Detection and Prediction of the Rainy Season Onset in West Africa

Lorenzo Occelli

Detection and Prediction of the Rainy Season Onset in West Africa

Master's Thesis

by

Lorenzo Occelli

MSc Climate Physics

Utrecht University

Study number: 0661678

July 3, 2023

Supervisor: Dr. Michiel L.J. Baatsen
Second Supervisor: Prof. Michiel R. van den Broeke
External Supervisor: MSc. Bob Ammerlaan

Universiteit Utrecht

Weather Impact

Preface

This thesis serves as conclusion of my MSc in Climate Physics and is the culmination of a seven-month internship at *Weather Impact*. In this period I explored the characteristics of the monsoonal climate in West Africa, focusing on the onset of the rainy season and its prediction. The convergence of academic research and operational weather forecasting made this project extremely dynamic, helping me to gain a more holistic perspective on climate science and its implications for agriculture and society. The intership was also a unique opportunity to engage with state-of-art weather forecasting in an non-academic environment.

I am extremely grateful of the experience and personal growth that resulted from this thesis, as well as the opportunities it opened up for me. My first thank you is for Bob Ammerlaan and Michiel Baatsen, who supervised me at *Weather Impact* and at *Utrecht University* with the utmost dedication and patience. I also extend my thanks to Stefan Ligtenberg for giving me the opportunity to work at *Weather Impact* and for sharing his vast expertise in the field. My daily experience in the office would not have been as enjoyable as it was without its welcoming team: thank you to Edith, Gerrit, Glenn, Janina, Paula and Tamara for the nice chats and the coffee breaks. A special thanks to Thirza and Joep, fellow interns, for sharing our experiences together. I must also acknowledge my classmates from the Master's program, who lightened the burden of long winter workdays with their genuine friendship.

Among the people that in some way contributed to this project, I want to express my gratitude for Michiel van den Broeke, Nick van de Giesen and Francesco Guardamagna. A special mention goes to my sister, Martina, who is always interested in my professional development and generously shares her academic experience to provide valuable advice.
Lastly, I want to express my appreciation to my parents. They always contribute the most, being supportive and enabling me to pursue my interests even if they take me far from home.

Lorenzo Occelli

Amersfoort, July 3, 2023

Abstract

The onset of the rainy season is crucial for rain-fed farming in tropical countries. The timing of the onset can drastically influence the performance of the growing season, which subsequently impacts the food security of rural communities. Despite the high societal importance, the availability of operational forecast of the onset is limited, especially across the African continent. While there is little ambiguity about the overall behaviour of the West African Monsoon (WAM), there are several man-made definitions of its onset. Those definitions can address the onset both regionally and locally, the latter ones being the most commonly used.

In this study we tackle the operational prediction of the WAM onset, focusing on Ghana and neighbouring countries. The regional approach is the main focus of both the climatological research and the development of the forecasting algorithm. However, the local perspective has been analysed too, representing the common practise for onset's forecasting. Firstly, the correlation between the onset and atmospheric variables is investigated in the period 1981-2021, using satellite-based rainfall records (CHIRPS dataset) and ERA5 re-analysed wind fields at different pressure levels. Afterwards, the operational prediction of the onset is produced by post-processing operational weather forecasts, issued by the European Center for Medium-range Weather Forecasts (ECMWF).

It is found that ECMWF rainfall forecasts (both medium-range and sub-seasonal) are the necessary ingredients of any prediction of the onset. Other atmospheric variables, such as wind fields at 925, 850, and 200 hPa, revealed to bring little to no advantage for the operational forecast of the onset, despite showing a clear correlation with the WAM in the climate. The best forecast performances are obtained with a threshold-based algorithm, which detects the onset imposing conditions on the amount and the temporal distribution of rainfall. Both regional and local onset's forecast display promising performances in specific areas of the analysed region. Those areas are characterised by high spatial coherence of the rainfall pattern and can also be influenced by the adopted onset's definition.

We conclude that the operational forecast of the onset is feasible with both the regional and the local approach, but only for a portion (about 47%) of the analysed domain. However, predicting the onset is far from straightforward and the obtained forecast are dependent on several semi-arbitrary choices, first and foremost on the chosen onset's definition. Due to this dependency, it is of utmost importance to include the end-user needs in the development of any rainy season onset's forecast.

