promise for enhancing weather predictions, but their success also depends critically on the quality and accessibility of observational data.

Beyond challenges related to data coverage and model performance, Africa is particularly vulnerable to climate change and variability. Increasingly frequent and intense droughts, heatwaves, and heavy precipitation events are already being observed (IPCC 2023). These events are expected to intensify in a warming climate, with severe consequences for population, infrastructure, and ecosystems, including floods and wildfires. Disruptions of the hydrological cycle pose a serious threat to food security and livelihoods, disproportionately affecting vulnerable communities. To address this, the United Nations (UN) and the World Meteorological Organization (WMO) launched the initiative “Early Warnings for All”, which aims to ensure universal access to early warning systems for climate hazards by 2027 (UN 2025). With

Africa's population projected to exceed four billion by 2100 (UN 2024), the demand for reliable weather forecasts is greater than ever. In sub-Saharan Africa, 55–62 % of the workforce is employed in agriculture, with roughly 95 % of croplands relying on rainfall rather than irrigation (IPCC 2023), making the region highly sensitive to changing weather patterns.

Given these challenges, improving weather prediction in Africa is essential, particularly for high-impact variables such as precipitation. Several approaches can help enhance forecast skill:

- Statistical postprocessing to correct systematic biases in NWP models (Vogel et al. 2020)
- Increasing model resolution to convection-permitting scales and reducing reliance on simplified parametrizations (Pante and Knippertz 2019)
- Optimizing uncertain model parameters (e.g., entrainment rates) to improve performance over tropical Africa (Fischer 2025)
- Expanding observational networks to provide more and higher-quality data for data assimilation (van der Linden et al. 2020; Borne et al. 2023)
- Applying machine learning techniques to improve model performance (Walz et al. 2024b; Rasheeda Satheesh et al. 2025).

This study investigates the influence of observational data availability on weather forecast accuracy using the TEstbed for Exploring Machine LEarning in Atmospheric Prediction (TEEMLEAP) developed by Wilhelm et al. (2025). The TEEMLEAP framework simulates the entire operational forecasting chain from observations through data assimilation to model predictions and verification. Unlike operational NWP systems that rely on real-world observations, the TEEMLEAP testbed uses pseudo-observations (PSO) derived from ERA5 reanalysis, simulating radiosonde measurements. This enables controlled experiments that systematically test how different observation configurations, including both global and Africa-specific setups, affect forecast accuracy.

Compared to field campaigns, such as African Monsoon Multidisciplinary Analysis (AMMA) (Redelsperger et al. 2006; Agustí-Panareda et al. 2010; Faccani et al. 2009) and Dynamics–Aerosol–Chemistry–Cloud Interactions in West Africa (DACCIWA) (Knippertz et al. 2017; van der Linden et al. 2020), the testbed offers greater flexibility and substantially lower cost. It allows researchers to design and evaluate idealized observational networks without the logistical and political constraints of field campaigns. The main limitation is that no real observations are used. While ERA5 provides one of the best available global atmospheric states, combining ERA5-based data (from ECMWF's Integrated Forecasting System, IFS) to initialize forecasts with the ICOsahedral Nonhydrostatic (ICON) model of the German Weather Service (DWD) may introduce inconsistencies between the two model frameworks. Nevertheless, this controlled setup provides valuable insights into how enhanced observational systems could improve forecast skill.

This work is designed to deliver useful insights to underpin international initiatives that aim to address the lack of observations. The Systematic Observations Financing Facility (SOFF), led by the UN and WMO (SOFF 2025), helps countries close critical gaps in weather and climate observations, focusing on least developed countries and small island states. ECMWF recently published a SOFF impact study that underscores the importance of enhanced observations

(ECMWF 2025b), using a similar approach based on simulated observation experiments. Another promising initiative is the Trans-African Hydro-Meteorological Observatory (TAHMO) (TAHMO 2025), which aims to establish a dense network of hydro-meteorological stations across sub-Saharan Africa — approximately one every 30 km.

The main focus of this study is tropical Africa, with Europe included for comparison to investigate differences in predictability between tropical and extratropical regimes (Judt 2020; Keane et al. 2025). Standard meteorological variables will first be evaluated, including mean sea level pressure (MSLP), temperature at 850 hPa (T850), total column water vapour (TCWV), zonal and meridional winds at 600 hPa (U600, V600) and zonal wind at 250 hPa (U250). The analysis will then focus on precipitation, a key variable for both regions. The simulation period covers September 2022, with forecasts generated using the ICON model for lead times of up to seven days, consistent with the setup of Wilhelm et al. (2025).

Figure 1.1: Global coverage of radiosonde observations assimilated by ECMWF at 00 UTC on 30 October 2025 (ECMWF 2025a). A total of 687 radiosondes were used, but the density of observations varies significantly across regions. The notably sparse coverage over tropical Africa highlights a key challenge for forecast accuracy in this region.

The overarching aim of this work is to quantify the sensitivity of weather forecasts to observational coverage, addressing the following research questions:

1. How sensitive are forecasts in tropical Africa to the availability of observations?

2. Which spatial scales need to be resolved and is there a point of diminishing returns beyond which additional observations provide little benefit?

3. What might an optimal observational network for tropical Africa look like?

The thesis is structured as follows: Chapter 2 introduces the theoretical background, Chapter 3 reviews the current state of research, and Chapter 4 describes the datasets, models, the TEEM-LEAP testbed, and verification metrics. Chapter 5 details the experimental setup, followed by the results in Chapter 6 and the conclusions in Chapter 7.

2 Fundamentals and Background

This chapter provides an overview of the mean climate and key synoptic features of tropical Africa, alongside an introduction to the basics of NWP.

2.1 Mean Climate and Synoptics in Tropical Africa

The atmospheric dynamics of tropical regions differ fundamentally from those at higher latitudes for two main reasons: First, the tropics are characterized by consistently high levels of solar radiation throughout the year, resulting in elevated temperatures. Second, at the equator, the vertical direction is nearly perpendicular to the Earth's axis of rotation, whereas at higher latitudes the vertical aligns more closely with the rotational axis. This geometry reduces the influence of the Coriolis force in tropical regions, compared to the stronger rotational effects at mid- and high latitudes (Sobel 2012).

Consequently, cloud formation and precipitation in the tropics occur more spontaneously and are less constrained by large-scale atmospheric dynamics than in the extratropics. While this increases the complexity of forecasting rainfall, temperature prediction is relatively straightforward due to the limited seasonal variability (Sobel 2012).

These fundamental differences manifest in several distinctive synoptic features across tropical Africa, shaping rainfall patterns and seasonal variability. Climatological MSLP and 10 m wind patterns over the region defined as tropical Africa in this study (10° S–25° N, 20° W–50° E) for September are shown in Figure 2.1.

Saharan Heat Low (SHL)

Thermal lows are characteristic climatological features over many desert and semi-arid regions during the warm season, especially at low latitudes where solar radiation is strong. One prominent example is the Saharan Heat Low (SHL), which shapes both diurnal and synoptic circulations across North and West Africa. The SHL is identified as a broad zone of low surface pressure and elevated near-surface temperature centred over the western Sahara during boreal summer. It arises primarily from strong surface heating of the desert, which deepens the boundary layer and produces a shallow cyclonic circulation (Cornforth et al. 2017).

In climatological means, the SHL appears as a closed low in surface pressure fields, typically centred around 22° N. Its position and intensity vary intraseasonally, constrained by the surrounding orography of the Atlas and Ahaggar mountains and the Atlantic Ocean to the west. These variations have important meteorological implications: when the SHL strengthens, the resulting enhancement of the meridional pressure gradient between the desert and the Gulf of Guinea intensifies the low-level monsoon inflow toward the Sahel. Variability in the SHL is

therefore closely linked to fluctuations in the monsoon circulation and regional rainfall patterns. Consequently, the SHL acts as a thermal and dynamical anchor for the West African monsoon and serves as a key indicator of its variability (Cornforth et al. 2017).

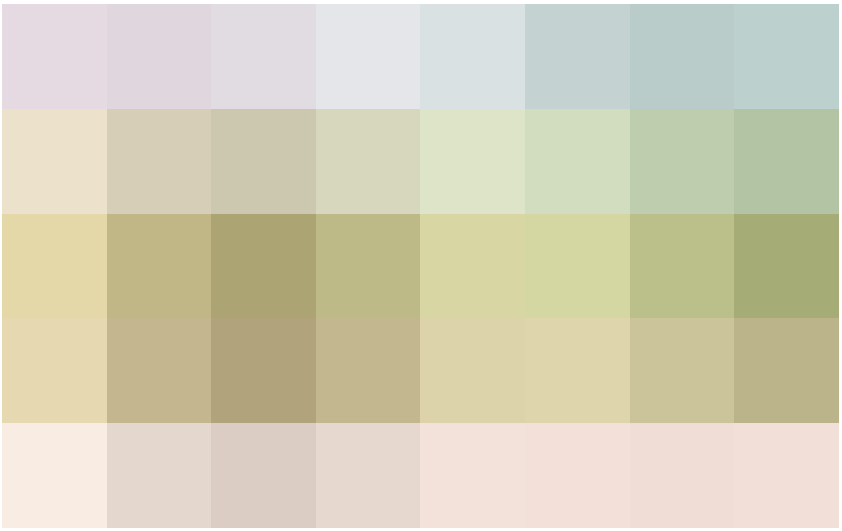


Figure 2.1: Climatology of MSLP (contours) and 10 m wind vectors for September, based on ERA5 data from 1990–2020. The plot domain represents tropical Africa as defined in this study (10° S–25° N, 20° W–50° E).

African Easterly Jet (AEJ)

The African Easterly Jet (AEJ) is a mid-tropospheric easterly wind maximum located near 600–700 hPa and roughly between 10° N and 15° N with a zonal wind peak 15 m s^{-1}. It arises from the horizontal temperature contrast between the hot, dry Saharan air to the north and the cooler, moister monsoon air to the south and is roughly in thermal wind balance. The vertical and horizontal wind shear associated with the AEJ provides the environment for the growth of African Easterly Waves, which are among the most significant synoptic disturbances in the region (Cornforth et al. 2017).

Tropical Easterly Jet (TEJ)

The Tropical Easterly Jet (TEJ) is an upper-tropospheric easterly flow, strongest near 150 hPa and centred between 5° and 15° N. It originates from the large-scale upper-level outflow associated with deep convection over South Asia and the Indian Ocean and extends westward across tropical Africa. Over West Africa, the TEJ promotes upper-level divergence and enhances the vertical development of convective systems (Cornforth et al. 2017).

Intertropical Discontinuity (ITD)

The Intertropical Discontinuity (ITD) marks the boundary between two contrasting air masses over West Africa: the hot, dry Saharan air to the north and the cooler, moister monsoon air

originating from the Gulf of Guinea to the south. The ITD is characterized by a sharp gradient in moisture and a confluence of near surface winds. The ITD shows a pronounced diurnal cycle. At night, as the low-level monsoon flow strengthens toward the SHL, the ITD advances northward. During the day, it retreats southward, leading to a displacement that can exceed a hundred kilometres within 24 hours (Cornforth et al. 2017).

African Easterly Waves (AEWs)

African Easterly Waves (AEWs) are westward-propagating synoptic-scale disturbances that develop along the AEJ during the summer season. They typically have wavelengths of several thousand kilometres and periods of two to six days. AEWs are observed as alternating regions of cyclonic and anticyclonic vorticity embedded within the mean easterly flow. These waves are most clearly identified in wind and pressure fields at mid-tropospheric levels and play an important role in organizing convection across West Africa. The amplitude of AEWs tends to peak near the level of the AEJ, and their evolution strongly influences day-to-day weather variability in the Sahel and coastal regions. As they propagate westward, AEWs can also extend into the tropical Atlantic, where they sometimes act as precursors to tropical cyclones. However, within West Africa, their main significance lies in modulating convective activity and rainfall patterns along the monsoon zone (Cornforth et al. 2017).

West African Monsoon (WAM)

The West African Monsoon (WAM) is the dominant feature of the regional climate, representing a seasonal reversal of surface winds over the greater part of West Africa. During boreal winter, the circulation is controlled by the dry northeasterly Harmattan flow, while in boreal summer southwesterly monsoon winds transport moist maritime air from the Atlantic Ocean deep into the continent. This reversal results from the establishment of the SHL over the Sahara and the associated meridional pressure gradient between the hot continent and the relatively cool ocean.

Following Fink et al. (2017) WAM circulation can be described through four characteristic weather zones, illustrated in Figure 2.2. From north to south, these are:

(A) Located north of the ITD, around 20° N, this zone is dominated by hot and dry northerly surface flow, with high temperatures and low dew point. Precipitation is rare and the SHL is located in this region.

(B) Shallow monsoon layer, intermittent convection can occur.

(C) Characterized by strong monsoon flow and enhanced mid-level easterlies (AEJ), this zone experiences deep convection and the highest rainfall within the WAM system.

(D) The southernmost zone along the Gulf of Guinea coast is influenced by persistent maritime air and low-level cloudiness, resulting in humid conditions and frequent precipitation.

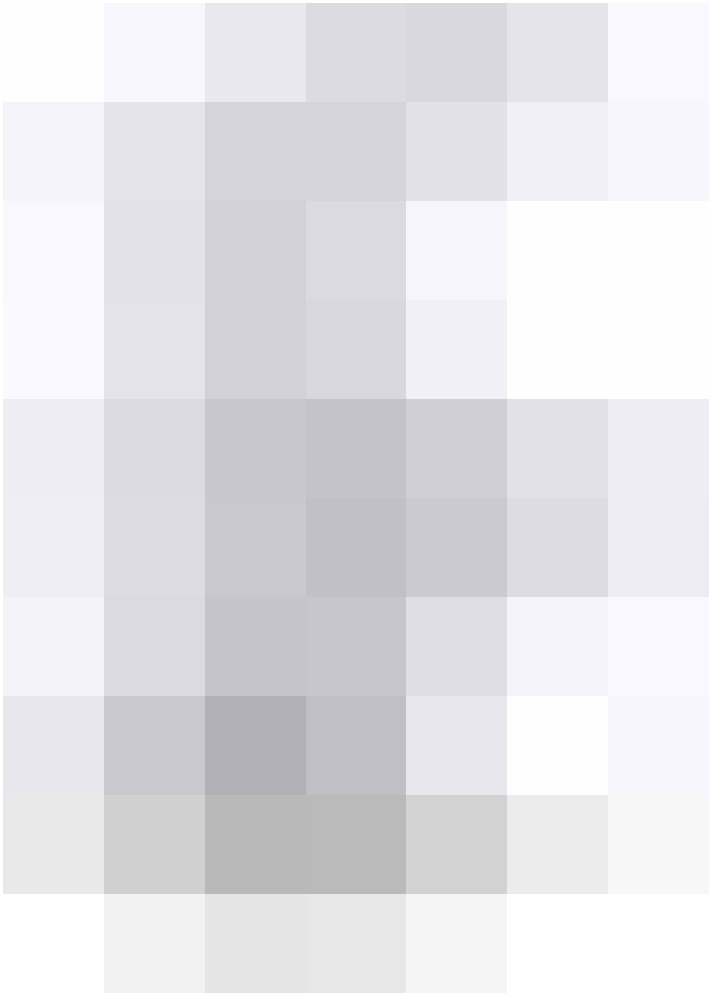


Figure 2.2: Schematic cross-section of the atmosphere over West Africa (10° W–10° E) in July, illustrating the main weather zones (A–D) associated with the West African Monsoon. The diagram shows the positions of the Intertropical Discontinuity (ITD), African Easterly Jet (AEJ), Tropical Easterly Jet (TEJ), and the monsoon layer (ML). Additional features include streamlines, cloud distribution, freezing level, isentropes (Θ), temperature extrema (T_n, T_x), mean temperature (T) and dew-point temperature (T_d), atmospheric pressure (p), and mean monthly rainfall totals (RR). Figure taken from Fink et al. (2017).

Deep Convection

Cumulonimbus clouds are responsible for the majority of seasonal rainfall in West Africa and form through atmospheric convection, a process driven by unstable vertical distributions of thermodynamic energy that naturally tend toward a stable state. Over the tropics, unstable conditions are generated universally due to the excess of moist energy from heated surfaces and energy loss through radiative processes in the free atmosphere. This widespread potential for convection, combined with its sensitivity to large-scale flows and local variations, makes predicting convective events challenging. The type and intensity of convection are governed by the interplay of updraughts, downdraughts, and wind shear, producing phenomena ranging from single-cell storms to highly organized systems, such as squall lines. Squall lines are the predominant form of mesoscale convective systems (MCSs) over West Africa, with lifetimes of several hours to a few days and horizontal extents exceeding 5000 km^2 (Lafore et al. 2017).

2.2 Numerical Weather Prediction (NWP)

The basic idea of NWP is to use physical laws to predict the weather and goes back to Abbe (1901) and Bjerknes (1904). Both recognized that predicting the state of the atmosphere could be framed as an initial value problem, where the future weather is determined by integrating the governing partial differential equations from the current observed state. While conceptually brilliant, at the beginning of the 20th century only a few routine atmospheric observations were available, computers did not yet exist to solve the equations, and the understanding of weather dynamics was limited. As a result, practical forecasting systems were hardly feasible (Bauer et al. 2015).

Today, NWP systems operate on large supercomputers worldwide, and forecast accuracy has significantly increased over the past 50 years (Figure 2.4). These improvements did not arise from a single breakthrough but from incremental advances across NWP components (Bauer et al. 2015). According to Kalnay (2002, p. 2), major factors include:

- Increased computational power, allowing higher numerical resolution and reduced approximations in operational models.
- Improved representation of small-scale physical processes, including clouds, precipitation, and turbulent transfer of heat, moisture, momentum, and radiation.
- Advanced data assimilation techniques yielding more accurate initial conditions.
- Expanded availability of observations, particularly from satellites and aircraft over the Southern Hemisphere and oceans.

As indicated above, a successful NWP system relies on several components, including an observation system, data assimilation, a NWP model, post-processing and verification. A schematic overview of this workflow is provided in Figure 2.3. These components are described in the following.

Since NWP is an initial value problem, it is crucial to obtain the best possible estimate of the current state of the atmosphere, providing the model with an accurate starting point. In addition, surface and boundary conditions, such as sea surface temperatures, sea ice extent, or vegetation coverage, must be specified. Modern systems use observations from both in situ instruments (ground stations, radiosondes, aircraft) and remote sensing (radar, lidar, satellites). Remote sensing often requires retrieval algorithms to convert raw measurements into meteorologically useful quantities (Warner 2010b).

Operational centres combine observations with a short-term model forecast (the background) to produce the best estimate of the initial state (analysis), a process known as data assimilation (Kalnay 2002, p. 136). Many approaches — such as optimal interpolation (OI), three-dimensional variational (3D-Var), or Kalman filtering — are based on the analysis equation

$$\mathrm{x}_a = \mathrm{x}_b + \mathrm{W}\left[\mathrm{y}_o - H(\mathrm{x}_b)\right], \tag{2.1}$$

where x_a is the analysis, x_b the background, y_o the observations, W the weighting matrix and H the observation operator. x_a and x_b are large vectors containing all model grid points and

prognostic variables, while y_o contains all observations. The observation operator H maps the model background into observational space. For example, for a radiosonde temperature measurement, H performs a spatial interpolation from model grid points to the observation location, while for satellite observations, H includes a radiative transfer calculation simulating what a satellite would measure if the model represented reality.

The weighting matrix W balances the contribution of model background and observations according to their uncertainty:

$$\mathrm{W} = \mathrm{BH}^T(\mathrm{R} + \mathrm{HBH}^T)^{-1}, \tag{2.2}$$

where B is the background error covariance and R observational error covariance.

The matrix B represents the expected errors in the model background, including their spatial correlations and relative magnitudes. It is often estimated using ensemble statistics or forecast-difference methods, such as the NMC method, which uses differences between forecasts valid at the same time but initialized at different times. More sophisticated flow-dependent approaches, often based on ensemble data assimilation, allow B to adapt to the current weather regime, representing the so-called "error of the day". The variances in R quantify observational uncertainty and are typically estimated from instrument calibration data, with additional contributions from representativeness errors (Warner 2010b).

In 3D-Var, a cost function J is defined, measuring the squared distance between a state vector x and both the background and the observations:

$$J(x) = \frac{1}{2}(\mathrm{x} - \mathrm{x}_b)^T\mathrm{B}^{-1}(\mathrm{x} - \mathrm{x}_b) + \frac{1}{2}\left[\mathrm{y}_o - H(x)\right]^T\mathrm{R}^{-1}\left[\mathrm{y}_o - H(x)\right] \tag{2.3}$$

The goal is to find the global minimum of J, usually achieved iteratively using gradient-based minimization. The location of this minimum corresponds to the analysis (Warner 2010b).

The variational approach has been extended to four dimensions (four-dimensional variational, 4D-Var) by incorporating observations over a time interval, known as the assimilation window. Unlike 3D-Var, which treats all observations as if they occurred at a single analysis time, 4D-Var uses the forecast model to evolve the atmospheric state, allowing it to make optimal use of non-simultaneous observations such as satellite and aircraft data. As a result, the analysis lies on a model trajectory and is dynamically consistent, reducing unrealistic imbalances in temperature, wind, and pressure fields, and providing a smoother transition from analysis to forecast. At ECMWF, 4D-Var assimilation has been operational since 1997 (ECMWF 2025c).

Many other data assimilation methods exist, including various Kalman filter variants and hybrid approaches, which will not be discussed further here.

Once an analysis is available, the NWP model integrates the governing equations forward in time. The dynamical core includes the momentum equations for a spherical Earth, the continuity equation, the first law of thermodynamics, the ideal gas law, and a water vapour

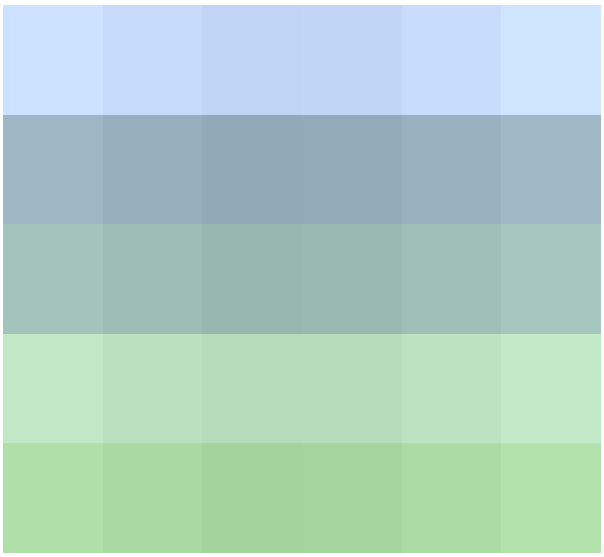


Figure 2.3: Schematic representation of the main components of a numerical weather prediction system. Observations and model background are combined through data assimilation to produce the analysis, which initializes the model. The dynamical core and physical parametrizations generate the forecast, followed by postprocessing and verification.

budget equation (Warner 2010e). Because these equations cannot be solved analytically, they are integrated numerically on a discrete grid (Warner 2010c).