Contents

Chapter 1

Introduction and Motivation

In tropical countries a number of human activities are tightly related to the rainy season and its timing. These activities, ranging from food and energy production to vector-borne diseases, significantly impact both population's life quality and national economies. The Onset date of the Rainy Season (**ORS**) is therefore a key information for agriculture, hydro-power production and healthcare facilities. However, the ORS is a difficult variable to predict in operational weather forecasting, not least due to the arbitrariness around the onset's definition itself. As a result, many African regions lack of reliable weather information concerning the rainy season onset to plan essential activities, such as the sowing season.

This study aims to explore different methods to forecast the ORS in a sub-region of West Africa, covering Ghana, Togo, large parts of Benin, Ivory Coast, and Burkina Faso. Unlike most of the available literature, this thesis gives the highest priority to the simulation of operational settings, in order to gain a realistic evaluation of the forecast's performance. Moreover, we adopted in most of the analyses a regional approach, which has been less investigated (especially for operational applications) than the more common local approach.

This chapter introduces the rain season's onset and its societal implications (Section 1.1), highlights the scope and research questions of this study (Section 1.2), discusses and reviews the possible definitions of onset's dates (Section 1.3), and gives an overview of the structure of the thesis (Section 1.4).

1.1 The Onset of the Rainy Season

According to the International Labour Organization (ILO), agriculture employs 52% of workforce in Sub-Saharan Africa (May 2023) [25]. Moreover, the Africa Human Development Report 2012 claims that, across Africa, up to 75% of cereals comes from domestic agriculture [42], suggesting that subsistence farming is wide-spread in the Continent. Additionally, You et al. [44] reports that only 6% of the total agricultural area of whole Africa is irrigated land, two-thirds of which are concentrated in Egypt, Madagascar, Morocco, South Africa, and Sudan. In Ghana, the Food and Agriculture Organization (FAO) estimates that only 0.3% of farming land is irrigated [18]. These factors lead to conclude that the population of West African countries is heavily dependent on rain-fed agriculture, which is a key issue for food security in the region [4].

In the context of rain-fed farming, the ORS is a crucial time of the year that greatly affects the yield of the following growing season. Planting too early may expose to dry spells, which could led to severe harvest failure in the worst scenario; on the other hand, a delayed sowing results in a reduced amount of available rainfall during the growing season, affecting negatively the crop yield.

The influence of ORS (and its information availability) on agricultural planning is supported by different studies. Baffour-Ata et al. [3] analysed the crop yields of groundnut, sorghum, millet, maize and rice in north Ghana, concluding that the number of dry days and the ORS explain respectively 32% and 30% of yields variations, second only to temperature (43%). Focusing on the semi-arid Southern Kenya, Hansen et al. [24] found that seasonal rainfall forecast (generated by a general circulation model (GCM) based on sea surface temperature) could potentially increase from 9% to 24% average gross margin for maize. In addition, heat-waves and heavy precipitation events are projected to increase in frequency and intensity across the entire African continent (IPCC [37]). Already in present days, climate change is putting the African farming under stress, lowering the crop productivity and enhancing risks related to extreme events [12] [16]. A focus study on

different regions of Ghana has shown that monthly minimum and maximum temperature oscillations accounts for 44% and 57% of variations in maize harvest [17], with yield decreasing for both minimum and maximum temperature's increase.

As many of the above-mentioned conditions can severely impact small stake-holder's food security, weather services constitute the backbone to increase the resilience of African farming. They could also improve the effectiveness of humanitarian aids, shifting the resources from reactive measures, in response to ongoing food crises, to proactive strategies, preventing such crises far ahead [10].

While this research is focused on the effect of the ORS on agriculture, it is important to mention that the beginning of the wet season has profound implications in many other contexts. Public health in West Africa is a prominent example: the start of monsoonal rainfalls coincides with the end of bacterial meningitis outbreaks [31] [41], but it also increases vector-borne diseases such as dengue fever and malaria [39]. Predictive information about the ORS could therefore help national health organisations in the planning of medical countermeasures.

1.2 Scope of the Thesis and Research Questions

This thesis project was promoted, funded, and hosted by *Weather Impact*, based in Amersfoort (NL). *Weather Impact* is a private company specialised in providing tailored weather and climate information to agribusinesses worldwide, with a specific focus on rain-fed farming in Africa and South-east Asia. With their different requirements and perspectives, the convergence of academic research and operational weather forecasting conferred to this project challenging but exciting settings. In line with these premises, the aim of the research and final goal of the intership has been the evaluation of possible operative algorithms able to predict the ORS for a given location. A rigorous scientific approach has been employed during the design and implementation of this algorithm, leading to an in-depth climatological analysis of the rainy season onset in West Africa.