Due to the limited spatial resolution of NWP models, not all atmospheric processes, such as cloud microphysics, radiation, or turbulence, can be resolved explicitly and must be parametrized. Parametrization schemes relate the effects of unresolved physical processes to the variables represented in the model (Warner 2010d).

The atmosphere is a highly non-linear and complex system, so forecasts are inherently limited in lead time and carry uncertainty. Even a simple two-dimensional flow model, much less complex than the real atmosphere, as shown by Lorenz (1969), is highly sensitive to small differences in initial conditions, producing diverging results after a certain time. Lorenz demonstrated that an intrinsic limit of predictability exists. Moreover, no NWP model is perfect, and additional uncertainties arise from numerical approximations in the dynamical core, as well as from imperfect parametrizations of physical processes (Bauer et al. 2015).

To quantify forecast uncertainty, ensemble prediction has become increasingly important, facilitated by growing computational power. Unlike deterministic forecasts, which produce a single prediction, ensemble forecasts consist of multiple runs under slightly varied conditions. These variations are introduced through perturbations in initial conditions, using methods such as singular vectors or breeding, or through stochastic parametrizations. The goal is to sample the uncertainty in the modeling process and assess its impact on the forecasts (Warner 2010a). At ECMWF, global forecasts are currently performed with 50 ensemble members (ECMWF 2025d).

Often, the direct model output from NWP systems is not immediately useful for end users and may contain systematic errors. To enhance the practical value of forecasts, the raw model output is typically refined through post-processing. Common post-processing techniques include bias correction, downscaling, and the derivation of additional quantities, such as wind chill.

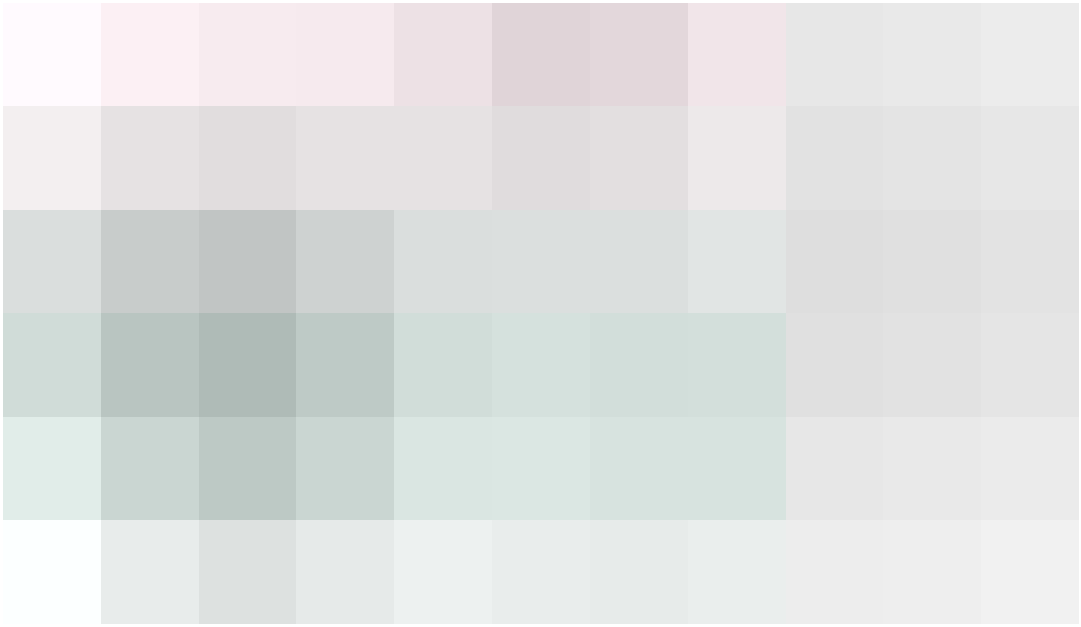


Figure 2.4: Anomaly correlation coefficients of 3-, 5-, 7-, and 10-day ECMWF 500 hPa geopotential height forecasts for the extratropical Northern and Southern Hemispheres, plotted as annual running means of archived monthly-mean scores from January 1, 1981, to January 31, 2025. The values for each month are averaged over that month and the 11 preceding months. Shading indicates differences in scores between the two hemispheres at the specified forecast ranges. Figure taken from (ECMWF 2025d).

These methods can be applied to both deterministic and ensemble forecasts to improve their accuracy and relevance for specific applications.

Forecast verification evaluates the quality and performance of forecasts. It involves comparing forecasts with corresponding observations to quantify attributes such as bias, reliability, resolution, discrimination, and skill. The choice of verification metric depends strongly on the forecast type and user requirements, ranging from basic measures such as the root mean square error (RMSE) or simple categorical scores to more advanced probabilistic metrics used for ensemble forecasts, such as the Brier score (BS) or continuous ranked probability score (CRPS). Verification provides valuable insights into the strengths and limitations of forecasts, supporting both model development and informed use by end users (Wilks 2011).

3 Current State of Research

Tropical Africa poses unique challenges for weather forecasting due to a combination of intense convective systems, limited observational networks, and complex interactions between the monsoon and larger-scale atmospheric waves (Lamptey et al. 2024). This chapter reviews the current understanding of atmospheric predictability in the tropics and examines factors limiting forecast skill in the region.

The fundamental limits of weather prediction were first articulated by Lorenz (1969), who proposed that deterministic fluid systems containing many interacting scales of motion may be observationally indistinguishable from indeterministic systems. Even if the governing equations are perfectly known, small uncertainties in the initial state inevitably amplify with time until the predicted and actual states become effectively uncorrelated. Importantly, Lorenz showed that reducing the amplitude of the initial error cannot indefinitely extend the forecast horizon — the atmosphere possesses an intrinsic finite range of predictability.

Using a simplified spectral model derived from the two-dimensional vorticity equation, Lorenz demonstrated that each scale of motion has a characteristic predictability time that depends on the energy distribution across scales. When the kinetic energy spectrum does not decrease too rapidly with decreasing wave length, smaller scales evolve more quickly and lose predictability sooner than larger ones. With parameters comparable to the Earth's atmosphere, Lorenz estimated approximate limits of about one hour for cumulus-scale motions, a few days for synoptic-scale systems, and a few weeks for the largest planetary waves.

Building on Lorenz's intrinsic perspective, Judt (2020) investigated the intrinsic and practical predictability of the atmosphere across different climate regions using outputs from a global storm-resolving model. They find that error growth is strongly latitude- and scale-dependent. At smaller scales (below $\sim$300 km wavelength), the tropics exhibit lower predictability than the mid-latitudes, while at planetary scales, tropical predictability extends up to 20 days. Mid-latitudes and polar regions generally reach saturation after slightly more than two weeks. This variation reflects the underlying dynamics: equatorial Kelvin, Rossby, and mixed Rossby–gravity waves in the tropics are more resistant to error growth than baroclinic disturbances in the mid-latitudes, explaining the extended tropical predictability. Judt's results indicate that current NWP systems have not yet reached the theoretical limits of predictability, particularly in the tropics, and that higher-resolution models with improved dynamical representation — such as global storm-resolving models — can help exploit these extended horizons. The scale- and region-dependent predictability, based on differences in kinetic energy (DKE), is illustrated in Figure 3.1.

Consistent patterns were also found by Keane et al. (2025), who showed that, after approximately 5 to 7 days, predictability in the tropics tends to remain higher than in the mid-latitudes. In addition to DKE, Keane evaluated precipitation using the fractions skill score (FSS). Their results indicate that mid-latitudes are more predictable at small spatial scales and short lead times, whereas tropical predictability is greater at larger scales and longer lead times. This

behaviour is illustrated in Figure 3.2, which shows FSS maps for different lead times and spatial scales. These patterns are robust across multiple systems, including both traditional NWP models and machine-learning-based approaches, suggesting that the observed latitude- and scale-dependent predictability reflects fundamental atmospheric dynamics rather than model-specific characteristics.

Figure 3.1: Red dots indicate scale-dependent limits of atmospheric predictability (90 % error saturation), while orange dots show the limits of practically useful forecast skill (60 % saturation), based on DKE. Star symbols at 20 days mark cases where the respective saturation level was not reached within that period. Data for scales smaller than the model's effective resolution of 24 km (vertical gray lines) are excluded. Figure taken from Judt (2020).

Poor short-range precipitation forecast performance over tropical Africa has also been documented by Vogel et al. (2020). While other tropical regions show some forecast skill, tropical Africa stands out with even negative skill for 1-day accumulated precipitation forecasts. Vogel et al. (2020) attribute this to a fundamental mismatch between modelled and observed rainfall characteristics: convection parametrizations in current NWP systems tend to produce too frequent and too weak precipitation events, whereas rainfall in reality is dominated by intense, long-lived MCSs. Postprocessing — through statistical correction of ensemble forecasts — can substantially increase skill and largely remove negative biases, but even after such correction, forecast skill remains only neutral over tropical Africa. This suggests that the structural discrepancies between model convection and observed organized convective systems are too large to be fully corrected statistically. Figure 3.3 shows the continuos ranked probability skill core (CRPSS) for the tropical belt before and after postprocessing.

In response to the persistently poor precipitation forecast performance in the tropics, Walz et al. (2024b) developed a data-driven approach based on a U-Net convolutional neural network (CNN). The model is trained on Integrated Multi-satellitE Retrievals for GPM (IMERG) data to produce deterministic forecasts, which are subsequently transformed into probabilistic forecasts using the recently introduced non-parametric Easy Uncertainty Quantification (EasyUQ) method (Walz et al. 2024a). In general, purely statistical approaches perform comparably to postprocessed ECMWF ensemble forecasts. However, the CNN–EasyUQ approach clearly outperforms all competitors in terms of both precipitation occurrence and amount (Figure 3.4). A hybrid configuration combining CNN–EasyUQ with physics-based ensemble forecasts yields slight additional improvements. These results indicate that the CNN–EasyUQ

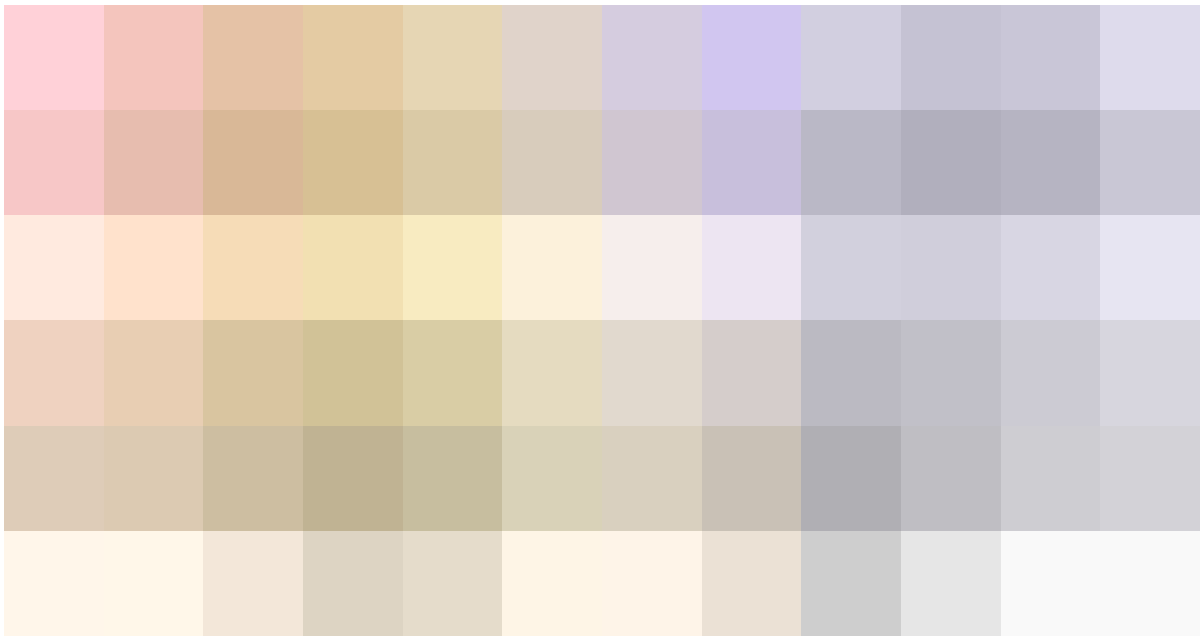


Figure 3.2: Spatial maps of the FSS for the year 2022, comparing the ECMWF operational forecasting system (top panels) with the GraphCast machine-learning model (bottom panels). The maps display forecast skill at varying lead times and spatial averaging scales, as noted in the panel titles (2 days / 2 gridpoints on the left; 10 days / 12 gridpoints on the right). The FSS is calculated using a 90th percentile threshold, determined separately for forecasts and observations at each grid point. Areas shaded in brown represent higher skill while purple shading indicates lower skill. Figure taken from Keane et al. (2025).

framework can enhance operational probabilistic rainfall forecasting in tropical Africa and may offer benefits beyond this region.

In a complementary study, Rasheeda Satheesh et al. (2025) investigate machine-learning models for daily rainfall forecasting in northern tropical Africa using predictors derived from tropical-wave activity. Their CNN and gamma-regression models outperform benchmark ensemble and climatology-based forecasts, demonstrating that tropical-wave information contains substantial predictive potential for rainfall in this region. By relying on physically interpretable large-scale tropical-wave predictors, their approach offers a cost-effective alternative to numerical weather prediction, particularly in areas where conventional forecasting systems show limited skill. These findings reinforce the growing evidence that data-driven methods can enhance tropical rainfall prediction.

Pante and Knippertz (2019) demonstrate that explicitly resolving Sahelian MCSs in the ICON model using a two-way nesting approach significantly improves weather forecasts not only over West Africa but also in remote tropical and extratropical regions. Their simulations show that the convection-parametrized run (PARAM) produces widespread but weak and slowly moving rainfall across the Sahel, while explicitly resolving convection (EXPLC) captures distinct, intense MCSs propagating westward with lifetimes of 2–4 days (Figure 3.5). The explicitly resolved MCSs in EXPLC closely replicate the timing, intensity, and movement of those observed by the Tropical Rainfall Measuring Mission (TRMM), while substantially reducing the wet bias present in PARAM. Improved representation of precipitation and its diurnal cycle over the Sahel leads to better moisture balance and enhances forecast accuracy across large-

Figure 3.3: CRPSS for 1-day accumulated precipitation forecasts from the ECMWF ensemble during 2009–2017, relative to extended probabilistic climatology (EPC). Panel (a) shows forecast skill before postprocessing and panel (b) after statistical correction was applied. Figure taken from Vogel et al. (2020).

scale atmospheric patterns such as Rossby wave trains. This work highlights the potential of convective-permitting models to enhance both regional and global weather prediction, emphasizing the importance of resolving tropical convection for more accurate rainfall forecasts.

As mentioned in the Introduction (Chapter 1), tropical Africa remains one of the most data-sparse regions in the world (see Figure 1.1). During the boreal summer months, AMMA (Redelsperger et al. 2006) and DACCIWA (Knippertz et al. 2017) field campaigns deployed a dense network of radiosonde observations across West Africa. While improving forecast skill was not the primary aim of these campaigns, the collected observations were subsequently used in NWP experiments to evaluate their influence on analyses and forecasts through data denial studies comparing simulations with and without these additional soundings.

The impact of the AMMA radiosonde data on the analyses is substantial. Faccani et al. (2009) and Agustí-Panareda et al. (2010) find that assimilating the extra soundings significantly improved low-level temperature, moisture, and wind fields over the Sahel. In particular, the AEJ is better represented, with enhancements along its south-easterly flank (Faccani et al. 2009; Agustí-Panareda et al. 2010). Humidity bias correction further improved surface relative humidity, TCWV, and precipitation fields.

Improvements in forecasts are more limited. Agustí-Panareda et al. (2010) report that the positive impacts are largely confined to the first 24 hours, due to persistent model biases in the boundary layer, convection, and cloud processes. Both studies indicate that, while dense radiosonde networks can substantially improve NWP analyses, longer-term forecast benefits remain constrained by model errors.

Building on the insights from AMMA, van der Linden et al. (2020) evaluate the impact of the DACCIWA campaign, which deployed roughly 900 soundings from 12 stations across southern West Africa during June and July 2016. Similar to AMMA, data-denial experiments

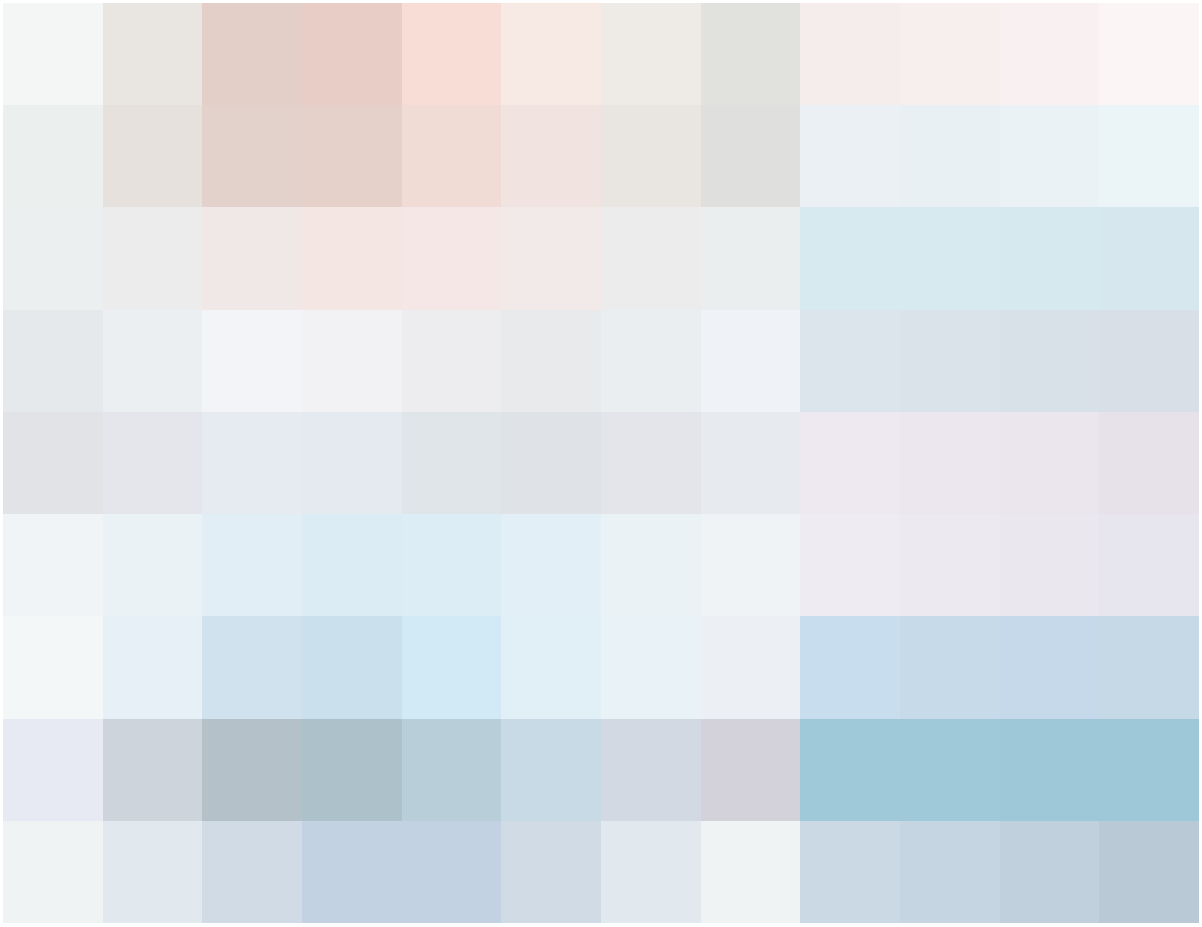


Figure 3.4: CRPSS for 1-day accumulated precipitation forecasts relative to the monthly probabilistic climatology (MPC) baseline. Panels show results for (a) the full ECMWF ensemble prediction system (EPS), (b) EPS with ensemble model output statistics (EMOS) postprocessing, (c) the high-resolution (HRES) run combined with EasyUQ, (d,e) two logistic regression-based distributional index models (DIM), (f) CNN with EasyUQ and (g) the hybrid forecast combining CNN–EasyUQ with physics-based forecasts. Results are averaged over July, August and September and evaluation folds from 2011–2019. Figure taken from Walz et al. (2024b).

are conducted using ECMWF's IFS to assess how these additional observations influenced both analyses and forecasts.

The DACCIWA radiosondes have a noticeable impact on the analyses, particularly on wind fields throughout the troposphere and lower stratosphere. Low-level temperature and humidity biases remained largely unchanged. The largest observation impact occurs at night and in the early morning, while daytime boundary-layer turbulence limits their effect. High-level cloud cover and the tropical easterly jet also show modest improvements, indicating a better representation of larger-scale circulation features.

The effect on forecasts is, as with AMMA, limited in duration. Positive impacts on low-level wind, temperature, and rainfall generally last no more than 12–24 hours, with only weak but consistent improvements in precipitation downstream of the radiosonde stations. Biases in outgoing long-wave radiation and rainfall persist, highlighting that model errors and data assimilation limitations remain the primary constraints on forecast quality.

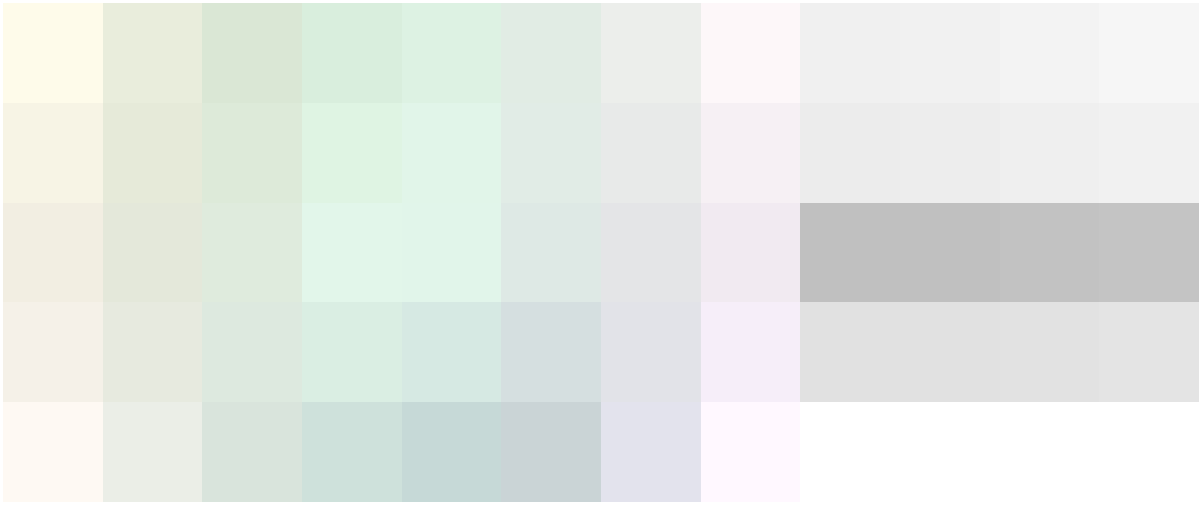


Figure 3.5: Precipitation simulation over West Africa. Panels (a–c) show Hovmöller diagrams of 3-hourly precipitation averaged from 8° N–18° N between 12 UTC 5 August and 12 UTC 15 August 2017 for simulations with the convection parametrization (PARAM, a), explicitly resolved convection (EXPLC, b) and TRMM observations (c). Panel (d) shows the diurnal cycle averaged from 8° N–18° N and 10° W–10° E (region indicated by dashed lines in a–c. Figure taken from Pante and Knippertz (2019).

Following the regional field campaigns AMMA and DACCIWA, more recent studies assess the impact of satellite-based wind profiling on weather prediction over Africa and beyond. The Aeolus mission, launched by the European Space Agency in 2018, is the first satellite to provide global vertical profiles of horizontal winds, thereby filling a crucial gap in the tropical observing system. Using Aeolus data, Borne et al. (2023) investigated its influence on the WAM circulation during the summers of 2019 and 2020 within the ECMWF and DWD forecasting systems. Assimilating Aeolus winds improves the representation of key circulation features such as the AEJ and TEJ, with the strongest benefits in the upper troposphere.

Building on these regional findings, Borne et al. (2025) conduct a three-year global observing system experiment using the improved reprocessed Aeolus dataset (Baseline 16). They report significant reductions in zonal wind forecast errors — up to 1–1.5 % in the tropical upper troposphere — and downstream improvements in precipitation forecasts, particularly for longer lead times of 5–10 days. The most pronounced rainfall benefits are found in the Southern Hemisphere winter, while in the Tropics, improvements are generally weaker and less dependent on the season. Together, these studies demonstrate that assimilating Aeolus wind profiles enhances the representation of large-scale circulation features such as jet streams and Rossby waves, leading to more accurate medium-range forecasts of wind and heavy rainfall.

To complement insights from regional campaigns and satellite wind profiling, recent global experiments explore the potential impact of additional observations in data-sparse regions. The ECMWF SOFF impact experiments aim to quantify the value of hypothetical observing stations, with a particular focus on Africa and the Pacific. These experiments do not use real-world observations; instead, simulated datasets are generated to represent the effect of adding new stations. Using the Ensemble of Data Assimilations (EDA), the experiments assess reductions in analysis and short-range forecast uncertainty, limited to 12-hour forecasts. The results show a very large impact in Africa, where forecast uncertainty is reduced by more than 30 %, demonstrating the critical importance of filling observational gaps in under-sampled

regions (see Figure 3.6). While these experiments confirm the potential benefits of targeted observational investments, longer lead times beyond 12 hours are not considered, leaving the medium-range impacts of additional observations as an important area for future research.

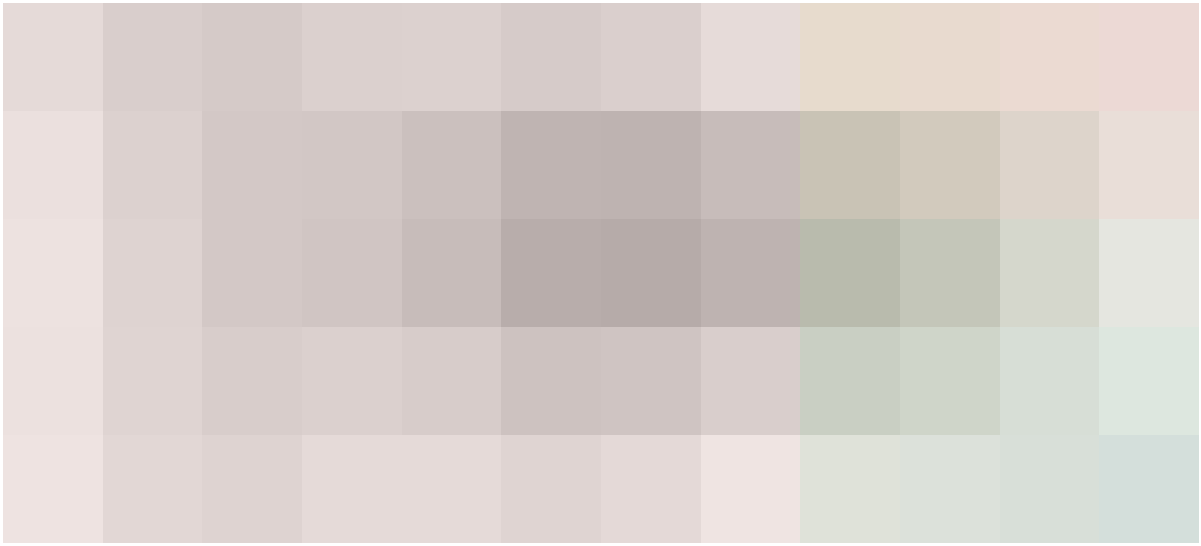


Figure 3.6: Percentage reduction in surface pressure analysis uncertainty for Scenario 1, which adds additional surface and upper-air observing stations in less developed countries and small island developing states, compared to the control experiment with the baseline observing network, for 1–30 June 2023. Blue shading indicates regions where Scenario 1 reduces analysis uncertainty relative to the control, and diagonal hatching denotes areas where the improvement is statistically significant at the 95 % confidence level. Figure taken from ECMWF (2025b).

In summary, weather prediction in tropical Africa is limited by intrinsic predictability, model biases, model resolution, and observational gaps. Both regional campaigns and satellite missions demonstrate measurable improvements, while simulated expansions of observing networks suggest large untapped potential for further reducing forecast uncertainty.

4 Data and Methodology

4.1 Datasets and Models

This chapter provides an overview of the datasets and models employed in the study, followed by a description of the TEEMLEAP testbed. Finally, the evaluation metrics used to assess the performance of the forecasts are outlined.

4.1.1 ERA5 Reanalysis

The ERA5 dataset is a state-of-the-art global atmospheric reanalysis produced by ECMWF as part of the Copernicus Climate Change Service (C3S) (Hersbach et al. 2020). ERA5 provides hourly estimates of a wide range of atmospheric, land, and oceanic climate variables on a native model grid with a horizontal resolution of 31 km (TL639). The dataset covers the period from 1940 onwards (Soci et al. 2024). ERA5 is produced using the ECMWF IFS, specifically cycle 41r2, which was operational at ECMWF in 2016. The reanalysis employs a 4D-Var data assimilation system to optimally combine model forecasts with a large volume of diverse observational data, including satellite and in situ measurements. The vertical discretization consists of 137 hybrid sigma-pressure levels, extending from the Earth's surface up to 0.01 hPa, corresponding roughly to an altitude of 80–100 km.

4.1.2 Integrated Multi-satellite Retrievals for GPM (IMERG)

The Integrated Multi-satellitE Retrievals (IMERG) product of the Global Precipitation Measurement (GPM) mission (Huffman et al. 2020) provides high-resolution precipitation estimates derived from a constellation of multiple satellites. IMERG primarily utilizes passive microwave (PMW) sensors on low Earth orbit (LEO) platforms, which are highly sensitive to rainfall and thus provide the most accurate satellite-based precipitation estimates. PMW sensors measure the natural microwave radiation emitted by the Earth's surface, which is strongly influenced by the presence of liquid water in the atmosphere. As raindrops absorb and emit microwave radiation, the intensity of the observed signal increases with rainfall, allowing for direct estimation of precipitation rates, particularly in deep convective systems.

In addition, the product incorporates infrared (IR) sensor data from geosynchronous Earth orbit (GEO) satellites. IR sensors detect thermal radiation emitted by cloud tops, with colder cloud tops generally indicating higher altitudes associated with more intense rainfall. Although IR-based estimates provide global coverage, their accuracy tends to be lower compared to PMW estimates, as they rely on indirect relationships between cloud top temperature and precipitation intensity.

IMERG integrates these diverse data sources to produce half-hourly precipitation estimates with a spatial resolution of $0.1° \times 0.1°$ over a latitude range of 60° S–60° N, with partial coverage extending beyond this latitude band. To enhance the accuracy and reliability of the satellite-derived estimates, the product is calibrated against surface-based precipitation gauge observations. This calibration process helps correct for systematic biases and improves the quality of the data.

IMERG is delivered in three processing stages: Early (approximately 4 hours after observation), Late (approximately 14 hours after observation), and Final (3.5 months after observation). Each stage reflects different levels of data refinement and incorporation of additional sources of information. For this study, the final version of IMERG V07A is used (Huffman 2023).

4.1.3 Basic Cycling Environment (BACY)

The Basic Cycling Environment (BACY) developed at DWD is a data assimilation and forecast framework designed for rapid experimentation (Schraff et al. 2016). Compared to DWD's standard operational system, BACY offers significantly faster cycling by using hard drive storage instead of tape-based archives. It is fully portable to standard Linux environments, facilitating collaboration. BACY is flexible and easy to modify, allowing for rapid implementation of changes and extensions. In recent years, BACY has been employed in several assimilation-related studies (Bick et al. 2016; Ruckstuhl and Janjić 2020; Zeng et al. 2021).

The BACY framework operates through three key cycles: the assimilation cycle (ASM), the forecast cycle (MAIN) and the verification cycle (VERI). The ASM cycle performs data assimilation using DACE and ICON, providing short-range forecasts and surface analyses for variables like snow, soil moisture, and sea surface temperature. The MAIN cycle generates medium-range forecasts based on the conditions from ASM. The VERI cycle assesses forecast accuracy by comparing model outputs with observational data, including radiosoundings and remote sensing observations.

4.1.4 Data Assimilation Coding Environment (DACE)

The Data Assimilation Coding Environment (DACE) is a modular framework developed by DWD to facilitate data assimilation in numerical weather prediction models. It supports multiple assimilation methods, including 3d-Var, the Local Ensemble Transform Kalman Filter (LETKF), and a hybrid variational ensemble Kalman Filter (VarEnKF) (Potthast 2019).

4.1.5 ICOsahedral Nonhydrostatic (ICON) Model

The ICOsahedral Nonhydrostatic (ICON) model is a state-of-the-art numerical weather prediction and climate simulation system jointly developed by DWD and the Max Planck Institute for Meteorology (MPI-M) (Zängl et al. 2015). It is designed as a flexible, unified modelling framework capable of simulating the atmosphere across a wide range of spatial and temporal scales.

ICON employs an icosahedral–triangular Arakawa C grid, which subdivides the sphere into triangular cells of equal size, providing uniform horizontal resolution while avoiding the singularities and grid convergence issues common to traditional latitude–longitude grids. The dynamical core solves the non-hydrostatic equations using a finite-volume discretization method. Time integration is performed with a two-time-level predictor–corrector scheme that is explicit in the horizontal directions and implicit for terms that describe vertical propagation of soundwaves.

The model uses a terrain-following height coordinate to accurately represent orographic effects and vertical motions. Furthermore, ICON supports both one-way and two-way nesting, allowing high-resolution limited-area domains to be embedded seamlessly within the global simulation domain without requiring external boundary conditions.

For this study, ICON release 2.6.6 is used.

4.2 TEEMLEAP Testbed

4.2.1 Overview

In response to the rapid developments in the last few years in NWP, increasingly incorporating machine learning methods, the Karlsruhe Institute of Technology (KIT) and DWD developed the TEEMLEAP testbed to systematically investigate potential improvements in weather forecasting. The testbed simulates the entire operational weather forecasting chain and is implemented on the high performance computing system HoreKa at KIT. It is built on DWD's BACY (subsection 4.1.3), which integrates data assimilation workflows with DACE (subsection 4.1.4) and the ICON model (subsection 4.1.5) for NWP. Alternatively, the forecast can be generated using the Fourier Forecasting Neural Network (FourCastNet) (Pathak et al. 2022), which is already integrated into the testbed, although it is not utilized in this study. A verification cycle can be performed to assess the accuracy of the predictions against ERA5 data. However, for this work, partly more specific evaluation metrics are instead employed, outside the scope of the testbed. The structure of the testbed is summarized in Figure 4.1.

Unlike NWP systems used by weather services, the TEEMLEAP testbed does not rely on real observational data from in situ or remote sensing instruments. Instead, it utilizes reanalysis data from ECMWF to generate PSOs (subsection 4.2.2). This approach is inspired by Observing System Simulation Experiments (OSSEs), which assess potential improvements in observing systems (e.g., Andersson and Masutani 2010; Masutani et al. 2010; Privé et al. 2013; Privé et al. 2023). The use of PSOs in the testbed allows for efficient data handling and flexible experimentation. However, it introduces the drawback of indirectly incorporating observations, which may lead to model errors and biases. Moreover, relying on a single observation type limits the transferability of results to operational systems, where diverse observation types and more complex assimilation techniques are used (Wilhelm et al. 2025).

Nonetheless, the testbed enables the systematic exploration of various weather forecasting challenges, including improvements to observation systems, understanding forecast uncertainties, and developing hybrid systems that combine machine learning with traditional models.

The surface analysis procedures for snow, soil moisture, and sea surface temperature are excluded from the current version of the testbed. Despite the absence of these surface analyses, reasonable results have been obtained in global simulations during periods with minimal variation in snow cover and sea surface temperatures, such as in September (Wilhelm et al. 2025).

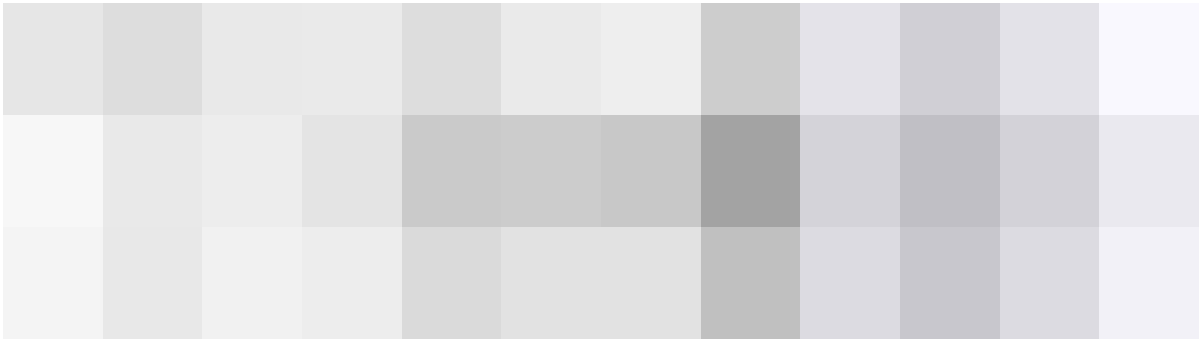


Figure 4.1: Overview of the TEEMLEAP testbed structure. PSOs are generated by simulating radiosonde launches from ERA5 data. Within BACY, these vertical profiles are assimilated with DACE to create the atmospheric analysis. The ICON model is used to calculate both the assimilation background (first guess) in the assimilation cycle and the forward integration in the MAIN cycle. The verification cycle evaluates the accuracy of the forecasts using ERA5 data. Figure taken from Wilhelm et al. (2025).

4.2.2 Pseudo-Observation Profiles

To emulate real-world measurements from radiosondes as closely as possible, vertical profiles of temperature T, relative humidity r, horizontal wind components u, v, and geopotential height at the surface (Z_{sfc}, at the lowermost model level) are used — the same variables as in the operational assimilation case. The vertical profiles are created from ERA5 data ($0.25° \times 0.25°$) on hybrid sigma-pressure model levels, from near the Earth's surface (level 137) up to level 40, corresponding to roughly 24 km altitude. In contrast to real radiosoundings, which provide data at high vertical resolution (typically one measurement per second), the PSO profiles have a coarser vertical spacing (Wilhelm et al. 2025).

The ERA5 values are horizontally interpolated to predefined latitude–longitude positions at which the PSO profiles are generated using a distance-weighted method. The spatial distribution of PSOs is easily configurable. In the default setup for global sensitivity experiments, PSOs are placed on a Fibonacci lattice (González 2010), ensuring an almost uniform and isotropic distribution across the globe.

4.2.3 Pseudo-Observation Error Profiles

Before the assimilation, each vertical PSO profile requires an associated description of observational uncertainty. These error profiles are crucial for the optimization procedure as they

determine how much weight the observations receive relative to the model background. Without them, ERA5-based PSOs would be treated as perfect and the model information would be largely disregarded during the analysis at locations with PSOs.

Within the TEEMLEAP testbed, the necessary error characteristics are derived from ERA5 itself. Following the approach of Desroziers et al. (2005) and adapted by Wilhelm et al. (2025), typical ERA5-intrinsic standard error profiles are diagnosed from observation-minus-background, observation-minus-analysis, and background-minus-analysis differences over several cycling periods. These profiles represent the minimal intrinsic uncertainty expected for the PSOs and are denoted by the pressure-dependent standard deviation $\sigma_{ERA5}(p)$. For experiments assuming lower observational quality, these intrinsic errors can be scaled using a pressure-dependent factor $f(p)$, yielding a total standard deviation

$$\sigma(p) = f(p) \cdot \sigma_{ERA5}(p), \tag{4.1}$$

which reflects the desired observation error level.

To incorporate this total assumed observation error into the assimilation, the ERA5-based PSO profiles are perturbed such that the combined uncertainty from intrinsic ERA5 errors and the added perturbations matches $\sigma(p)$. The perturbations are generated for each PSO profile with zero mean and standard deviations $\sigma_{pert}(p)$. Under the assumption of Gaussian-distributed errors, which allows variances to be added linearly, these satisfy:

$$\sigma(p)^2 = \sigma_{ERA5}(p)^2 + \sigma_{pert}(p)^2 \tag{4.2}$$

To ensure physically plausible and vertically smooth perturbation profiles, the perturbations are constructed using a vertical covariance matrix approach following Houtekamer (1993) and Houtekamer et al. (1996). A key parameter in this process is the vertical decorrelation length L_v, which controls the vertical scale of smoothing: longer values produce smoother profiles, while shorter values allow finer-scale variability. Following Errico et al. (2013), $L_v = 500$ m is used for temperature and horizontal wind components, and $L_v = 180$ m for relative humidity, reflecting the smaller-scale variability typically observed in moisture fields due to clouds, precipitation, and turbulent mixing (Wilhelm et al. 2025). Figure 4.2 shows illustrative examples of perturbation profiles.

The PSO error profiles are constructed under the assumption of Gaussian-distributed errors, which is reasonable for temperature and wind components but only an approximation for relative humidity, which is left- and right-bounded and often non-Gaussian. Despite this limitation, Gaussian perturbations are applied for all variables to generate the total assumed observational uncertainty, combining intrinsic ERA5 errors and additional perturbations. For relative humidity, unphysical perturbed values below 0 % or above 105 % are removed, balancing the need to allow for possible supersaturation while keeping a realistic number of valid samples. Overall, this procedure leads to a marginally positive mean in humidity perturbations (1–4 %), but the assimilation itself prevents the development of a persistent bias in the troposphere (Wilhelm et al. 2025).

Figure 4.2: Illustration of exemplary ERA5-based PSO perturbation profiles for (a) temperature T, (b) relative humidity r and (c) zonal wind u. Light grey lines depict $N = 100$ individual perturbations, while the dark grey line highlights a single profile. The purple line shows the sample mean and the ochre line indicates the sample standard deviation of the perturbations. The dashed ochre line represents the theoretical perturbation standard deviation $\sigma_{\mathrm{pert}}(p)$, and the turquoise line shows the total observation error profile $\sigma(p)$, combining intrinsic ERA5 error and perturbations. Vertical decorrelation lengths L_v used in the generation of perturbations are indicated in each subplot. Figure taken from Wilhelm et al. (2025).

4.3 Verification Metrics

In this chapter, several verification metrics used to evaluate forecast performance are presented. These metrics provide quantitative measures of forecast accuracy and skill, addressing various aspects, including error magnitude, bias, categorical agreement, and spatial verification.

4.3.1 Root Mean Square Error (RMSE)

The Root Mean Square Error (RMSE) is a widely used verification metric that quantifies the average magnitude of the forecast error. It is defined as the square root of the mean squared differences between the forecast values f_i and the observed values o_i:

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(f_i - o_i)^2} \tag{4.3}$$

Due to the squaring of errors, RMSE gives greater weight to larger deviations, making it particularly sensitive to outliers.

4.3.2 Mean Error (ME)

The Mean Error (ME), also known as the bias, measures the average signed difference between forecast and observation:

$$\mathrm{ME} = \frac{1}{n}\sum_{i=1}^{n}(f_i - o_i) \tag{4.4}$$

It indicates whether the forecast systematically overestimates or underestimates the observed values.

4.3.3 Stable Equitable Error in Probability Space (SEEPS)

The Stable Equitable Error in Probability Space (SEEPS) is a categorical verification score designed to evaluate precipitation forecasts in a way that is sensitive to the regional climate, equitably weighted, and robust to sampling uncertainty (Rodwell et al. 2010).

SEEPS addresses the challenges of verifying precipitation, which often follows a skewed and intermittent distribution. Rather than comparing raw precipitation amounts, SEEPS operates in probability space by mapping observations and forecasts through the climatological cumulative distribution function (CDF). This ensures the score accounts for the local climatology and allows consistent interpretation across different regions.