Research Questions

The research questions leading the project are:

1. Can large-scale atmospheric fields serve as predictors for the start of the rainy season? (Addressed in Chapter 3)
2. Is it feasible to forecast the ORS regionally, and how this compares to local predictions? (Addressed in Chapter 4)
3. What are the possibilities of applying Supervised Learning techniques to the ORS prediction? (Addressed in Chapter 5)

Related to research question 1:

- Which atmospheric field shows distinguishable signals to suggest an imminent change to a wet regime? (Addressed in Section 3.3)
- Are those signals strong enough to justify their operational usage? (Addressed in Section 3.4)
- What can be derived by large-scale atmospheric patters about the characteristics of the rainy season and of its onset? (Addressed in Section 3.5)

1.3 Onset Dates Definitions

The aim of this section is to provide a general overview on the different ORS definitions, while it is scope of section 2.3 to describe the onset's detection techniques used in the research.

The onset of the rainy season is far from having a unique interpretation: a literature review by Fitzpatrick et al. [19] has identified at least 18 different definitions for the West African Monsoon (WAM) onset, which can be applied to every other region showing a tropical rainfall regime. Part of this huge variety is accountable to different scientific proxies which can identify the beginning of the rainy season, while part is due to

the societal importance of the rainfalls, which leads to the implementation of onset definitions tailored on agronomic needs, both location and crop-specific.

Two distinctions can be drawn to create order in this forest of definitions: a first line separates *local* from *regional* onset definitions, while a second divides the definitions based on rainfall thresholds from the ones based on a larger variety of atmospheric fields and indices. It is important to note that local onset definitions are almost exclusively based of specific rainfall thresholds at the given location.

1.3.1 Local Onset Definitions

Local onset definitions are based on rainfall data collected from single station's gauges or from satellites' products that provide information for every cell of a grid. Intrinsic to this method is therefore the ignorance of the onset in every neighbouring location. For agronomic scopes the definition of a location-specific onset is preferable and of most applicability to local stake-holders' needs [11]. On the other hand, local onsets show concurrently the highest inter-annual and spatial variability, which poses limits to their predictability. These onset dates are extremely sensitive to single convective events acting on the local scale, which are for their own nature unpredictable: an isolated event can cause the anticipation of the onset date by several weeks, leading to an erroneous onset's estimate.

Trying to limit the effect of isolated rainfall events, some methods employ a combination of different empirical indices, accounting for the total precipitation amount received by a specific location, its temporal distribution and the number of consecutive days without dry spells [8]. The testing and validation of such algorithms has shown that finding suitable thresholds is not a trivial task and that their performances are often not satisfying. Moreover, each of these thresholds is extremely costumed on the location and type of crop, creating a strong limitation to the up-scaling of the method.

Noteworthy is the approach introduced by Liebmann et al. [30] and later improved by Bombardi at al. [5]. This method implements a local ORS detection technique based on precipitation data which avoids the use of fixed rainfall threshold. The definition relies on cumulative precipitation anomalies, identifying the ORS at the time when the first derivative of the time-smoothed anomaly changes sign. More details about this method can be found in Section 2.2.1.

1.3.2 Regional Onset Definitions

The term "*regional onset definition*" can have different interpretations (particularly concerning the size of a *region*) and once again there is no consensus regarding a specific one. In the following research we consider as *regional* every ORS definition not relying on parameters obtained from a single ground-station or from a single cell of a gridded data.

Defining a regional onset entails capturing the start of the rainy season as a phenomenon transcending the characteristics of localized convective system, instead interpreting it as part of atmospheric events with meso to synoptic scale features. Several atmospheric and oceanographic indices can be uses to implement a regional ORS definition:

- Sultan and Janicot [40] looked at zonally-averaged precipitation data, defining the onset of WAM as the time when the 10-days moving average of precipitation at 5°N decreases simultaneously with an increase at 10°N and 15°N;
- Fontaine, Louvet, and Roucou [20] employed a time-filtered, latitudinal-averaged, outgoing long-wave radiation (OLR) observations, claiming the onset when a set of conditions (concerning location and movement of the OLR minimum) are fulfilled;
- Nguyen, Renwick, and McGregor [32] defined an atmospheric index based on mean sea level pressure and 850 hPa zonal wind, identifying the ORS when the index changes sign, from negative to positive.

Among the valuable advantages of using a regional onset definition there is its suitability to perform climatological research on the physical mechanisms leading to the start of the rainy season. This task would be extremely difficult using local onsets, too much influenced by local-scale phenomena.