Forecasts are verified by categorizing daily precipitation into three classes:

- p_1, dry $\leq 0.2\,\mathrm{mm}$
- p_2, light precipitation
- p_3, heavy precipitation

Based on the local climatology, the threshold separating light and heavy precipitation is defined such that light precipitation occurs twice as often as heavy precipitation at each individual grid point ($p_1/p_2 = 2$). The score is negatively oriented, with zero representing the best possible score — this occurs when the forecast and observation fall into the same category. If the forecast and observation belong to different categories, the corresponding error term is greater than zero. The error matrix is defined as

$$S_{f,o} = \frac{1}{2}\begin{pmatrix} 0 & \dfrac{1}{1-p_1} & \dfrac{1}{p_3}+\dfrac{1}{1-p_1} \\ \dfrac{1}{p_1} & 0 & \dfrac{1}{p_3} \\ \dfrac{1}{p_1}+\dfrac{1}{1-p_3} & \dfrac{1}{1-p_3} & 0 \end{pmatrix}, \tag{4.5}$$

where rows f and columns o correspond to the forecast and observed categories, respectively: dry, light precipitation, and heavy precipitation. The matrices are inherently asymmetric,

meaning that incorrect forecasts in climatologically frequent categories receive greater penalties than those in less frequent categories. This asymmetry is desirable because it encourages the model to improve its accuracy for the most common weather conditions, thereby enhancing its ability to distinguish between different categories. As the probability of "dry" weather p_1 (or "wet" weather $1 - p_1$) approaches zero, elements of the SEEPS error matrix (Equation (4.5)) become extreme due to the presence of reciprocals of p_1 and $1 - p_1$. Therefore, a limiting range of $p_1 \in [0.10, 0.85]$ is applied.

4.3.4 Fractions Skill Score (FSS)

Highly varying spatio-temporal fields, such as precipitation, are often difficult to verify using classical pointwise verification metrics, especially as model resolution increases. Small spatial errors result in both missed events and false alarms, a problem known as the "double penalty". This occurs when the forecast system predicts the correct feature but slightly displaced in space. Traditional metrics like the RMSE may paradoxically yield a smaller error if the feature is not forecasted at all, rather than being forecasted inaccurately. To overcome this limitation, Roberts and Lean (2008) developed the Fractions Skill Score (FSS), a spatial verification metric that compares the fractions of grid points exceeding a threshold (e.g., daily rainfall > 1 mm) within defined neighbourhoods.

First, the observed rainfall field O_r and the forecast rainfall field M_r both defined on the same grid, are converted into binary fields I_o and I_M, respectively. Grid points where the rainfall exceeds a specified threshold q are assigned a value of 1, while all other points are set to 0:

$$I_o = \begin{cases} 1 & O_r \geq q \\ 0 & O_r < q \end{cases} \qquad and \qquad I_M = \begin{cases} 1 & M_r \geq q \\ 0 & M_r < q \end{cases} \tag{4.6}$$

To generate the fractions at each grid point in the binary fields, the proportion of surrounding points within a square neighbourhood of length n that have a value of 1 is calculated. The resulting fraction fields are denoted as $O(n)(i,j)$ and $M(n)(i,j)$ for the observation and forecast, respectively:

$$O(n)(i,j) = \frac{1}{n^2} \sum_{k=1}^{n} \sum_{l=1}^{n} I_O \left[i + k - 1 - \frac{n-1}{2}, j + l - 1 - \frac{n-1}{2} \right] \tag{4.7}$$

$$M(n)(i,j) = \frac{1}{n^2} \sum_{k=1}^{n} \sum_{l=1}^{n} I_M \left[i + k - 1 - \frac{n-1}{2}, j + l - 1 - \frac{n-1}{2} \right] \tag{4.8}$$

Here, i ranges from 1 to N_x, the number of columns in the domain, and j ranges from 1 to N_y, the number of rows. Fractions are calculated at different spatial scales by varying

the neighbourhood size n. Finally, the Mean Square Error (MSE) for observed and forecast fractions is given by:

$$\mathrm{MSE}_{(n)} = \frac{1}{N_x N_y} \sum_{i=1}^{N_x} \sum_{j=1}^{N_y} \left(M_{(n),ij} - O_{(n),ij} \right)^2 \tag{4.9}$$

The MSE alone is not particularly informative, as it strongly depends on the frequency of the event being evaluated. To address this, a skill score is computed by comparing the MSE to that of a low-skill reference forecast:

$$\mathrm{MSE}_{(n),\mathrm{ref}} = \frac{1}{N_x N_y} \sum_{i=1}^{N_x} \sum_{j=1}^{N_y} \left(M_{(n),ij}^2 + O_{(n),ij}^2 \right) \tag{4.10}$$

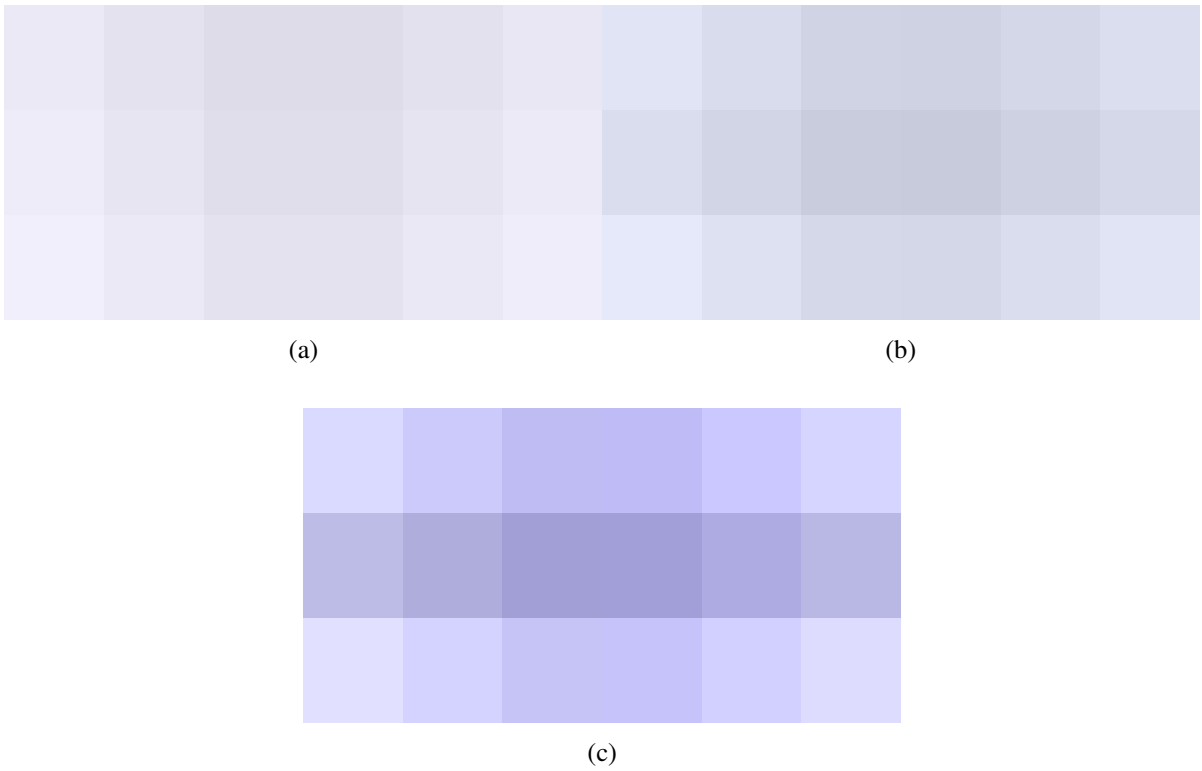
$$\mathrm{FSS}_{(n)} = 1 - \frac{\mathrm{MSE}_{(n)}}{\mathrm{MSE}_{(n),\mathrm{ref}}} \tag{4.11}$$

The FSS ranges from 0 to 1, where a score of 1 indicates perfect skill, and a score of 0 indicates no skill. To account for the varying size of the grid boxes on a latitude–longitude grid, the fraction fields are weighted by the cosine of the latitude during the computation of the MSE.

Figure 4.3: Illustrative example of neighbourhood fraction calculations used in the FSS. Grey grid points indicate exceedance of the chosen threshold, whereas transparent points lie below it. In this example, the observation (radar) field and the forecast field produce identical neighbourhood fractions (6/25). Consequently, although the fields do not coincide at every grid point, the resulting FSS is perfect. Figure taken from Rodwell et al. (2010).

5 Experiment Setup

A total of six sensitivity experiments are conducted using the TEEMLEAP testbed. Three experiments use a globally homogeneous distribution of PSOs: Global500, Global1000 and Global2000, where the number indicates the total number of PSOs distributed over the globe. The other three experiments focus on tropical Africa: Africa2000, Tropics2000 and Tropics730. In these regionally focused experiments, the number indicates the PSO density applied either directly over Africa (Africa2000) or across the entire tropics (Tropics2000 and Tropics730), equivalent to the density that would be applied in a Global2000 or Global730 experiment, while the rest of the globe retains the baseline Global500 density. All experiments are carried out on the HoreKa high-performance computing system. The globally homogeneous experiments were performed by Dr. Jannik Wilhelm, who kindly provided access to the raw forecast files.



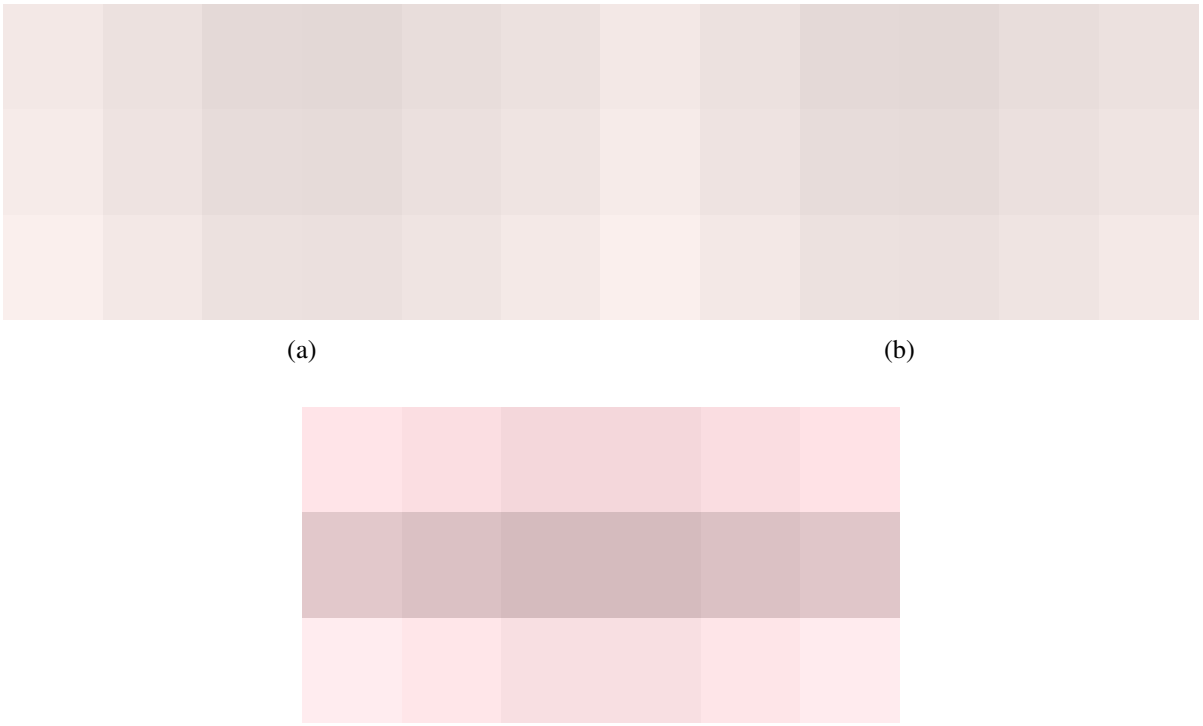
(a) (b)

(c)

Figure 5.1: Location of the PSOs for the global experiments with (a) Global500, (b) Global1000 and (c) Global2000.

Since tropical Africa is the area of interest, the regional experiments are designed to increase observational density there. For Africa2000, a global density of 2000 PSOs is applied within tropical Africa (10° S–25° N, 20° W–50° E), while the rest of the globe retains the baseline Global500 density. Two separate distributions are generated — one including the tropical African box and one with it masked — and then merged to produce the final configuration.

The Tropics2000 and Tropics730 experiments follow the same procedure, but the higher-density region covers the entire tropical belt (23.5° S–23.5° N). Within this belt, densities correspond to Global2000 and Global730, respectively, while the remainder of the globe retains the baseline Global500 density. Tropics730 is specifically designed so that the total number of PSOs matches Africa2000 (589 in total), allowing a controlled comparison of whether forecast skill over Africa benefits more from a denser local network or from moderately increasing observational coverage across the entire tropics. These regional configurations are generated using additional functions implemented in a Python script within the TEEMLEAP testbed.



(a) (b)

(c)

Figure 5.2: Location of PSOs in the experiments with focus on Africa: (a) Africa2000, (b) Tropics730, and (c) Tropics2000.

The general model setup is identical for all experiments, differing only in the number and spatial distribution of PSOs. An assimilation cycle is initialized on 27 August 2022 (00 UTC) using the operational ICON background from the DWD database, interpolated to a horizontal resolution of 40 km (R02B06) with 90 vertical levels.

A two-day spin-up period follows, during which ICON is continuously nudged toward ERA5 by assimilating a dense network of 7200 globally homogeneous and unperturbed PSOs. After the spin-up, PSO profiles (each with 98 vertical levels) from the respective experiment configurations (Figures 5.1 and 5.2) are assimilated every three hours using the 3DVar-based physical-space assimilation system.

During both the spin-up and main experiment phases, the assimilation window and frequency are fixed at three hours, consistent with the availability of background error covariances derived from 3-hour ICON backgrounds. To ensure smooth temporal evolution, analysis increments are applied using ICON's incremental analysis update (IAU) within a symmetric three-hour time window.

Table 5.1: Summary of the six sensitivity experiments, including total PSO number, average distance to the nearest neighbour, and the short names used in this study. For the Africa-specific configurations, two average distances are reported because the PSOs are not uniformly distributed globally: the first value refers to the region with higher PSO density (tropical Africa or the entire Tropics), and the second value refers to the remainder of the globe.

Total observation number	**Average distance to nearest neighbour (km)**	**Experiment name**
500	968.8	Global500
1000	678.8	Global1000
2000	482.6	Global2000
589	482.6 (tropical Africa), 968.8	Africa2000
589	797.2 (Tropics), 968.8	Tropics730
1097	482.6 (Tropics), 968.8	Tropics2000

The PSO profiles are perturbed such that their global observation error statistics match those of real radiosondes. This is achieved by applying level-wise perturbations to the standard PSO profiles, using the corresponding radiosonde error tables from DACE (subsection 4.2.2) as a reference. After each assimilation step, a new 3-hour ICON background is generated.

The assimilation cycling continues for slightly more than one month, ending on 1 October 2022 (00 UTC). Following the assimilation experiments, medium-range deterministic forecasts are generated with ICON (R02B06, 40 km horizontal resolution) for lead times of up to seven days, initialized at 00 UTC over the period 29 August–1 October 2022 (excluding the spin-up period). To allow the background error matrix to adjust, the first three days of forecasts are excluded from validation. Consequently, only forecasts initialized from 1 September onwards are considered for evaluation. The evaluation periods consequently vary with lead time and are summarized in Table 5.2.

Table 5.2: Overview of the evaluation period for different lead times.

Lead time in hours	**Evaluation period**
0h	1 Sept.–1 Oct.
24	2 Sept.–2 Oct.
48	3 Sept.–3 Oct.
72	4 Sept.–4 Oct.
96	5 Sept.–5 Oct.
120	6 Sept.–6 Oct.
144	7 Sept.–7 Oct.
168	8 Sept.–8 Oct.

6 Results

For evaluation, all forecasts and the IMERG data (subsection 4.1.2) are remapped to a regular latitude–longitude grid of $0.5° \times 0.5°$. ERA5 data (subsection 4.1.1) are retrieved from the Climate Data Store directly at the same horizontal resolution. The remapping is performed using the Climate Data Operator (CDO) (Schulzweida 2023). Most variables are remapped using bilinear interpolation (remapbil), while precipitation and TCWV are remapped using first-order conservative interpolation (remapcon) to preserve integral quantities.

All results presented in this chapter correspond to the temporal mean of the 31 forecasts initialized from 1 September to 1 October 2022 at 00 UTC, with evaluation periods depending on lead time as summarized in Table 5.2.

This chapter first presents results for standard meteorological variables, including MSLP, T850, TCWV, U600, V600, and U250. Subsequently, a more detailed evaluation of the precipitation forecast is provided.

The main focus lies on tropical Africa (10° S–25° N, 20° W–50° E). To assess the effect of additional PSOs in a region with markedly different synoptic dynamics, evaluations are also performed for Europe (30° N–75° N, 15° W–42.5° E).

6.1 Standard Variables

Standard meteorological variables are evaluated to quantify the impact of additional PSOs on forecast accuracy in tropical Africa and Europe.

The forecasts are compared to ERA5 data for lead times between 0 and 168 hours (up to 7 days). Note that the 0-hour forecast does not represent the analysis field. In the testbed, an IAU scheme (Reinert et al. 2024) is used, where the analysis increment is applied gradually over a three-hour window (from 90 minutes before to 90 minutes after the nominal start time). This approach improves numerical stability. Consequently, the model integration effectively begins 90 minutes before the nominal start, and the 0-hour forecast corresponds to the model state after 90 minutes of forward integration — that is the midpoint of the IAU period.

To quantify the impact of additional PSOs, Global500 is used as the control experiment. For all other experiments, the RMSE reduction relative to the control is computed as follows:

$$\text{RMSE reduction} = \frac{\text{RMSE}_{\text{control}} - \text{RMSE}_{\text{experiment}}}{\text{RMSE}_{\text{control}}} \times 100 \tag{6.1}$$

Figure 6.1 presents the area-weighted mean RMSE of the control experiment over tropical Africa. As expected, RMSE increases with lead time. For most variables the growth is most

Figure 6.1: RMSE for the control experiment (Global500) over tropical Africa. Shown are (a) MSLP, (b) T850, (c) TCWV, (d) U600, (e) V600 and (f) U250 as a function of forecast lead time.

rapid during the first 24 hours, followed by a more gradual increase. TCWV is a notable exception, exhibiting an almost perfectly linear rise throughout the forecast range. After the initial jump in RMSE during the first few hours, the subsequent error growth is approximately linear for all variables.

A comparison with Europe (Figure 6.2) reveals two striking differences. First, the overall error amplitude is larger over Europe. Variables such as MSLP and T850 are spatially more homogeneous over tropical Africa, making them easier to predict than their extratropical counterparts, where strong frontal gradients in pressure and temperature dominate. Wind speeds are generally higher in the extratropics — at both 600 hPa and 250 hPa — which results in stronger advection. Together with the sharper horizontal gradients typical of mid-latitudes, this means that even small displacement errors translate into larger RMSE values over Europe.

Figure 6.2: RMSE for the control experiment (Global500) over Europe. Shown are (a) MSLP, (b) T850, (c) TCWV, (d) U600, (e) V600 and (f) U250 as a function of forecast lead time.

Second, the shape of the error-growth curves differs between the regions. Over Europe, RMSE increases more slowly during the first 2–3 days (up to roughly 72 hours), then accelerates before gradually levelling off toward days 6–7. This produces a subtle S-shape, consistent with the forecast approaching an error-saturation regime at longer lead times. In contrast, the tropical Africa curves remain more linear throughout. These regional differences reflect the well-established contrast between tropical and extratropical predictability (Judt 2020; Keane et al. 2025).

RMSE reductions of the other experiments relative to the control run for tropical Africa are shown in Figure 6.3. The additional PSOs yield substantial improvements at 0 hour forecasts, reaching up to 35 %. However, the benefit diminishes quickly and by 168 hours the improvements are around 5 % (Global2000). MSLP shows a weaker and noisier response. After roughly 96 hours, most of the advantage of Global2000 over Global1000 has largely vanished.

Figure 6.3: Relative RMSE reduction (in %) with respect to the Global500 control experiment over tropical Africa. Panels show (a) MSLP, (b) T850, (c) TCWV, (d) U600, (e) V600 and (f) U250 as a function of forecast lead time. Results are shown for the Global1000, Global2000, Africa2000, Tropics2000 and Tropics730 experiments.

A notable effect also emerges when comparing two African configurations: the added value of a dense regional network (Africa2000) relative to the broader tropical belt (Tropics730) diminishes with lead time and can disappear or even reverse for some variables (T850, TCWV, U250). The reduced benefit at longer lead times likely reflects tropical teleconnections and wave propagation, highlighting the need for observations across the broader tropical belt.

Over Europe (Figure 6.4), the global experiments (Global1000 and Global2000) are the most relevant cases; African configurations are included for completeness. The initial improvement is comparable to tropical Africa, except for MSLP, where Europe shows larger gains. Unlike in the tropics, however, the RMSE reduction over Europe decreases more linearly with lead time and remains around 10 % after 168 hours (Global2000). The gap between Global2000 and

Figure 6.4: Relative RMSE reduction (in %) with respect to the Global500 control experiment over Europe. Panels show (a) MSLP, (b) T850, (c) TCWV, (d) U600, (e) V600 and (f) U250 as a function of forecast lead time. Results are shown for the Global1000, Global2000, Africa2000, Tropics2000 and Tropics730 experiments.

Global1000 remains evident even on day 7, whereas in tropical Africa the benefit of doubling the number of PSOs is confined largely to the first 96 hours.

Spatial patterns of RMSE are examined for selected variables. Here, TCWV is highlighted because it is a key precursor to precipitation and most of the subsequent analysis focuses on rainfall. Figure 6.5 shows that the largest errors occur along the ITD, where sharp moisture gradients make the field particularly sensitive to small positional or amplitude errors. Although the spatial pattern is broadly similar across all experiments (African-specific experiments not shown), the overall magnitude decreases as the number of PSOs increases, indicating that additional observations primarily reduce the amplitude rather than the spatial structure of the error.