Many studies (such as Fitzpatrick et al. [19], Gbangou et al. [22]) have investigated the relation between the regionally-defined onsets, which are theoretically more predictable, and the locally-defined ones, which are more useful for agronomic applications. The findings often challenge the possibility to establish a strong correlation between the two. Nevertheless, regional onsets provide a natural approach to detach the prediction of the ORS from precipitation forecast only, allowing the introduction of wind-based predictors. Therefore, this work focuses primarily on regional ORS, as it serves both the academic research and operative forecast goals outlined in Section 1.2.

1.4 Structure of this Study

The thesis is divided in two parts, each of them beginning with an introduction to the data used in the subsequent analysis. *Part I* contains the climatological research effort, while *Part II* concerns the exploration of different operative forecasting routines, all centred around the onset of the rainy season in Ghana and neighbouring countries. In *Part I* the reader will find how the ORS is detected in historical precipitation's data, using both local and regional onset's definitions (Chapter 2). Employing the obtained climatology of the ORS, *Chapter 3* explores the relationship between the start of the wet season and the large-scale atmospheric circulation.

Moving to Part II, we tackle the challenging task of forecasting the ORS in operational settings. A threshold-based forecasting routine is presented in Chapter 4, where it is tested on reproducing operational-like forecast for the ORS of six years (2016-2021); while the regional perspective is the main focus, a local alternative is also analysed for comparison. The final chapter (Chapter 5) describes an attempt to apply Supervised Learning, in specific the Random Forest model, to predict the ORS. The model is trained and tested on both historical data (1981-2021) and operational forecasts (2016-2021), highlighting the opportunities and limitations of this new approach. Eventually, a brief Conclusion section reviews the key findings of the thesis and produces a final assessment on the prediction of the rainy season onset.

Part I

Climatological Research

Prior to devoting ourself to the forecasting the ORS in the future, we commence with examining the past. First, through a detailed analysis of historical rainfall records, we aim to get a climatology of onset dates: this will be the stepping stone to investigate the relationship between the start of the rainy season and the large-scale atmospheric fields governing the tropical circulation. Such relationships can be a key information to predict the onset's dates and to study seasonal outlook and characteristics of different rainy seasons.

Data

The study of the rainy season onset and its relationships with the atmospheric circulation requires historical records of both precipitation and wind fields. We employed:

- CHIRPS daily precipitation data (0.25° resolution, 1981-2021) [21];
- ERA5 reanalysis hourly data on pressure levels (0.25° resolution, 1981-2021) [23].

In this study the climatology is everywhere intended as the 1981-2021 time window. CHIRPS and ERA5 data share the same resolution, but on different grids (which are shifted of 0.125°). The ERA5 data have been therefore re-gridded to match the CHIRPS grid.

CHIRPS (Climate Hazards group Infra-Red Precipitation with Station data) is developed by the US Geological Survey (USGS) and the Climate Hazards Group of the University of California (UCSB). The dataset contains rainfall estimates from rain gauge and satellite observations, post processed in order to remove systematic biases. In our analysis we treat CHIRPS estimates as the ground truth. This is justified by Atiah et al. [2], who compared several satellite-based precipitation products over Ghana with rain-gauge observations between 1998-2012, concluding that CHIRPS is among the best performing one. Dembélé and Zwart [14] came to a similar conclusion performing the same analysis over Burkina Faso, between 2001-2014. However, the studies refer mainly to monthly and seasonal rainfall data, recording poor performances for daily products (which are the ingredient of our forecast). Caution is therefore still needed when interpreting results obtained from satellite-based precipitation products.

ERA5 constitutes the fifth generation of atmospheric reanalysis produced by the Copernicus Climate Change Service (CCCS) at the European Centre for Medium-Range Weather Forecasts (ECMWF). It blends observations from satellites, weather stations, ships, buoys, and other sources with advanced weather models to create a consistent and high-resolution representation of the Earth's climate. From ERA5 reanalysis (hourly data) the following fields were analysed:

- total precipitation (large-scale + convective precipitation, 1981-2021);
- meridional wind speed at 200, 500, 850 and 925 hPa (1981-2021);
- zonal wind speed at 200, 500, 850 and 925 hPa (1981-2021);
- vertical wind speed at 200, 500 and 850 hPa (1981-2021);
- wind divergence at 200, 500 and 850 hPa (1981-2021).

The hourly wind speeds were post-processed into daily averages, while the hourly total precipitation was transformed into daily accumulated rainfall amount.

Chapter 2

Building a Climatology of the Rainy Season Onset

2.1 Climatology of Rainfalls

This section aims to provide the reader with basic information about the development of the rainy season in West Africa, also known as the West African Monsoon (WAM). The focus area of the research is Ghana and parts of the neighbouring countries, in particular Togo, Benin, Burkina Faso, and Ivory Coast. This climatological picture will help to interpret the coming results and to support some of the choices made by the author during the study.

The core of this research is analysing and predicting the inter-annual variability of the ORS. Perhaps, such point of view leads to depict the rainy season itself as a phenomenon highly variable and difficult to predict: while this is often true for its onset, the rainy season itself displays different characteristics. The West African Monsoon shows a remarkably stable yearly occurrence, making West Africa a suitable location for agricultural activities. It comes as no surprise that, around the Globe, every area with a monsoonal climate is highly populated and it is home of the majority of World's population living in tropical and equatorial regions.

The seasonal unfolding of the WAM involves the movement of the Inter-Tropical Front (ITF), which is the line of convergence between moist and cooler southerly winds with dry and warmer northerly Harmattan winds. The ITF migrates north-ward during the boreal spring and south-ward during the autumn, carrying a maximum of convective precipitation called "*Inter-Tropical Convergence Zone*" (ITCZ) or, more generically, "*Tropical Rain Belt*". Two important remarks must be added: firstly, the location of the ITF does not correspond with the maximum of precipitation, which is usually located around 400 km south [19]; secondly, the migration of the ITCZ is far from being steady and smooth, resembling instead an alternation of high and low convective activity phases [26].

Over the area described in Figure 2.1 (**a**), the rainy season begins in February or March with convective activity coming from the Atlantic Ocean and therefore impacting first the coastal areas. Subsequently, it proceeds inland, reaching by late April and May the north of Ghana. This region is located approximately on the northern edge of the ITCZ oscillation range, entailing that these latitudes (above 10°N) experience the wet season only once per year. Conversely, the center and the south of Ghana experience wet conditions twice per year, in spring and in autumn. This is shown by the climatological daily precipitation time series (Figure 2.1 **b** to **d**), where the bimodal distribution of `Ghana South` progressively merges into the single-peak distribution of `Ghana North`. In every region the boreal winter is the driest season. The ORS in spring is the most well defined across every sub-region, therefore it is the focus of this study.