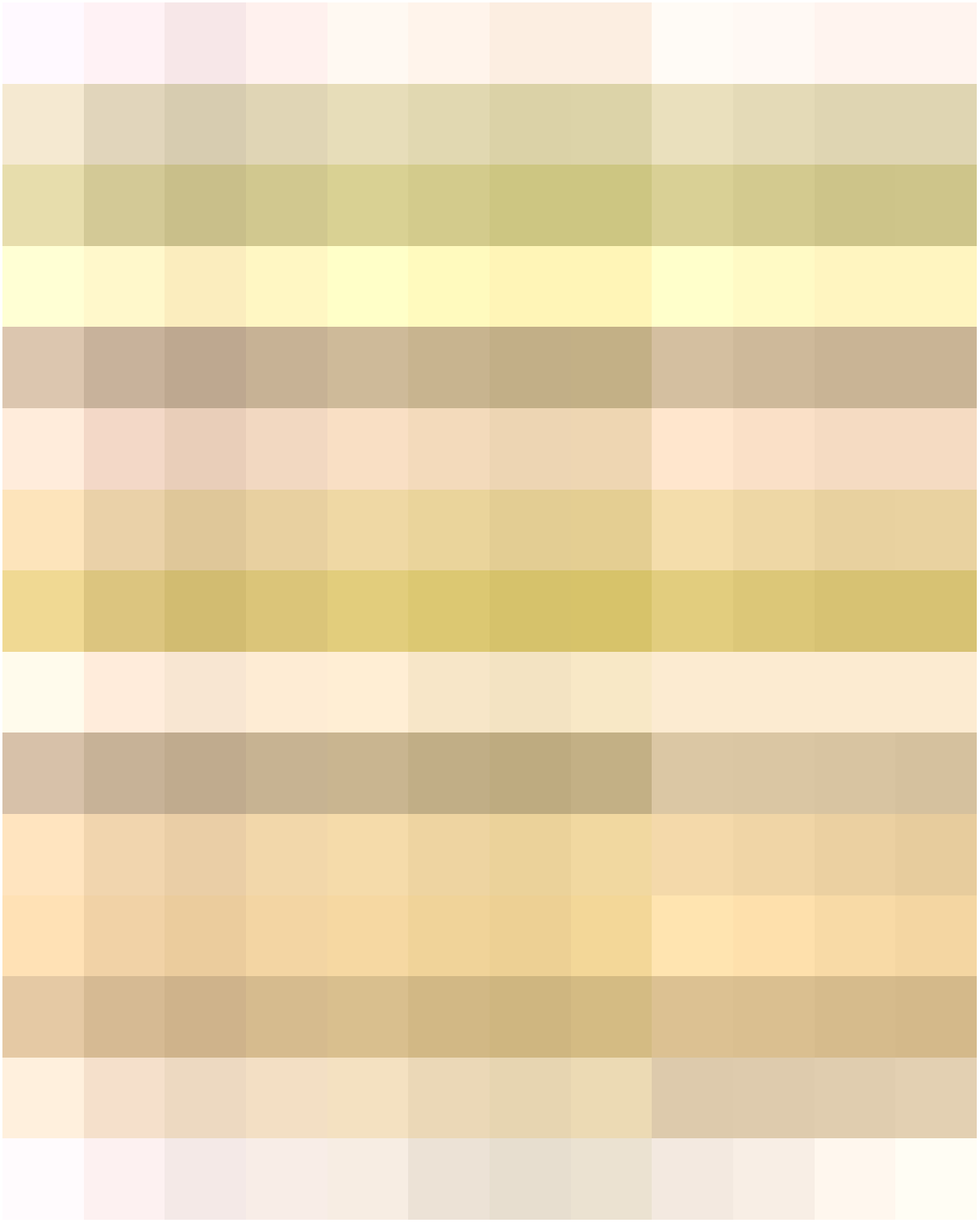


Figure 6.5: Spatial RMSE maps of TCWV over tropical Africa for experiments Global500, Global1000 and Global2000. Lead times are (a–c) 0 h, (d–f) 24 h, (g–i) 48 h, (j–l) 72 h and (m–o) 96 h. Area-weighted mean RMSE values are shown in the bottom-left corner of each panel.

To complement the RMSE, which is unsigned, the corresponding ME fields (Figure 6.6) are analysed to identify systematic positive or negative biases. Initially, the ME is noisy and spatially incoherent. As lead time increases, a clear and persistent pattern emerges: most continental regions exhibit a positive TCWV bias, while parts of the eastern tropical Atlantic,

notably the Congo Basin, show a negative bias. Previous work has shown that humidity fields over West Africa are particularly sensitive to model formulation. For example, Roberts et al. (2015) identify substantial discrepancies in low-level water vapour across different (re)analysis datasets, often linked to shifts in the ITD and to differences in the representation of moist convection. These findings suggest that differences in the way ICON and IFS represent convective and cold-pool processes may also contribute to the TCWV biases observed in this study.

A similar overall positive bias is found in Europe (see Figure A.1 and Figure A.2), but without the pronounced land–ocean contrast observed in the tropics. Instead, the largest positive ME appears over the Mediterranean.

Additional spatial features (not shown) include:

- A persistent RMSE maximum in T850 near Ascension Island, likely linked to boundary-layer and low-cloud uncertainties (Zhang and Zuidema 2019);
- A pronounced RMSE band in U600 over West Africa near10° N, associated with the African Easterly Jet, though slightly displaced southward;
- Elevated RMSE in MSLP over mountainous regions, especially the Ethiopian Highlands, likely reflecting differences in topography between ICON and ERA5;
- In Europe, a strong north–south RMSE gradient in MSLP, with larger errors in the north and smaller in the south, and a distinct maximum over the Icelandic Low region.

In summary, the impact of additional PSOs is evident in all variables examined. The magnitude and persistence of the improvement strongly depend on the region. While spatial error patterns are highly correlated across experiments, error magnitudes decrease systematically with increasing PSO density.

6.2 Precipitation

Precipitation is a difficult quantity to verify for several reasons. Firstly, it is sparsely observed by surface networks and still imperfectly estimated by radar and satellite. Secondly, a point observation may not be representative of a model grid-box average. Thirdly, precipitation has a complex spatio-temporal distribution, often featuring a large number of dry days and occasional very extreme events (Rodwell et al. 2010). Because of this, basic metrics like RMSE are poorly suited for evaluating precipitation. More advanced scores, such as SEEPS and FSS, are therefore applied. SEEPS and FSS are evaluated using 24-hour accumulated precipitation for all lead times. Because the model output provides precipitation accumulated since initialization, the raw forecast values can be used directly only for the 24-hour lead time. For longer lead times, 24-hour accumulations are obtained by differencing forecasts with a 24-hour lead-time separation but the same initialization time, thereby isolating the precipitation accumulated during the most recent 24-hour period.

Especially in tropical Africa, ground-based precipitation records are limited, which introduces substantial uncertainty in rainfall time series. As a result, satellite-based products such as IMERG play a central role in evaluating model performance.

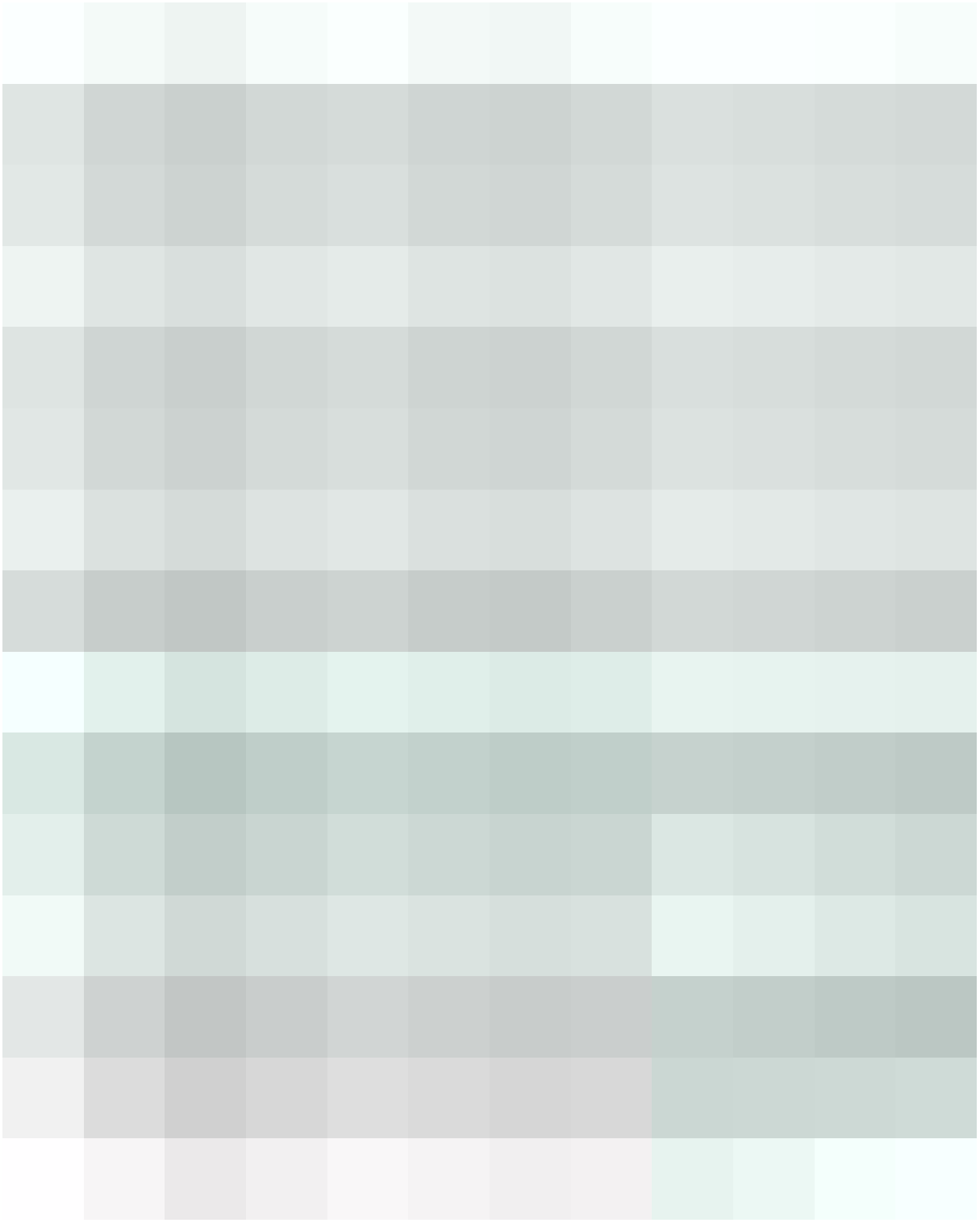


Figure 6.6: Spatial ME maps of TCWV over tropical Africa for experiments Global500, Global1000 and Global2000. Lead times are (a–c) 0 h, (d–f) 24 h, (g–i) 48 h, (j–l) 72 h and (m–o) 96 h. Area-weighted mean ME values are shown in the bottom-left corner of each panel.

6.2.1 IMERG vs. ERA5 data

Comparing IMERG and ERA5 provides important insight into the uncertainty of precipitation estimates in tropical Africa, as differences between the two datasets highlight the limitations of both satellite retrievals and reanalysis products. Figure 6.7 presents monthly accumulated precipitation from IMERG and ERA5 for September, shown as a climatological average and for September 2022. Both datasets consistently capture the position of the West African rain belt around 5°–10° N. Although the area-weighted climatological mean difference between the products is small (1.3 mm), substantial local discrepancies occur.

Figure 6.7: Accumulated monthly precipitation over tropical Africa for September. Panels show (a) IMERG climatology (2000–2021), (b) IMERG 2022, (c) ERA5 climatology (2000–2021), (d) ERA5 2022, (e) differences between IMERG and ERA5 climatology and (f) differences between IMERG and ERA5 in 2022. Area-weighted mean values are shown in the bottom-right corner of each panel.

The largest differences appear over complex terrain. Orographic regions such as the Ethiopian Highlands and the mountains along the Albertine Rift (on the eastern border of the Democratic Republic of the Congo) exhibit much higher precipitation totals in ERA5 than in IMERG. This behaviour is well documented: IMERG often struggles in mountainous regions where retrievals are affected by heterogeneous surface emissivity and shallow orographic rainfall, typically leading to an underestimation of precipitation (Rojas et al. 2021; Bulovic et al. 2020).

Systematic contrasts also emerge between land and ocean. Over land, IMERG tends to produce slightly wetter conditions compared to ERA5 — except in mountainous areas where its underestimation dominates — whereas over the tropical Atlantic and Indian Oceans, ERA5 generally shows higher precipitation amounts. The wetter ocean signal in ERA5 is consistent with findings that the reanalysis tends to overestimate precipitation frequency, particularly over oceanic regions (Duque et al. 2023). This behaviour is linked to the way reanalyses generate precipitation: in ERA5, precipitation is produced by the IFS forecast model, which is known to generate overly frequent drizzle and light rain while underrepresenting intense convective events. This behaviour results in smoother precipitation fields and can explain the higher ERA5 rainfall over the oceans (Wu et al. 2022; Light et al. 2022). IMERG typically performs better over the ocean, where the homogeneous and radiometrically cold surface facilitates the detection of precipitation, while the radiometrically warm and highly variable land surface complicates retrievals (Derin et al. 2021).

Coastal transition zones represent an additional source of disagreement. In September 2022, pronounced differences appear along the Upper Guinea Coast. Such features are consistent with the known challenges IMERG faces in regions where satellite footprints simultaneously include land and ocean surfaces, often resulting in misses or false alarms in these transition zones (Derin et al. 2021). In the same month, IMERG also exhibits a positive anomaly relative to ERA5 within and slightly north of the rain belt, yielding an area-mean difference of 6.2 mm. This pattern is consistent with ERA5's documented underestimation of peak intensities associated with MCSs and tropical convection more generally (Lavers et al. 2022). ERA5 typically performs better in the extratropics, whereas its accuracy declines in the tropics, particularly in summer, when convection is strongest.

Differences between the datasets may also be influenced by IMERG's known cloud-type sensitivities. Previous studies have shown that IMERG performs more reliably for precipitation associated with ice-phase clouds, whereas warm-rain processes are more difficult to detect and can lead to missed events and an underestimation of rainfall (Maranan et al. 2020; Rojas et al. 2021).

Overall, despite notable discrepancies in mountainous and coastal regions, IMERG and ERA5 show broad agreement in the large-scale precipitation patterns over tropical Africa. However, local differences underscore the importance of considering the strengths and limitations specific to the dataset, especially in regions dominated by complex terrain, strong convection, or land–ocean contrasts.

An equivalent plot for Europe is provided in Figure A.3, but is not further discussed here.

6.2.2 Accumulated precipitation over experiment period

To start the evaluation, the rainfall accumulated over the entire experiment period is calculated and compared to ERA5 and IMERG. It should be noted that this period is shifted by one day for each additional day in lead time (see Table 5.2). Figure 6.8 shows the relative difference between area-weighted averages of the forecast and the respective dataset for tropical Africa and Europe as a function of lead time. The relative difference is calculated as:

$$\text{Relative difference} = \frac{\text{Forecast}_{\text{accumulated}} - \text{IMERG/ERA5}_{\text{accumulated}}}{\text{IMERG/ERA5}_{\text{accumulated}}} \times 100 \qquad (6.2)$$

In tropical Africa, the pattern is very similar for IMERG and ERA5, with a pronounced rainfall deficit at short lead times. This deficit diminishes until roughly 96 hours, after which the negative anomaly gradually increases again as lead time extends toward 168 hours. The main difference between IMERG and ERA5 lies in the magnitude of the anomalies, with forecasts showing a larger deficit relative to IMERG than to ERA5. At shorter lead times, experiments with higher PSO counts exhibit a stronger negative bias. However, this effect weakens with increasing lead time.

In Europe, a negative bias is also evident at short lead times, though it is less pronounced than in tropical Africa. For longer lead times, the negative bias gradually shifts to a positive bias, crossing zero between 48 and 144 hours. Interestingly, after 168 hours, the Global2000 experiment almost perfectly matches the accumulated rainfall of IMERG and ERA5. Similar to tropical Africa, more PSOs lead to a stronger negative bias at 24 hours, but the difference between the Global1000 and Global2000 experiments are minimal. The Global500 experiment, on the other hand, generally produces the wettest forecasts, except at 168 hours.

The consistent negative rainfall bias across both regions and all experiments at short lead times may reflect a potential adjustment process between two different "model" worlds. During the assimilation cycle, the ICON background is nudged toward the ERA5 state, which can create an initial "shock" that requires some time to stabilize. Rainfall depends on multiple meteorological variables, including TCWV, winds, pressure, and convergence, as well as model-specific factors such as convection parametrization and rain efficiency. As noted in section 6.1, TCWV exhibits a positive bias over most continental regions at short lead times, consistent with previous findings that low-level water vapour is highly sensitive to model formulation (Roberts et al. 2015). This suggests a possible mechanism for the rainfall deficit: ICON may require more atmospheric moisture to generate rainfall, but nudging toward ERA5 temporarily suppresses moisture, delaying the model's return to its typical conditions. While speculative, this interpretation highlights a potentially interesting interaction between assimilation-driven initial conditions and rainfall forecasts.

6.2.3 Evaluation of precipitation forecasts using SEEPS

To derive probabilities of dry (p_1), light rain (p_2), and heavy rain (p_3) categories, as well as the threshold separating p_2 and p_3, climatological reference data from September 2000–2021 is used. This is done separately for IMERG and ERA5 so that forecasts can be evaluated

Figure 6.8: Relative difference (in %) between accumulated rainfall in forecasts and reference datasets as a function of lead time over the 31-day experiment period. Upper panels show results for tropical Africa using (a) IMERG and (b) ERA5 as reference datasets. Lower panels show results for Europe using (c) IMERG and (d) ERA5. Positive values indicate overestimation by the forecasts, while negative values indicate underestimation.

consistently against each dataset using its own climatology. Grid points with a dry-category probability of $p_1 > 0.85$ are masked and excluded from the evaluation, as such locations are dominated by dry conditions.

Values with $p_1 < 0.1$ are not masked — otherwise many tropical regions would be excluded simply because rainfall occurs frequently and truly dry days are rare. Instead, these low p_1 values are set to a fixed minimum of 0.1 to ensure numerical stability of the SEEPS score.

Figure 6.9 presents the resulting fields of p_1 and the threshold between light and heavy rain for tropical Africa. Blank areas indicate grid points where $p_1 > 0.85$; since these locations are dominated by dry conditions, they are excluded from the evaluation. Over the tropical Atlantic, IMERG and ERA5 differ markedly: ERA5 shows far fewer masked points, implying many more rainy days, whereas IMERG classifies much of the region as predominantly dry. The threshold between light and heavy rain differs substantially in mountainous regions, where ERA5 produces considerably higher values. IMERG, in contrast, identifies light rainfall even in parts of the western Sahara. An equivalent plot for Europe is provided in Figure A.4, but is not discussed here.

Figure 6.9: Climatological classification of dry probability (p_1) and the threshold between light (p_2) and heavy rain (p_3) used for the SEEPS score over tropical Africa. The upper row shows p_1 for (a) IMERG and (b) ERA5. The bottom row presents the thresholds between p_2 and p_3 for (c) IMERG and (d) ERA5. Blank areas are excluded from the SEEPS evaluation, as the frequency of rainy days is too low in these regions.

Figure 6.10 shows the area- and time-averaged SEEPS score for the control experiment over tropical Africa, evaluated against IMERG and ERA5. In addition to the overall score, the figure shows the frequencies of the combinations of forecast and observed categories (dry = 1, light rain = 2, heavy rain = 3) in percent.

It should be noted that the SEEPS scores against IMERG and ERA5 are not strictly comparable. Differences in underlying climatology influence the entries of the error matrix, and due to restrictions in p_1, the number of grid points considered for evaluation differs between the two datasets. This is reflected in the scores: combinations involving forecast errors generally occur more frequently with ERA5, but the corresponding SEEPS scores are slightly lower than those against IMERG.

Figure 6.10: Area- and time-averaged SEEPS score for the control experiment (Global500) evaluated against IMERG (blue) and ERA5 (red) over tropical Africa, shown as a function of forecast lead time. The top-left panel shows the SEEPS score, while the top-right panel displays the frequency of forecasts and observations falling into the same category. The remaining six panels show the frequencies of different combinations of forecast and observed categories, where 1 represents dry, 2 light rain and 3 heavy rain. All frequencies are expressed in percent.

Despite these differences, the temporal evolution of the SEEPS score with lead time shows a similar pattern between the two evaluations, although absolute levels differ. The score is lowest at the shortest lead times, increases rapidly during the first 24 hours, and then exhibits a more gradual change at longer lead times. The frequency of exact matches between forecast and observation (top-right panel) decreases steadily with lead time, starting from values above 50 %.

Notable differences between the evaluation against IMERG and ERA5 are observed in the cross-category combinations between dry (1) and light rain (2):

- Forecast 1, Observation 2 occurs more frequently when evaluated against ERA5;
- Forecast 2, Observation 1 occurs more frequently when evaluated against IMERG.

These findings highlight differences in the representation of light rainfall events between the two evaluation datasets.

Also note that forecasts that are too dry — with the exception of Forecast 2, Observation 3 — are more frequent at shorter lead times. This is consistent with the findings in subsection 6.2.2, which showed a deficit in accumulated rainfall for short lead times.

Over Europe, the SEEPS score of the forecasts is substantially lower when evaluated against IMERG compared to ERA5, particularly at short lead times (Figure A.5). The percentage of perfect matches initially exceeds 60 %. Compared to tropical Africa, the SEEPS score of the forecast with respect to IMERG is significantly lower at short lead times, but increases steadily and is higher at a lead time of 168 hours. When evaluated against ERA5, the SEEPS score evolves more gradually, starting from a higher initial level.

Relative improvements in the SEEPS score and in the frequencies of category combinations, evaluated against IMERG over tropical Africa, are presented in Figure 6.11. The relative improvement is defined such that positive values indicate an improvement. Accordingly, for the SEEPS score and for the frequencies of cross-category occurrences (which correspond to error terms), an improvement corresponds to a reduction and is therefore computed using Equation (6.3). Conversely, for perfect matches, an improvement corresponds to an increase and is computed using Equation (6.4).

$$\text{Relative improvement} = \frac{\text{SEEPS}_{\text{control}} - \text{SEEPS}_{\text{experiment}}}{\text{SEEPS}_{\text{control}}} \times 100 \tag{6.3}$$

$$\text{Relative improvement} = \frac{\text{SEEPS}_{\text{experiment}} - \text{SEEPS}_{\text{control}}}{\text{SEEPS}_{\text{control}}} \times 100 \tag{6.4}$$

Compared to the standard meteorological variables discussed in section 6.1, the signals in the SEEPS-based evaluation are noisier and more difficult to interpret. The SEEPS score shows a slight improvement for most experiments; only Tropics730 exhibits a negative impact at 24-hour lead time. The largest improvements occur between 48 and 96 hours, with gains of up

Figure 6.11: Relative SEEPS score improvement (in percent) for all experiments compared to the control experiment (Global500), evaluated against IMERG over tropical Africa. Panel layout is the same as in Figure 6.10, with the top-left panel showing the overall SEEPS score improvement and the other panels showing changes in category combination frequencies as a function of lead time. Positive values indicate an improvement relative to the control run.

to 3 %. Beyond this range, the improvement decreases and at 168 hours the scores converge toward those of the control experiment.

The percentage of perfect matches increases by more than 3 %, this benefit weakens after 96 hours of lead time. In several experiments, too-dry forecasts become more frequent at short lead times.

The most substantial improvements are found in the heavy-rain category. Mismatches between heavy rain and the lower categories (no rain or light rain) are reduced by up to 30 % and 15 %, respectively, at short lead times. This benefit diminishes with increasing lead time, but generally remains positive.

For longer lead times, Tropics730 generally outperforms Africa2000, likely because the wider observational coverage of the tropical belt provides a more robust constraint on the forecast than the locally dense network in tropical Africa.

Overall, additional PSOs provide a limited but noticeable improvement in tropical Africa. The Global2000 experiment performs slightly better than Global1000 in most cases. However, when evaluated against ERA5 (not shown), these improvements become even less pronounced.

The results over Europe are more distinct and generally show larger improvements, particularly in the evaluation against IMERG (Figure 6.12). For the Global2000 experiment, the SEEPS score improves consistently by 5–10 % across all lead times. The frequency of perfect matches increases by approximately 5 %. Similar to the findings for tropical Africa, the largest improvements occur in the heavy-rain category at short lead times, where the reduction in the forecast 3, observation 1 mismatch exceeds 30 %. The corresponding forecast 1, observation 3 mismatch also shows substantial improvement, reaching up to 30 %.

In Europe, increasing the amount of PSO is clearly beneficial: Global2000 outperforms both Global1000 and Global500 across lead times. When evaluated against ERA5 (not shown), the improvements are smaller, particularly at short lead times.

6.2.4 Evaluation of precipitation forecasts using FSS

The second metric used to evaluate precipitation forecasts is the FSS. The FSS is evaluated primarily against IMERG; results using ERA5 are explicitly noted where relevant. Starting from the native $0.5° \times 0.5°$ grid, the spatial neighbourhood size is gradually increased up to 10° in steps of 0.5°. Four daily precipitation thresholds are considered: 1 mm, 5 mm, 10 mm and 20 mm. Figure 6.13 shows the FSS over tropical Africa for the three global experiments (Global500, Global1000 and Global2000) as a function of lead time and spatial scale. Not all neighbourhood sizes are shown to maintain readability.

As expected, forecast skill increases with decreasing lead time and larger spatial scales. Higher precipitation thresholds reduce skill because distinguishing “event” from “non-event” becomes increasingly difficult. For the 1 mm threshold, the FSS remains above 0.5 across all lead times and scales (green shading), whereas for the 20 mm threshold the skill is substantially lower (predominantly brown shading).

Figure 6.12: Relative SEEPS score improvement (in percent) for all experiments compared to the control experiment (Global500), evaluated against IMERG over Europe. Panel layout is the same as in Figure A.5, with the top-left panel showing the overall SEEPS score improvement and the other panels showing changes in category combination frequencies as a function of lead time. Positive values indicate an improvement relative to the control run.

A notable feature is that for the 1 mm threshold the FSS at 48 hours occasionally exceeds that at 24 hours. This behaviour may partly reflect the dry bias at short lead times discussed in subsection 6.2.2. The evaluation against ERA5 (not shown) exhibits a similar structure but with slightly reduced skill.

Over Europe, the FSS increases consistently with shorter lead times and larger spatial scales (Figure A.6). For the 1 mm and 5 mm thresholds, skill is generally lower in Europe than in tropical Africa. This contrast may be related to the much sharper climatic gradients in tropical

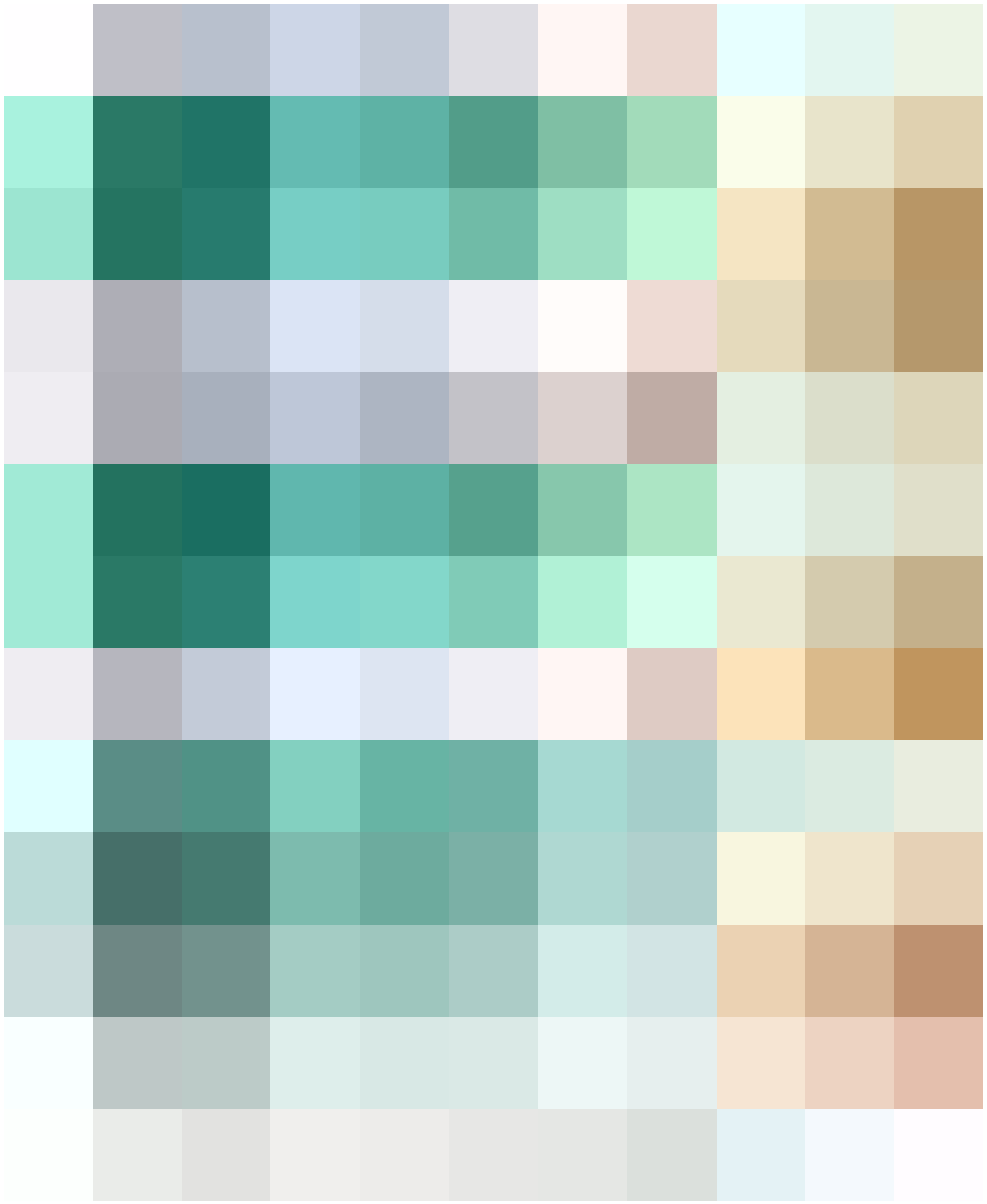


Figure 6.13: FSS over tropical Africa as a function of lead time and spatial scale for thresholds of 1 mm, 5 mm, 10 mm and 20 mm. Forecasts are evaluated against IMERG to compute the FSS. Results are shown for Global500 (top row), Global1000 (middle row) and Global2000 (bottom row).

Africa — from the dry Sahara to the convectively active rainbelt — whereas European precipitation is more uniformly distributed and strongly influenced by synoptic-scale dynamics such as Rossby waves and baroclinic instability. This distinction is also visible in the climatological dry-day probabilities shown in Figure 6.9 and Figure A.4: tropical Africa exhibits a bimodal distribution (either very dry or very wet), whereas Europe contains many regions with intermediate probabilities.

When evaluated against ERA5 (not shown), the FSS over Europe decreases substantially for all thresholds, consistent with the larger IMERG–ERA5 discrepancies already observed in the SEEPS analysis (subsection 6.2.3).

To assess the impact of additional PSOs, differences in FSS between experiments are computed. Results for the global experiments are shown in Figure 6.14. Both Global1000 and Global2000 exhibit modest improvements relative to Global500, with the strongest gains at higher thresholds. Differences between Global2000 and Global1000 are small: at lower thresholds (1 mm and 5 mm) the experiments perform similarly, while at higher thresholds (10 mm and 20 mm) the relative advantage oscillates between the two. This suggests the presence of a saturation point over tropical Africa, beyond which additional increases in observation density do not yield noticeable improvements in precipitation forecast skill.

Results for the regional experiments are given in Figure 6.15. Tropics2000 consistently outperforms Tropics730 (top row), with differences increasing toward higher thresholds. The comparison between Tropics2000 and Africa2000 (middle row) also favours Tropics2000, especially at longer lead times. The comparison between Africa2000 and Tropics730 (bottom row) indicates that a dense local network over tropical Africa (Africa2000) provides the greatest benefit at short lead times (1 day). However, this benefit weakens and reverses at longer lead times, particularly for higher thresholds.

In Europe, the benefit of additional PSOs is considerably larger (Figure 6.16). Here, forecast skill increases more systematically with observation density, and Global2000 clearly outperforms Global1000. Improvements are strongest at longer lead times and for high thresholds, suggesting that better representation of baroclinic systems and frontal structures in the initial conditions has a substantial impact on precipitation forecast performance.

Using ERA5 as the evaluation dataset (not shown) yields little change for tropical Africa but leads to a different ranking over Europe: for thresholds of 5 mm, 10 mm and 20 mm, Global1000 outperforms Global2000, whereas for the 1 mm threshold Global2000 remains the best-performing experiment. This again reflects the sensitivity of precipitation verification to the choice of the observational dataset.

Overall, the FSS results suggest that additional PSOs provide clear benefits for European precipitation forecasts, whereas improvements in tropical Africa are more modest and less systematic. This indicates that model errors related to convection play a large role in the tropics, potentially exceeding the influence of initial-condition uncertainty.

Figure 6.14: Differences in FSS over tropical Africa, computed against IMERG, as a function of lead time and spatial scale. Shown are Global1000–Global500 (top row), Global2000–Global500 (middle row) and Global2000–Global1000 (bottom row). Green shading indicates better performance of the first experiment in each pair, and purple shading indicates better performance of the second.

Figure 6.15: Differences in FSS over tropical Africa, computed against IMERG, as a function of lead time and spatial scale. Shown are Tropics2000–Tropics730 (top row), Tropics2000–Africa2000 (middle row), and Africa2000–Tropics730 (bottom row). Green shading indicates better performance of the first experiment in each pair, while purple shading indicates better performance of the second.

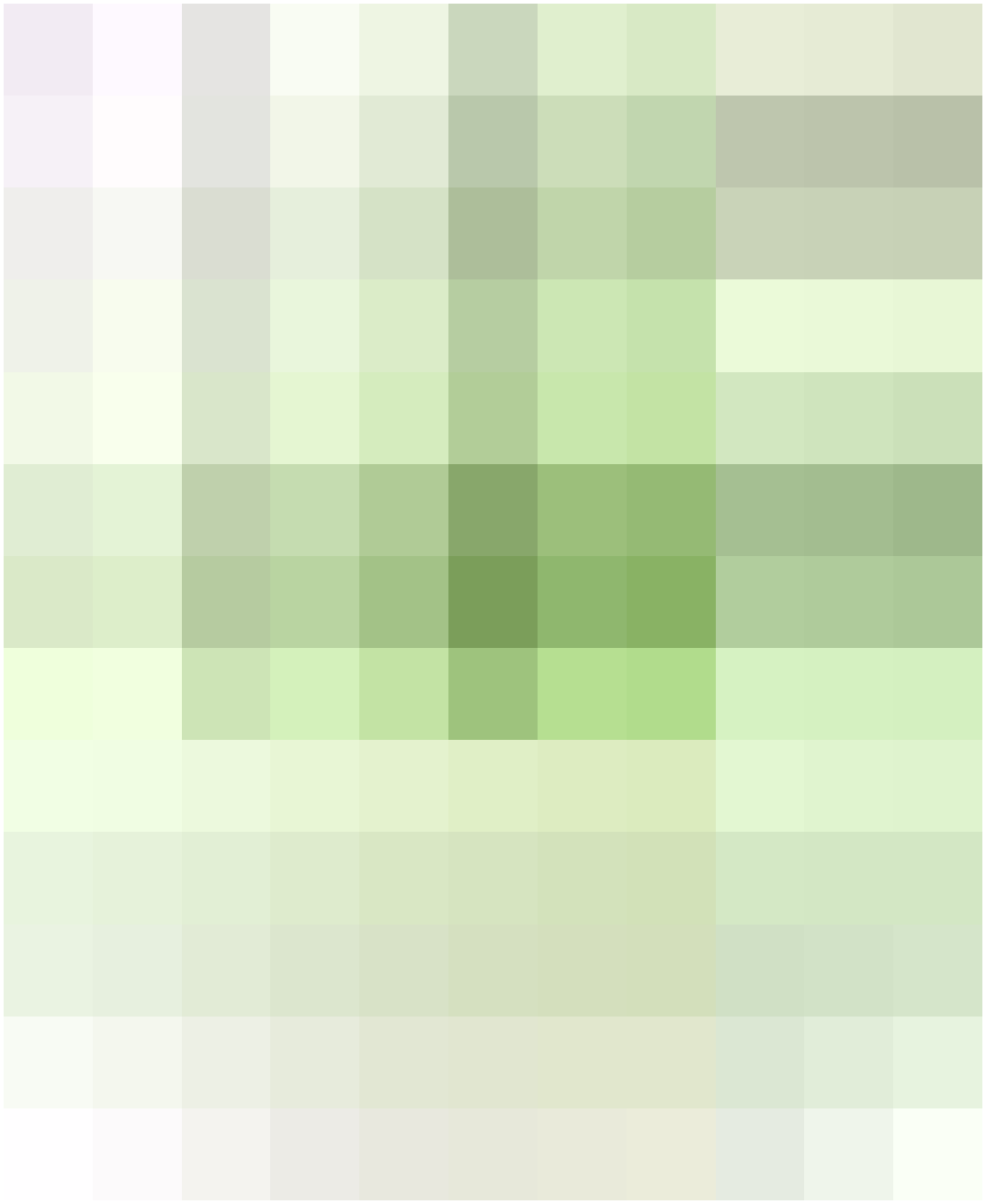


Figure 6.16: Differences in FSS over Europe, computed against IMERG, as a function of lead time and spatial scale. Shown are Global1000–Global500 (top row), Global2000–Global500 (middle row) and Global2000–Global1000 (bottom row). Green shading indicates better performance of the first experiment in each pair, while purple shading indicates better performance of the second.

7 Conclusions

This study investigated how observational coverage influences the accuracy of weather forecasts over tropical Africa, with Europe included for comparison. Using the TEEMLEAP testbed, a set of controlled observing-system experiments was conducted in which the number and spatial distribution of radiosonde-like pseudo-observations (PSOs) derived from ERA5 were systematically varied. The experiments included three globally homogeneous configurations (Global500, Global1000 and Global2000) and three regional setups focused on tropical Africa (Africa2000, Tropics2000 and Tropics730). Forecasts were produced with the ICON model for lead times of up to seven days, using Global500 as the baseline to assess relative changes in forecast skill. The evaluation covered key meteorological variables, including mean sea level pressure (MSLP), temperature at 850 hPa (T850), total column water vapour (TCWV), zonal and meridional winds at 600 hPa (U600, V600) and zonal wind at 250 hPa (U250), with a particular focus on precipitation due to its high societal relevance in tropical Africa.

For variables such as MSLP, T850, TCWV, and winds, error growth in tropical Africa is largest during the first two forecast days. In contrast, error growth in Europe is initially slower but accelerates between days 2 and 5. This latitude-dependent evolution is consistent with the findings of Judt (2020) and Keane et al. (2025).

Across variables, increasing the number of PSOs reduces RMSE by up to 30 % at short lead times. The benefit decreases with time — approximately exponentially in tropical Africa and more linearly in Europe — yet remains positive for most experiments and variables, typically around 5 %. After roughly four days, the advantage of Global2000 over Global1000 in tropical Africa largely vanishes, while substantial differences persist across the full seven-day forecast horizon in Europe.

The exponential decay of forecast improvements in tropical Africa agrees with findings from data-denial experiments conducted with observations of the AMMA and DACCIWA campaigns (Agustí-Panareda et al. 2010; van der Linden et al. 2020). These similarities suggest that sparse observational coverage contributes to forecast errors in the region but is only one component among several limiting factors.

Precipitation is a central focus of this study. The limited availability of ground-based rainfall observations in tropical Africa introduces considerable uncertainties, clearly reflected in discrepancies between IMERG and ERA5, particularly in mountainous and coastal regions. Forecasts of accumulated rainfall over the entire experiment period exhibit a dry bias at short lead times relative to both datasets, with the deficit becoming larger as the number of assimilated PSOs increases. Since ICON is tuned for operational settings using diverse real-world observations, replacing these with ERA5-derived PSOs may have introduced imbalances during assimilation. The TCWV bias identified in this study may partly contribute to the dry precipitation bias; however, fully understanding these interactions requires further investigation.

Moreover, global ICON forecasts cannot resolve convection explicitly and key precipitation-related processes must be parametrized.

Verification with SEEPS shows that increasing the number of PSOs provide modest improvements in daily precipitation forecasts over tropical Africa. Typical improvements are 1–3 % relative to Global500, with much larger gains for heavy-precipitation events (up to 30 %). In Europe, SEEPS improvements are generally larger (5–10 %) and again most pronounced for heavy rainfall.

The FSS results are broadly consistent with the SEEPS findings. Over tropical Africa, both Global1000 and Global2000 perform slightly better than Global500, especially for higher thresholds (10 mm and 20 mm per day). Differences between Global2000 and Global1000 are small and often ambiguous.

A clearer picture emerges when comparing experiments with regionally targeted PSO distributions. Tropics2000 consistently outperforms Tropics730 and Africa2000, particularly at longer lead times. A dense local network (Africa2000) appears effective at short lead times, but from 48 hours onwards accurate information from surrounding regions (Tropics730) becomes essential — a result that highlights the interconnectedness of tropical weather systems.

In Europe, the impact of increased observational density on FSS is substantially larger. Skill improvements are strongest for higher thresholds and longer lead times. Global2000 clearly outperforms both Global1000 and Global500.

These regional contrasts align with findings by Borne et al. (2025), who also report noticeable impacts of improved wind observations from Aeolus on extratropical precipitation forecasts but negligible benefits in the tropics. Their study also shows a tendency for larger improvements at longer lead times, although this effect is more pronounced at smaller spatial scales than in the present work.

Taken together, these findings directly address the research questions posed in Chapter 1, providing insight into forecast sensitivity, the spatial scales that matter, and the characteristics of an effective observational network for tropical Africa. A schematic overview of the main results is provided in Figure 7.1.

In summary, additional PSOs improve forecast quality in both regions, but the magnitude and persistence of the improvements vary substantially. Benefits in Europe are larger and extend through longer lead times, whereas improvements in tropical Africa saturate quickly, particularly for precipitation. This suggests that resolving large-scale tropical waves with a spacing of around 700 km (comparable to Global1000) may already be sufficient for forecasts beyond a few days, whereas on smaller scales model errors and parametrization uncertainties dominate. This interpretation is consistent with studies showing that explicit convection or advanced data-driven approaches can substantially enhance tropical precipitation forecasts (Pante and Knippertz 2019; Walz et al. 2024b).

A central limitation of this study is the use of PSOs instead of real-world observations. Although this approach is necessary for a controlled observing-system experiment, it can introduce inconsistencies between the ICON model and the ERA5-derived profiles. ERA5 is produced with ECMWF's IFS, whose numerics, physical parameterizations, and data-assimilation system differ from those of ICON. As a result, PSOs may contain systematic biases incompatible with ICON's internal climatology, potentially leading to imbalances during assimilation.

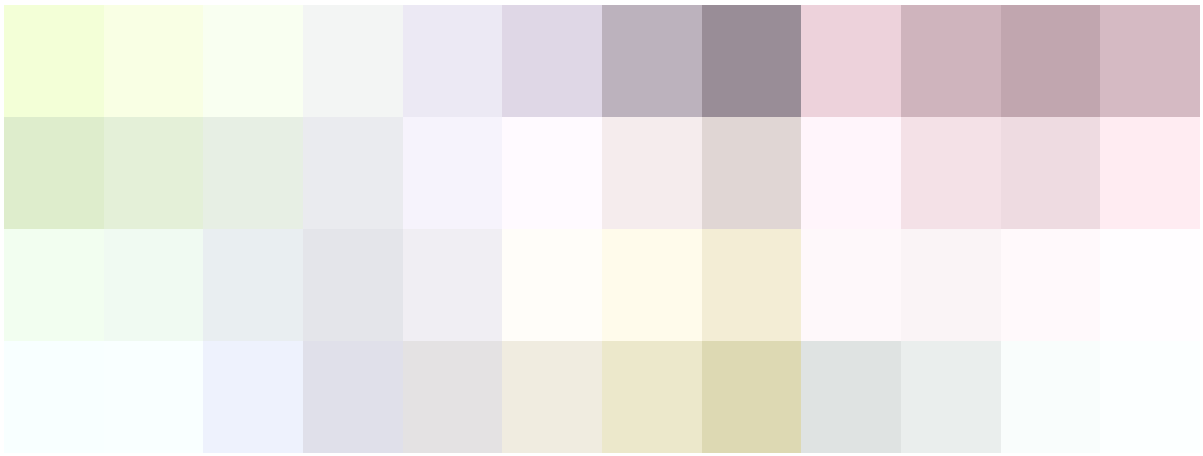


Figure 7.1: Summary of the key findings of this study.

Nevertheless, the TEEMLEAP testbed provides a powerful, low-cost environment for exploring hypothetical observational networks and evaluating their impact.

A further limitation is the restriction to a single month (September 2022) and two verification datasets (IMERG and ERA5), both of which carry considerable uncertainties in data-sparse regions. The pronounced IMERG–ERA5 discrepancies highlight the importance of ongoing initiatives such as SOFF and TAHMO for improving in-situ observations in Africa. Without high-quality reference datasets, verification is uncertain and model development becomes more challenging.