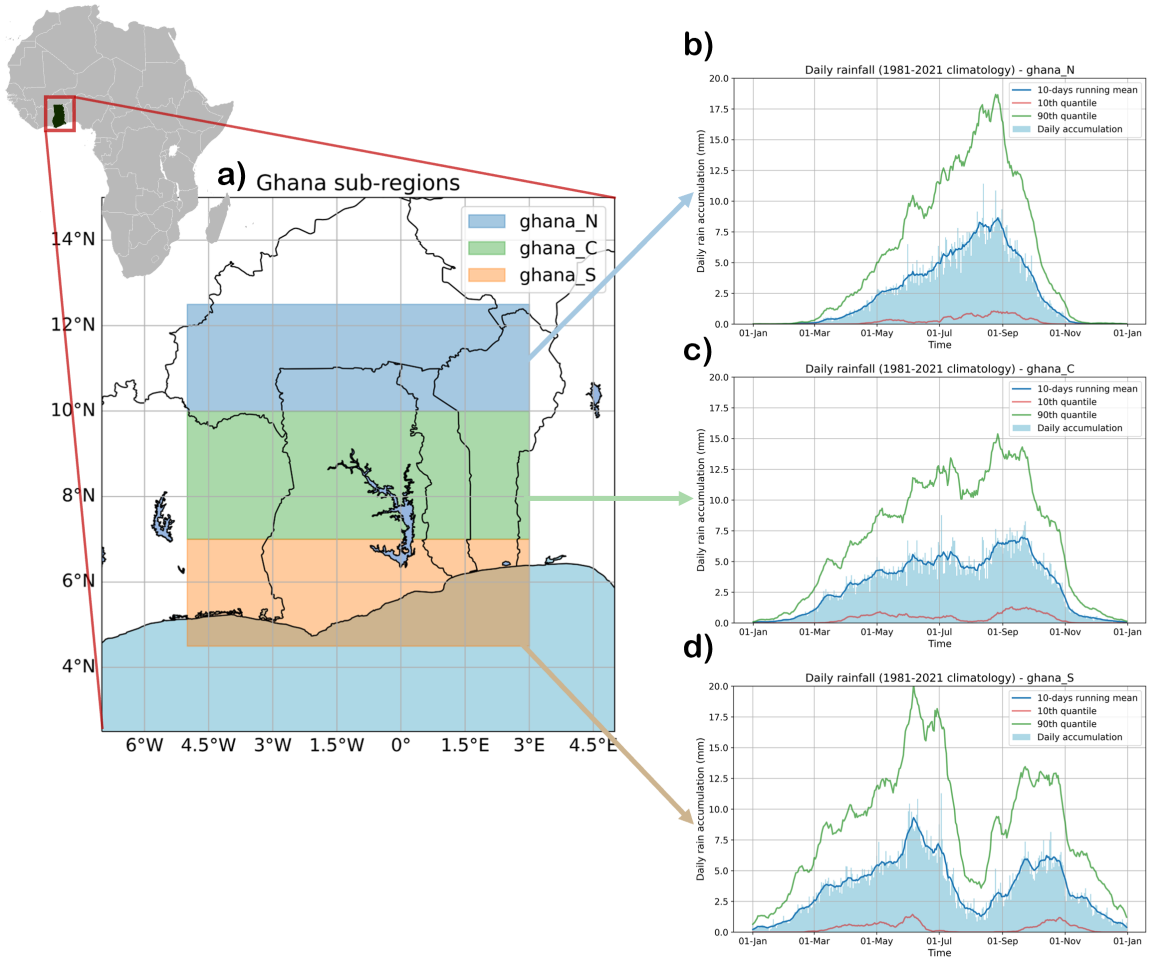

Figure 2.1: Climatology of daily rainfall accumulation over different regions of Ghana (CHIRPS 1981-2021). The plots on the right show both the daily values and the 10 days smoothed line. The rainfall accumulation is intended as the average of the area they refer to.

Figure 2.2 displays that the annual accumulation of precipitation is higher in the coastal areas and decreases moving northward. As the intensity of the area-averaged daily precipitation is similar across the region (Figure 2.1 **b** to **d**), the lower amount of rain in the north is not due to weaker precipitation intensity during the rainy season, but rather due to the different length of the rainy season itself. In addition, the surface vegetation and the orography can hugely influence the genesis of convective events and consequently the spatial distribution of accumulated precipitation. Such effect is recognisable in the east of the region, where the Togo-Akwapim mountains extend along the border between Ghana and Togo. According to Amekudzi et al. [1], the elevation-drive buoyancy instability of air masses rich in moisture leads to enhanced cloud formation and precipitations, affecting mostly the wind-ward side of the mountain range.

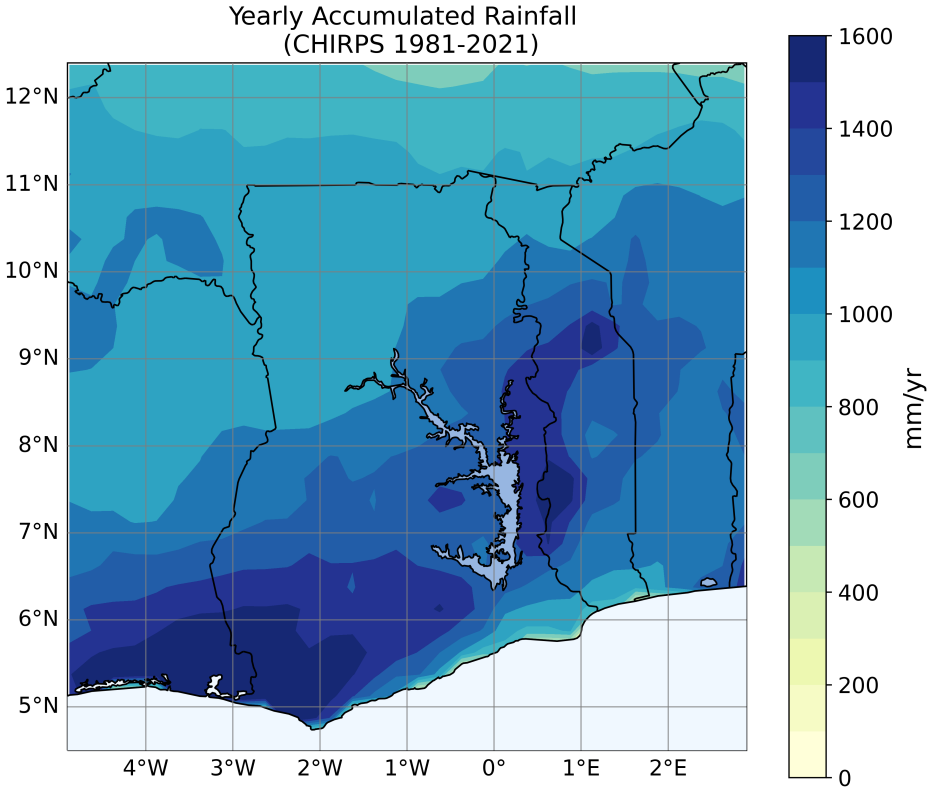

Figure 2.2: Climatology of yearly rain accumulation over Ghana (CHIRPS 1981-2021)