As noted by Judt (2020), the intrinsic limit of predictability in the tropics has not yet been reached by NWP systems, indicating substantial room for improvement. A key open question highlighted by this study is the extent to which the limited forecast skill in tropical Africa arises from insufficient observational coverage versus model-intrinsic errors. Future work could therefore repeat these experiments with convection-permitting models to better disentangle the relative contributions of initial condition uncertainty and model shortcomings — particularly those related to convective parametrization — to forecast performance. Testing data-driven rainfall-forecasting methods, such as the U-Net approach of Walz et al. (2024b), could provide additional insight into sensitivity to available observations. Extending the analysis to multiple seasons or years would further clarify the robustness of the conclusions reached here.

Given the societal importance of accurate rainfall forecasts and the ongoing challenges in achieving them, every incremental improvement is valuable. The results of this study highlight that enhanced observational coverage can contribute meaningfully to better forecasts, even if model limitations remain. Continued progress in observational networks, model physics, numerical resolution, and our understanding of tropical dynamics will hopefully pave the way toward more reliable precipitation predictions.

Abbreviations

3D-Var	Three-Dimensional Variational
4D-Var	Four-Dimensional Variational
AEJ	African Easterly Jet
AEW	African Easterly Wave
AMMA	African Monsoon Multidisciplinary Analysis
ASM	Assimilation Cycle
BACY	Basic Cycling Environment
CDO	Climate Data Operator
CNN	Convolutional Neural Network
CRPSS	Continuous Ranked Probability Skill Score
DACE	Data Assimilation Coding Environment
DACCIWA	Dynamics–Aerosol–Chemistry–Cloud Interactions in West Africa
DKE	Difference Kinetic Energy
DWD	Deutscher Wetterdienst (German Weather Service)
EasyUQ	Easy Uncertainty Quantification
ECMWF	European Centre for Medium-Range Weather Forecasts
ERA5	ECMWF Reanalysis v5
FSS	Fractions Skill Score
IAU	Incremental Analysis Update
ICON	ICOsahedral Nonhydrostatic
IFS	Integrated Forecasting System
IMERG	Integrated Multi-satellitE Retrievals for GPM
IPCC	Intergovernmental Panel on Climate Change
IR	Infrared
ITD	Intertropical Discontinuity
KIT	Karlsruhe Institute of Technology
MCS	Mesoscale Convective System

MSLP	Mean Sea Level Pressure
NWP	Numerical Weather Prediction
PMW	Passive Microwave
PSO	Pseudo-Observation
RMSE	Root Mean Square Error
SHL	Saharan Heat Low
SEEPS	Stable Equitable Error in Probability Space
SOFF	Systematic Observations Financing Facility
T850	Temperature at 850 hPa
TCWV	Total Column Water Vapour
TEEMLEAP	TEstbed for Exploring Machine LEarning in Atmospheric Prediction
TEJ	Tropical Easterly Jet
TRMM	Tropical Rainfall Measuring Mission
U250	Zonal Wind at 250 hPa
U600	Zonal Wind at 600 hPa
UN	United Nations
UTC	Coordinated Universal Time
V600	Meridional Wind at 600 hPa
WAM	West African Monsoon
WMO	World Meteorological Organization

Bibliography

Abbe, C., 1901: The physical basis of long-range weather forecasts. *Monthly Weather Review*, **29**, 551–561.

Agustí-Panareda, A., Beljaars, A., Cardinali, C., Genkova, I., and Thorncroft, C., 2010: Impacts of Assimilating AMMA Soundings on ECMWF Analyses and Forecasts. *Weather and Forecasting*, **25** (**4**), 1142–1160, https://doi.org/10.1175/2010WAF2222370.1.

Andersson, E. and Masutani, M., 2010: Collaboration on Observing System Simulation Experiments (Joint OSSE), https://doi.org/10.21957/62GAYQ76.

Bauer, P., Thorpe, A., and Brunet, G., 2015: The quiet revolution of numerical weather prediction. *Nature*, **525** (**7567**), 47–55, https://doi.org/10.1038/nature14956.

Bick, T. et al., 2016: Assimilation of 3D radar reflectivities with an ensemble Kalman filter on the convective scale. *Quarterly Journal of the Royal Meteorological Society*, **142** (**696**), 1490–1504, https://doi.org/10.1002/qj.2751.

Bjerknes, V., 1904: Das problem der wettervorhersage betrachtet vom standpunkt der mechanik und physik. *Meteorologische Zeitschrift*, **21**, 1–7.

Borne, M., Knippertz, P., Rennie, M., and Weissmann, M., 2025: The impact of Aeolus observations on wind and rainfall predictions. *EGUsphere*, 1–25, https://doi.org/10.5194/egusphere-2025-5219.

Borne, M., Knippertz, P., Weissmann, M., Martin, A., Rennie, M., and Cress, A., 2023: Impact of Aeolus wind lidar observations on the representation of the West African monsoon circulation in the ECMWF and DWD forecasting systems. *Quarterly Journal of the Royal Meteorological Society*, **149** (**752**), 933–958, https://doi.org/10.1002/qj.4442.

Bulovic, N., McIntyre, N., and Johnson, F., 2020: Evaluation of IMERG V05B 30-Min Rainfall Estimates over the High-Elevation Tropical Andes Mountains. *Journal of Hydrometeorology*, **21** (**12**), 2875–2892, https://doi.org/10.1175/JHM-D-20-0114.1.

Cornforth, R. et al. (2017), "Synoptic Systems", in: *Meteorology of Tropical West Africa*, John Wiley & Sons, Ltd, pp. 40–89, https://doi.org/10.1002/9781118391297.ch2.

Derin, Y., Kirstetter, P.-E., and Gourley, J. J., 2021: Evaluation of IMERG Satellite Precipitation over the Land–Coast–Ocean Continuum. Part I: Detection. *Journal of Hydrometeorology*, **22** (**11**), 2843–2859, https://doi.org/10.1175/JHM-D-21-0058.1.

Desroziers, G., Berre, L., Chapnik, B., and Poli, P., 2005: Diagnosis of observation, background and analysis-error statistics in observation space. *Quarterly Journal of the Royal Meteorological Society*, **131** (**613**), 3385–3396, https://doi.org/10.1256/qj.05.108.

Duque, E. M., Huang, Y., May, P. T., and Siems, S. T., 2023: An Evaluation of IMERG and ERA5 Quantitative Precipitation Estimates over the Southern Ocean Using Shipborne Ob-

servations. *Journal of Applied Meteorology and Climatology*, **62** (**11**), 1479–1495, https://doi.org/10.1175/JAMC-D-23-0039.1.

ECMWF (2025a), ECMWF | Charts, URL: https://charts.ecmwf.int/catalogue/packages/monitoring/products/dcover?Flag=used&base_time=202510300000&obs=Temp.

– (2025b), ECMWF SOFF Impact Experiments: June 2025, URL: https://www.un-soff.org/wp-content/uploads/2025/06/ECMWF-SOFF-Impact-Experiments-June-2025.pdf.

– (2025c), Fifty Years of Data Assimilation at ECMWF, URL: https://www.ecmwf.int/sites/default/files/elibrary/81650-fifty-years-of-data-assimilation-at-ecmwf.pdf.

– (2025d), Fifty Years of Earth System Modelling at ECMWF, URL: https://www.ecmwf.int/sites/default/files/elibrary/81651-fifty-years-of-earth-system-modelling-at-ecmwf.pdf.

Errico, R. M. et al., 2013: Development and validation of observing-system simulation experiments at NASA's Global Modeling and Assimilation Office. *Quarterly Journal of the Royal Meteorological Society*, **139** (**674**), 1162–1178, https://doi.org/10.1002/qj.2027.

Faccani, C. et al., 2009: The Impacts of AMMA Radiosonde Data on the French Global Assimilation and Forecast System. *Weather and Forecasting*, **24** (**5**), 1268–1286, https://doi.org/10.1175/2009WAF2222237.1.

Fink, A. H. et al. (2017), "Mean Climate and Seasonal Cycle", in: *Meteorology of Tropical West Africa*, John Wiley & Sons, Ltd, pp. 1–39, https://doi.org/10.1002/9781118391297.ch1.

Fischer, M. A. (2025), Surrogate-based uncertainty quantification and parameter optimization in simulations of the West African monsoon, https://doi.org/10.5445/IR/1000182304.

González, Á., 2010: Measurement of Areas on a Sphere Using Fibonacci and Latitude–Longitude Lattices. *Mathematical Geosciences*, **42** (**1**), 49–64, https://doi.org/10.1007/s11004-009-9257-x.

Hersbach, H. et al., 2020: The ERA5 global reanalysis. *Quarterly Journal of the Royal Meteorological Society*, **146** (**730**), 1999–2049, https://doi.org/10.1002/qj.3803.

Houtekamer, P. L., 1993: Global and Local Skill Forecasts. *Monthly Weather Review*, **121** (**6**), 1834–1846, https://doi.org/10.1175/1520-0493(1993)121<1834:GALSF>2.0.CO;2.

Houtekamer, P. L., Lefaivre, L., Derome, J., Ritchie, H., and Mitchell, H. L., 1996: A System Simulation Approach to Ensemble Prediction. *Monthly Weather Review*, **124** (**6**), 1225–1242, https://doi.org/10.1175/1520-0493(1996)124<1225:ASSATE>2.0.CO;2.

Huffman, G. J. (2023), NASA Global Precipitation Measurement (GPM) Integrated Multi-satellitE Retrievals for GPM (IMERG) Version 07, URL: https://gpm.nasa.gov/sites/default/files/2023-07/IMERG_V07_ATBD_final_230712.pdf.

Huffman, G. J. et al. (2020), "Integrated Multi-satellite Retrievals for the Global Precipitation Measurement (GPM) Mission (IMERG)", in: *Satellite Precipitation Measurement: Volume 1*, ed. by V. Levizzani, C. Kidd, D. B. Kirschbaum, C. D. Kummerow, K. Nakamura, and F. J. Turk, Cham: Springer International Publishing, pp. 343–353, https://doi.org/10.1007/978-3-030-24568-9_19.

IPCC (2023), "Africa", in: *Climate Change 2022 – Impacts, Adaptation and Vulnerability: Working Group II Contribution to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change*, Cambridge: Cambridge University Press, pp. 1285–1456, https://doi.org/10.1017/9781009325844.011.

Judt, F., 2020: Atmospheric Predictability of the Tropics, Middle Latitudes, and Polar Regions Explored through Global Storm-Resolving Simulations. *Journal of the Atmospheric Sciences*, **77** (**1**), 257–276, https://doi.org/10.1175/JAS-D-19-0116.1.

Kalnay, E. (7, 2002), Atmospheric Modeling, Data Assimilation and Predictability, Cambridge University Press, https://doi.org/10.1017/CBO9780511802270.

Keane, R. J., Parker, D. J., Dunn-Sigouin, E., Kolstad, E. W., and Marsham, J. H., 2025: Mid-Latitude Versus Tropical Scales of Predictability and Their Implications for Forecasting. *Meteorological Applications*, **32** (**4**), e70055, https://doi.org/10.1002/met.70055.

Kiladis, G. N., Wheeler, M. C., Haertel, P. T., Straub, K. H., and Roundy, P. E., 2009: Convectively coupled equatorial waves. *Reviews of Geophysics*, **47** (**2**), https://doi.org/10.1029/2008RG000266.

Knippertz, P. et al., 2017: A meteorological and chemical overview of the DACCIWA field campaign in West Africa in June–July 2016. *Atmospheric Chemistry and Physics*, **17** (**17**), 10893–10918, https://doi.org/10.5194/acp-17-10893-2017.

Knippertz, P. et al., 2022: The intricacies of identifying equatorial waves. *Quarterly Journal of the Royal Meteorological Society*, **148** (**747**), 2814–2852, https://doi.org/10.1002/qj.4338.

Lafore, J. P. et al. (2017), "Deep Convection", in: *Meteorology of Tropical West Africa*, John Wiley & Sons, Ltd, pp. 90–129, https://doi.org/10.1002/9781118391297.ch3, (visited on 10/29/2025).

Lamptey, B. et al., 2024: Challenges and ways forward for sustainable weather and climate services in Africa. *Nature Communications*, **15** (**1**), 2664, https://doi.org/10.1038/s41467-024-46742-6.

Lavers, D. A., Simmons, A., Vamborg, F., and Rodwell, M. J., 2022: An evaluation of ERA5 precipitation for climate monitoring. *Quarterly Journal of the Royal Meteorological Society*, **148** (**748**), 3152–3165, https://doi.org/10.1002/qj.4351.

Light, C. X. et al., 2022: Effects of grid spacing on high-frequency precipitation variance in coupled high-resolution global ocean–atmosphere models. *Climate Dynamics*, **59** (**9**), 2887–2913, https://doi.org/10.1007/s00382-022-06257-6.

Lorenz, E. N., 1969: The predictability of a flow which possesses many scales of motion. *Tellus*, **21** (**3**), 289–307, https://doi.org/10.1111/j.2153-3490.1969.tb00444.x.

Maranan, M., Fink, A. H., Knippertz, P., Amekudzi, L. K., Atiah, W. A., and Stengel, M., 2020: A Process-Based Validation of GPM IMERG and Its Sources Using a Mesoscale Rain Gauge Network in the West African Forest Zone. *Journal of Hydrometeorology*, **21** (**4**), 729–749, https://doi.org/10.1175/JHM-D-19-0257.1.

Masutani, M. et al., 2010: Observing system simulation experiments at the National Centers for Environmental Prediction. *Journal of Geophysical Research: Atmospheres*, **115** (**D7**), https://doi.org/10.1029/2009JD012528.

Pante, G. and Knippertz, P., 2019: Resolving Sahelian thunderstorms improves mid-latitude weather forecasts. *Nature Communications*, **10** (**1**), 3487, https://doi.org/10.1038/s41467-019-11081-4.

Parker, D. J. et al., 2022: The African SWIFT Project: Growing Science Capability to Bring about a Revolution in Weather Prediction. *Bulletin of the American Meteorological Society*, **103** (**2**), E349–E369, https://doi.org/10.1175/BAMS-D-20-0047.1.

Pathak, J. et al. (22, 2022), FourCastNet: A Global Data-driven High-resolution Weather Model Using Adaptive Fourier Neural Operators, https://doi.org/10.48550/arXiv.2202.11214, arXiv: 2202.11214 [physics], pre-published.

Potthast, R. (2019), Documentation of the Data Assimilation Coding Environment, URL: https://www.cosmo-model.org/content/support/software/dace.pdf.

Privé, N. C., Errico, R. M., and Tai, K.-S., 2013: Validation of the forecast skill of the Global Modeling and Assimilation Office Observing System Simulation Experiment. *Quarterly Journal of the Royal Meteorological Society*, **139** (**674**), 1354–1363, https://doi.org/10.1002/qj.2029.

Privé, N. C., McGrath-Spangler, E. L., Carvalho, D., Karpowicz, B. M., and Moradi, I., 2023: Robustness of Observing System Simulation Experiments. *Tellus*, **75** (**1**), https://doi.org/10.16993/tellusa.3254.

Rasheeda Satheesh, A., Knippertz, P., and Fink, A. H., 2025: Machine Learning Models for Daily Rainfall Forecasting in Northern Tropical Africa Using Tropical Wave Predictors. *Weather and Forecasting*, **40** (**10**), 1895–1916, https://doi.org/10.1175/WAF-D-24-0192.1.

Redelsperger, J.-L., Thorncroft, C. D., Diedhiou, A., Lebel, T., Parker, D. J., and Polcher, J., 2006: African Monsoon Multidisciplinary Analysis: An International Research Project and Field Campaign. *Bulletin of the American Meteorological Society*, **87** (**12**), 1739–1746, https://doi.org/10.1175/BAMS-87-12-1739.

Reinert, D., Rieger, D., and Prill, F., 2024: ICON Tutorial 2024: Working with the ICON Model, https://doi.org/10.5676/DWD_PUB/NWV/ICON_TUTORIAL2024.

Roberts, A. J., Marsham, J. H., and Knippertz, P., 2015: Disagreements in Low-Level Moisture between (Re)Analyses over Summertime West Africa. *Monthly Weather Review*, **143** (**4**), 1193–1211, https://doi.org/10.1175/MWR-D-14-00218.1.

Roberts, N. M. and Lean, H. W., 2008: Scale-Selective Verification of Rainfall Accumulations from High-Resolution Forecasts of Convective Events. *Monthly Weather Review*, **136** (**1**), 78–97, https://doi.org/10.1175/2007MWR2123.1.

Rodwell, M. J., Richardson, D. S., Hewson, T. D., and Haiden, T., 2010: A new equitable score suitable for verifying precipitation in numerical weather prediction. *Quarterly Journal of the Royal Meteorological Society*, **136** (**650**), 1344–1363, https://doi.org/10.1002/qj.656.

Rojas, Y., Minder, J. R., Campbell, L. S., Massmann, A., and Garreaud, R., 2021: Assessment of GPM IMERG satellite precipitation estimation and its dependence on microphysical rain regimes over the mountains of south-central Chile. *Atmospheric Research*, **253**, 105454, https://doi.org/10.1016/j.atmosres.2021.105454.

Ruckstuhl, Y. and Janjić, T., 2020: Combined State-Parameter Estimation with the LETKF for Convective-Scale Weather Forecasting. *Monthly Weather Review*, **148** (**4**), 1607–1628, https://doi.org/10.1175/MWR-D-19-0233.1.

Schraff, C. et al., 2016: Kilometre-scale ensemble data assimilation for the COSMO model (KENDA). *Quarterly Journal of the Royal Meteorological Society*, **142** (**696**), 1453–1472, https://doi.org/10.1002/qj.2748.

Schulzweida, U. (2023), CDO User Guide, version 2.4.0, URL: https://code.mpimet.mpg.de/projects/cdo.

Sobel, A. H. (2012), Tropical Weather, URL: https://www.nature.com/scitable/knowledge/library/tropical-weather-84224797/.

Soci, C. et al., 2024: The ERA5 global reanalysis from 1940 to 2022. *Quarterly Journal of the Royal Meteorological Society*, **150** (**764**), 4014–4048, https://doi.org/10.1002/qj.4803.

SOFF (2025), The Systematic Observations Financing Facility, SOFF, Systematic Observations Financing Facility, URL: https://un-soff.org/.

TAHMO (2025), TAHMO, TAHMO, URL: https://tahmo.org/.

UN (2024), World Population Prospects, URL: https://population.un.org/wpp/.

– (2025), Early Warnings for All, United Nations, URL: https://www.un.org/en/climatechange/early-warnings-for-all.

Van der Linden, R., Knippertz, P., Fink, A. H., Ingleby, B., Maranan, M., and Benedetti, A., 2020: The influence of DACCIWA radiosonde data on the quality of ECMWF analyses and forecasts over southern West Africa. *Quarterly Journal of the Royal Meteorological Society*, **146** (**729**), 1719–1739, https://doi.org/10.1002/qj.3763.

Vogel, P., Knippertz, P., Fink, A. H., Schlueter, A., and Gneiting, T., 2020: Skill of Global Raw and Postprocessed Ensemble Predictions of Rainfall in the Tropics. *Weather and Forecasting*, **35** (**6**), 2367–2385, https://doi.org/10.1175/WAF-D-20-0082.1.

Walz, E.-M., Henzi, A., Ziegel, J., and Gneiting, T., 2024: Easy Uncertainty Quantification (EasyUQ): Generating Predictive Distributions from Single-Valued Model Output. *SIAM Review*, **66** (**1**), 91–122, https://doi.org/10.1137/22M1541915.

Walz, E.-M., Knippertz, P., Fink, A. H., Köhler, G., and Gneiting, T., 2024: Physics-Based vs Data-Driven 24-Hour Probabilistic Forecasts of Precipitation for Northern Tropical Africa.

Monthly Weather Review, **152** (**9**), 2011–2031, https://doi.org/10.1175/MWR-D-24-0005.1.

Warner, T. T. (2010a), "Ensemble Methods", in: *Numerical Weather and Climate Prediction*, Cambridge: Cambridge University Press, pp. 252–283.

– (2010b), "Model Initialization", in: *Numerical Weather and Climate Prediction*, Cambridge: Cambridge University Press, pp. 198–251.

– (2010c), "Numerical Solutions to the Equations", in: *Numerical Weather and Climate Prediction*, Cambridge: Cambridge University Press, pp. 17–118.

– (2010d), "Physical-Process Parameterizations", in: *Numerical Weather and Climate Prediction*, Cambridge: Cambridge University Press, pp. 119–170.

– (2010e), "The Governing Systems of Equations", in: *Numerical Weather and Climate Prediction*, Cambridge: Cambridge University Press, pp. 6–16.

Wilhelm, J. et al., 2025: TEEMLEAP—A New Testbed for Exploring Machine Learning in Atmospheric Prediction for Research and Education. *Journal of Advances in Modeling Earth Systems*, **17** (**7**), e2024MS004881, https://doi.org/10.1029/2024MS004881.

Wilks, D. (2011), "Chapter 8 - Forecast Verification", in: *Statistical Methods in the Atmospheric Sciences*, ed. by D. S. Wilks, International Geophysics, Academic Press, pp. 301–394, https://doi.org/10.1016/B978-0-12-385022-5.00008-7.

Wu, G., Qin, S., Mao, Y., Ma, Z., and Shi, C., 2022: Validation of Precipitation Events in ERA5 to Gauge Observations during Warm Seasons over Eastern China. *Journal of Hydrometeorology*, **23** (**5**), 807–822, https://doi.org/10.1175/JHM-D-21-0195.1.

Zängl, G., Reinert, D., Rípodas, P., and Baldauf, M., 2015: The ICON (ICOsahedral Non-hydrostatic) modelling framework of DWD and MPI-M: Description of the non-hydrostatic dynamical core. *Quarterly Journal of the Royal Meteorological Society*, **141** (**687**), 563–579, https://doi.org/10.1002/qj.2378.

Zeng, Y., Janjić, T., de Lozar, A., Welzbacher, C. A., Blahak, U., and Seifert, A., 2021: Assimilating radar radial wind and reflectivity data in an idealized setup of the COSMO-KENDA system. *Atmospheric Research*, **249**, 105282, https://doi.org/10.1016/j.atmosres.2020.105282.

Zhang, J. and Zuidema, P., 2019: The diurnal cycle of the smoky marine boundary layer observed during August in the remote southeast Atlantic. *Atmospheric Chemistry and Physics*, **19** (**23**), 14493–14516, https://doi.org/10.5194/acp-19-14493-2019.