2.2 Detection of the Local Onset in Historical Data

As previously mentioned (Section 1.3.1), the local ORS is of primary importance for farming activities. While this study will mostly focus on the regional perspective, it is useful to be able to compare regional observations and predictions with their local counterpart. Consequently, the 1981-2021 climatology of local ORS has been produced for the entire area of interest (Figure 2.1 **a**).

2.2.1 Local Onset Detection

The local ORS was computed reproducing the method described by Bombardi et al. [5] with CHIRPS precipitation data (0.25°resolution). This method is not based on precipitation thresholds, therefore it is easily applicable in different regions and is robust against false starts. The beginning of the detection window is set on a day t_0, well within the dry season of the specific region (in case of Ghana is set on the 1st of January). From that date, the cumulative precipitation anomaly S is computed according to:

$$S = \sum_i (P_i - \bar{P}), \tag{2.1}$$

where i is the i^{th} day following the day t_0, P_i is the rainfall recorded on day i^{th} and $\bar{P}$ is the long-term mean annual precipitation. The S curve shows a decreasing trend until the dry season persists, while it deflects upwards at the beginning of the wet period (Figure 2.3). Consequently, the ORS is claimed when the first derivative of the S curve becomes positive. However, two precautions are taken to avoid too early ORS detections due to isolated rainfall events outside the actual rainy season:

- the S curve is smoothed with a 20 days running mean;
- the algorithm checks that the value of the first derivative is positive for 3 consecutive days before claiming the ORS.

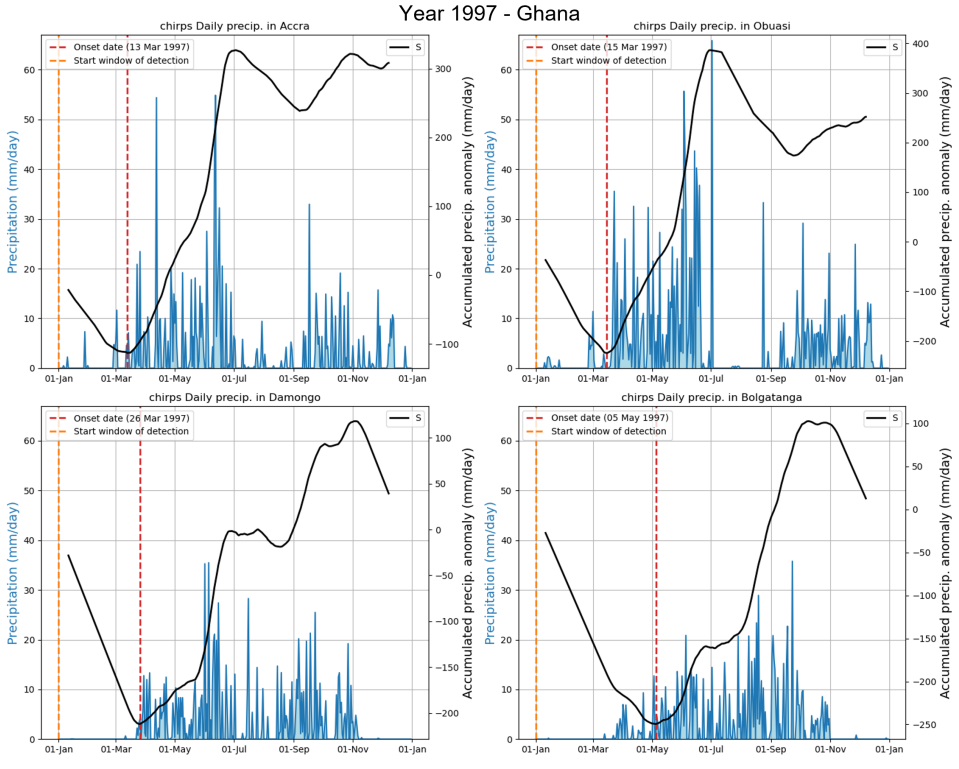

Figure 2.3: ORS detection in 1997 using the method introduced by Bombardi et al. [5]. Four different locations spread across Ghana are showcased following a south-to-north order (from top-left to bottom-right). Focusing of the rainfall time series, the reader can recognise the same transition from bi-annual to single rainy season regime observed in Figure 2.1.