A Figures

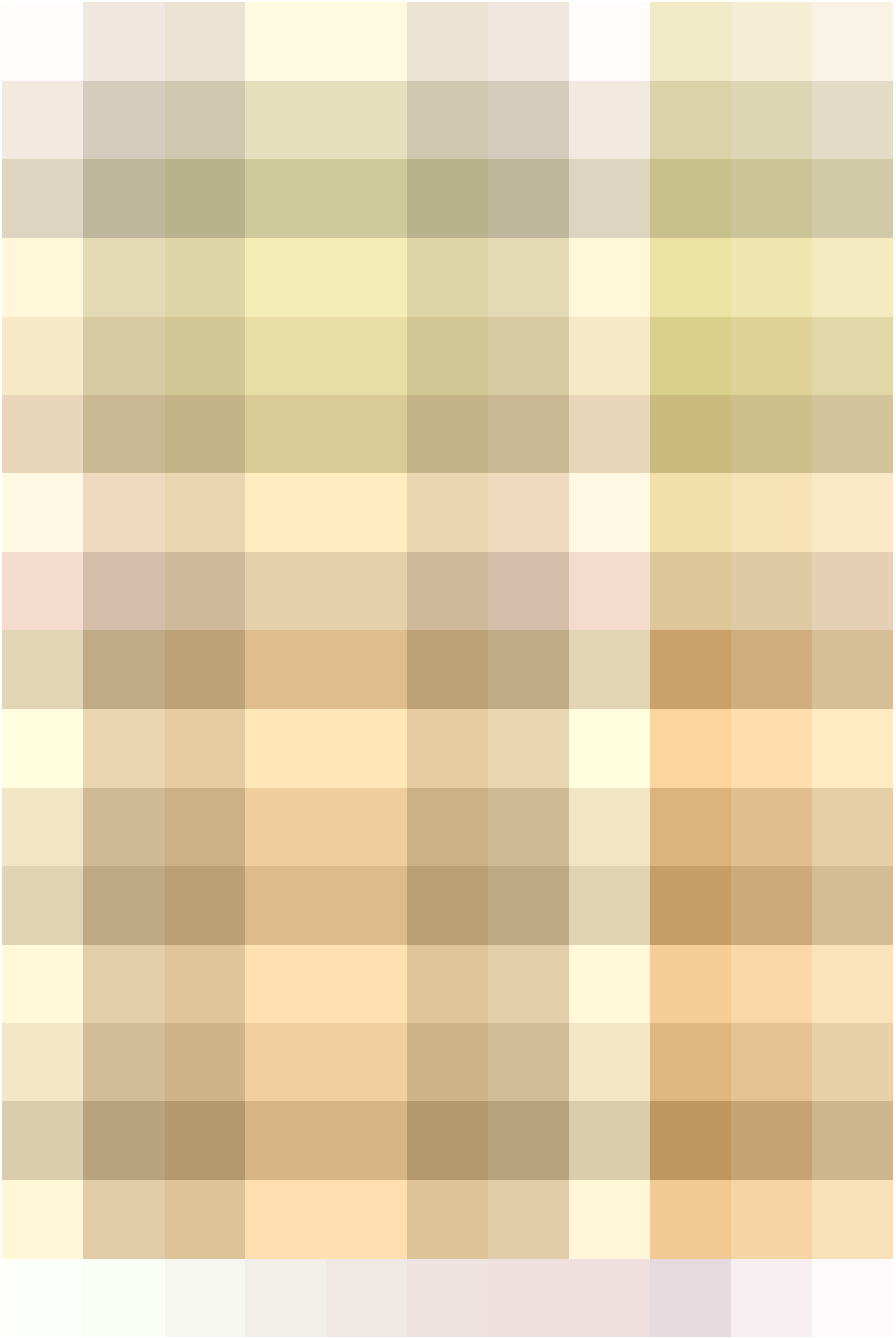


Figure A.1: Spatial RMSE maps of TCWV over Europe for experiments Global500, Global1000 and Global2000. Lead times are (a–c) 0 h, (d–f) 24 h, (g–i) 48 h, (j–l) 72 h and (m–o) 96 h. Area-weighted mean RMSE values are shown in the top-right corner of each panel.

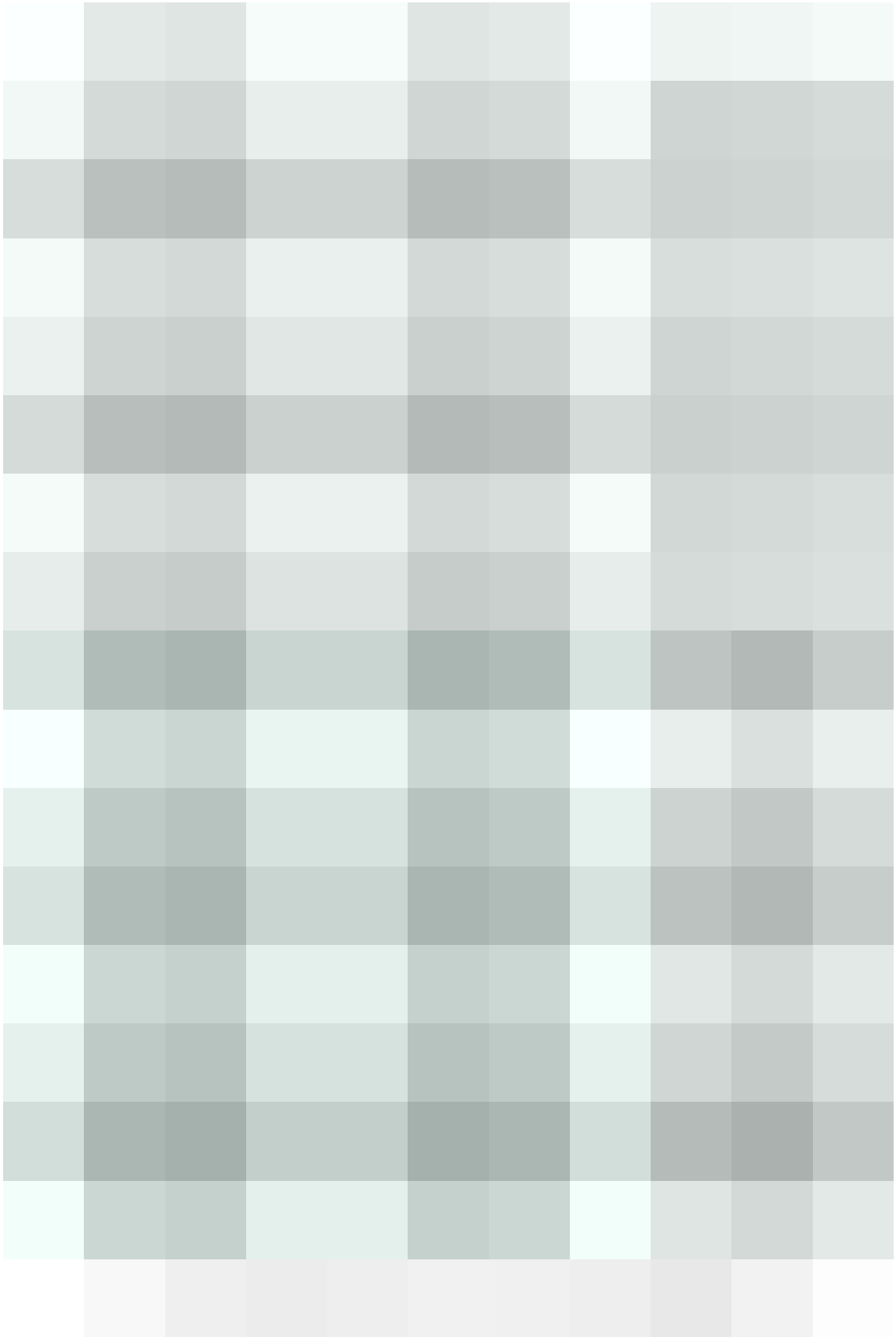


Figure A.2: Spatial ME maps of TCWV over Europe for experiments Global500, Global1000 and Global2000. Lead times are (a–c) 0 h, (d–f) 24 h, (g–i) 48 h, (j–l) 72 h and (m–o) 96 h. Area-weighted mean ME values are shown in the top-right corner of each panel.

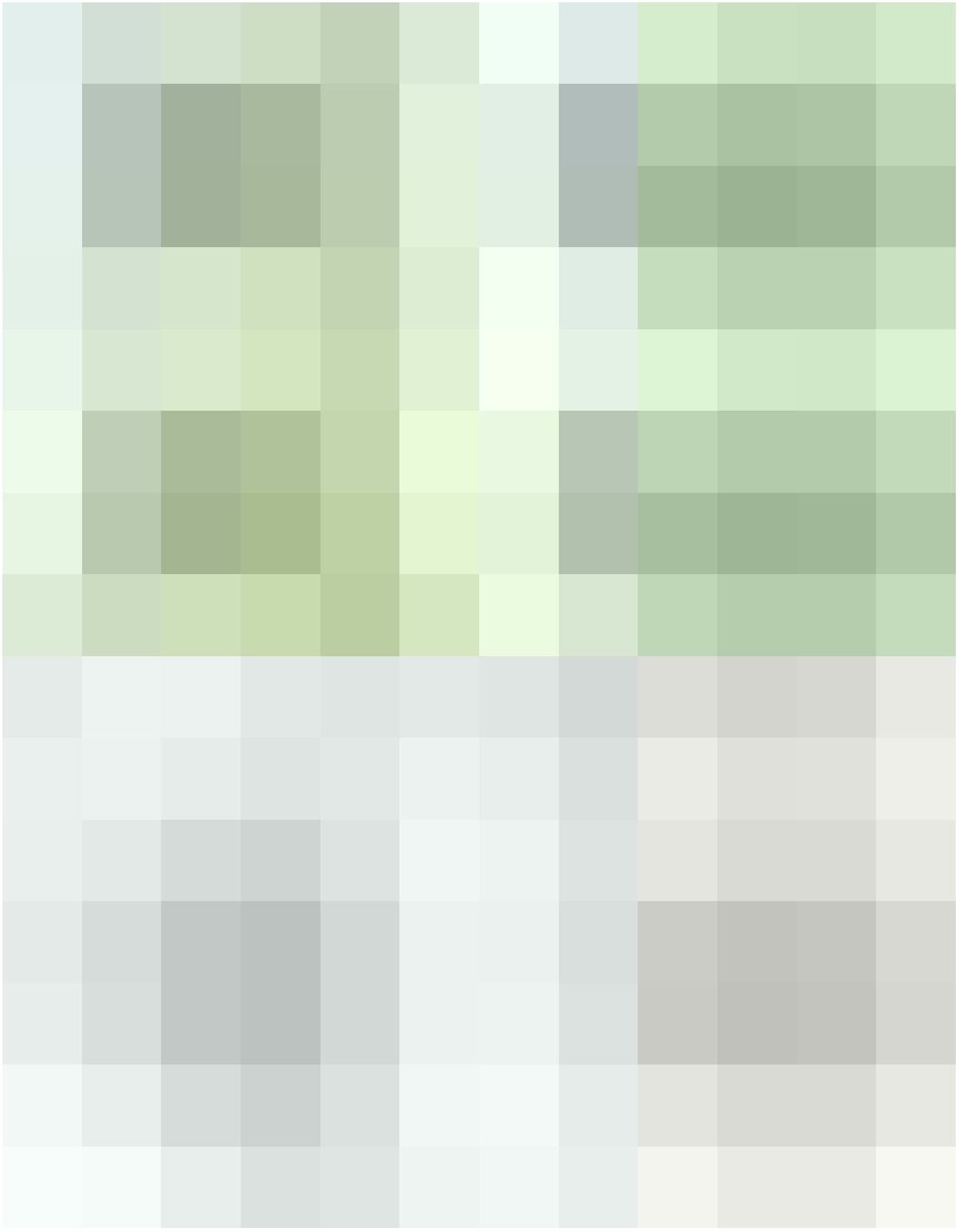


Figure A.3: Accumulated monthly precipitation over Europe for September. Panels show (a) IMERG climatology (2000–2021), (b) IMERG 2022, (c) ERA5 climatology (2000–2021), (d) ERA5 2022, (e) differences between IMERG and ERA5 climatology and (f) differences between IMERG and ERA5 in 2022. Area-weighted mean values are shown in the upper-right corner of each panel.

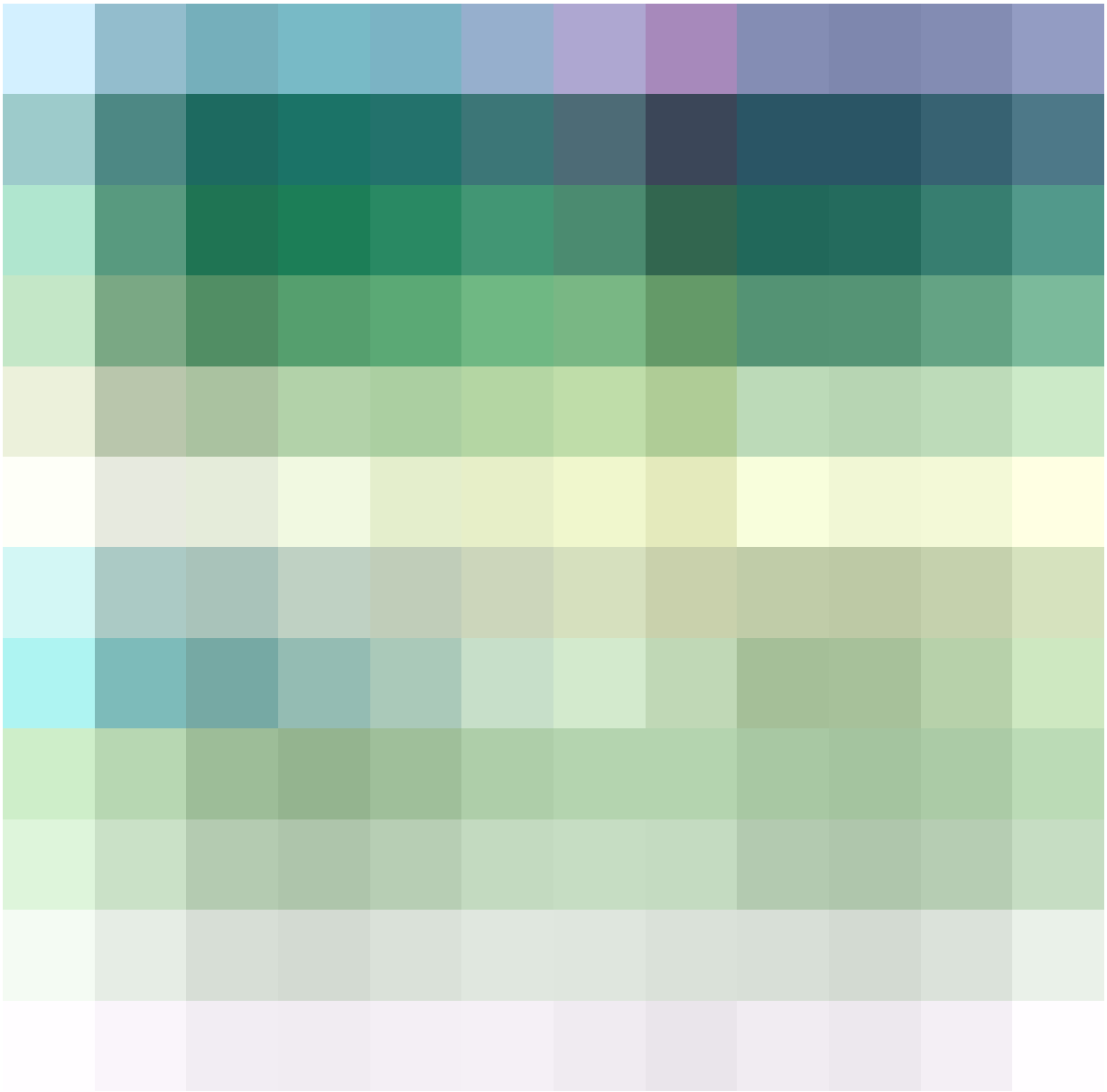


Figure A.4: Climatological classification of dry probability (p_1) and the threshold between light (p_2) and heavy rain (p_3) used for the SEEPS score over Europe. The upper row shows p_1 for (a) IMERG and (b) ERA5. The bottom row presents the thresholds between p_2 and p_3 for (c) IMERG and (d) ERA5. Blank areas are excluded from the SEEPS evaluation, as the frequency of rainy days is too low in these regions.

Figure A.5: Area- and time-averaged SEEPS score for the control experiment (Global500) evaluated against IMERG (blue) and ERA5 (red) over Europe, shown as a function of forecast lead time. The top-left panel shows the SEEPS score, while the top-right panel displays the frequency of forecasts and observations falling into the same category. The remaining six panels show the frequencies of different combinations of forecast and observed categories, where 1 represents dry, 2 light rain and 3 heavy rain. All frequencies are expressed in percent.

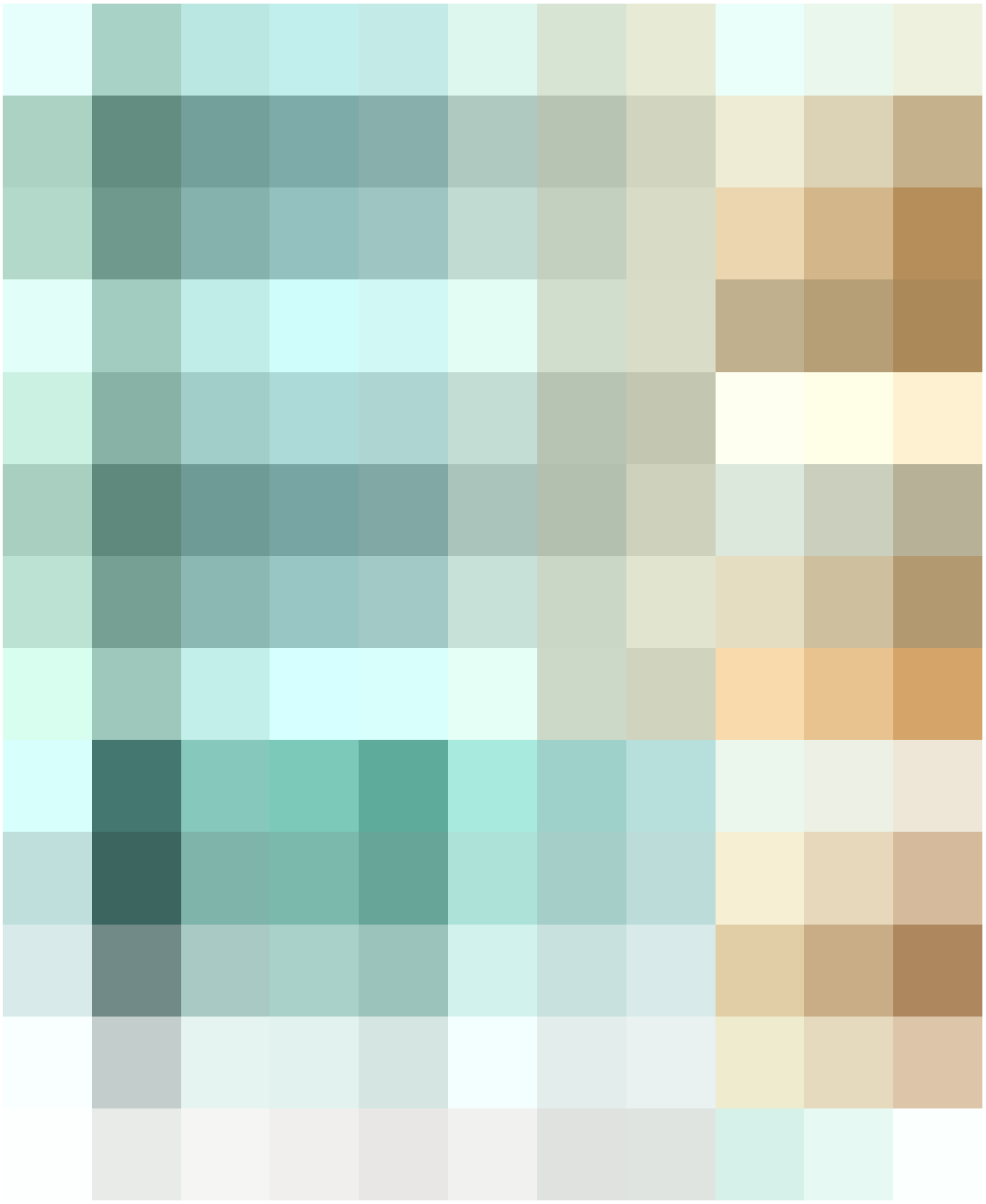


Figure A.6: FSS over Europe as a function of lead time and spatial scale for thresholds of 1 mm, 5 mm, 10 mm and 20 mm. Forecasts are evaluated against IMERG to compute the FSS. Results are shown for Global500 (top row), Global1000 (middle row) and Global2000 (bottom row).

Acknowledgments

I would like to express my sincere gratitude to Peter Knippertz for giving me the opportunity to conduct my master's thesis in his "Atmospheric Dynamics" group on an interesting and challenging topic. The discussions about next steps and results were always insightful and inspiring.

Special thanks go to Jannik Wilhelm, who introduced me to the world of the TEEMLEAP testbed. Without his support, the successful completion of the experiments within the testbed would not have been possible. I greatly appreciated the time and effort he dedicated to helping me, even after leaving KIT — often in his free time.

Many thanks go to Michael Krayer for his quick and competent support following the HoreKa update. The update required several software adaptations within the TEEMLEAP testbed as well as a new compilation of the ICON model. His assistance was essential — without it, I would not have been able to complete my own runs successfully.

I gratefully acknowledge the computing time provided on the high-performance computer HoreKa by the National High-Performance Computing Center at KIT (NHR@KIT). This center is jointly supported by the Federal Ministry of Education and Research and the Ministry of Science, Research and the Arts of Baden-Württemberg, as part of the National High-Performance Computing (NHR) joint funding program (`https://www.nhr-verein.de/en/our-partners`). HoreKa is partly funded by the German Research Foundation (DFG).

Various software tools supported the preparation of this thesis. Data processing and analysis were performed using Python and the Climate Data Operators (CDO). For improving the clarity and correctness of the written language, the free version of ChatGPT (`https://chat.openai.com`) was employed.

Erklärung

Ich versichere wahrheitsgemäß, die Arbeit selbständig verfasst, alle benutzten Hilfsmittel vollständig und genau angegeben und alles kenntlich gemacht zu haben, was aus Arbeiten anderer unverändert oder mit Abänderungen entnommen wurde sowie die Satzung des KIT zur Sicherung guter wissenschaftlicher Praxis in der jeweils gültigen Fassung beachtet zu haben.

Karlsruhe, den 11.12.2025

(Cédric Froidevaux)