2.2.2 Results: Local Rainy Season's Onset Climatology

The methodology described in the previous section has proven to be very successful when applied to our study area, allowing to identify the local ORS for each grid-cell and each year between 1981 and 2021, without exceptions.

The local ORS time series have been averaged to obtained the climatological onset date of each location, shown in Figure 2.4. Consistent with the literature ([19], [26]), the onset of the rainy season proceeds chronologically from the coast to the inner part of the region, along the south-north axis. The only exception is represented by the coastal area between 5°W and 2°W, which experiences the ORS later than the corresponding interior. The progression of the rainy season is not uniform across all latitudes, with a faster advancing pace observed in the center of the region respect to the north and south. In addition, it is likely that the Volta Lake (the body of water dominating the center of Ghana along the Greenwich's meridian) greatly affects the local ORS with small-scale processes.

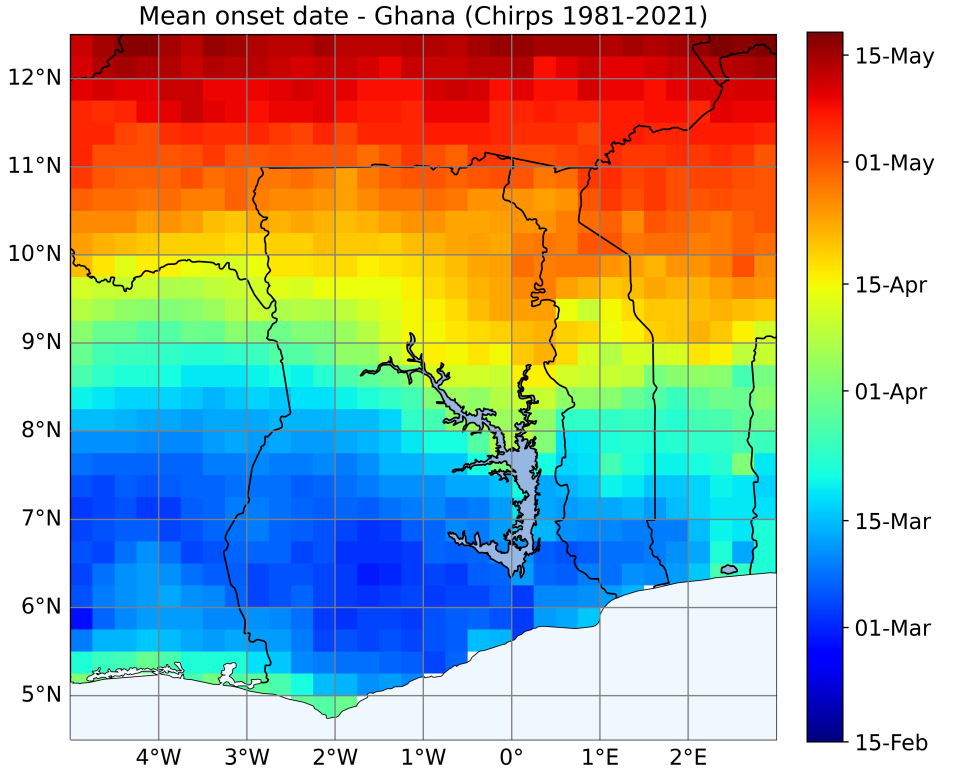

Figure 2.4: 1981-2021 climate of the ORS obtained with the Bombardi method [5] from CHIRPS daily precipitation records. The rainy season is found to start around early-March in the south of the region and to progressively shift north-ward.

2.3 Detection of the Regional Onset in Historical Data

We adopt here a regional approach to detected the ORS. This implies assigning a single onset's date to an entire region, covering several ground-stations/grid-cells. A more detailed explanation of this concept can be found in Section 1.3.2.

2.3.1 Sub-region's Definition

Throughout this study we call "*sub-region*" a spatial domain sharing the same onset date.
We want each sub-region to fulfil two requirements:

- within the sub-region, the precipitation spatial patter should be as homogeneous as possible;
- the sub-region must be large enough to minimize the influence of individual rainfall events and small-scale feedbacks.

The first condition aims to retain the same advantages of the local approach in terms of agronomic usefulness and suitability to satisfy the end-users' needs; the second requirement, on the opposite, is imposed to find correlations between the ORS and atmospheric variables, reducing ORS inter-annual variability and enhancing the signal-to-noise ratio. These conditions are clearly competing, meaning that choosing a sub-region entails finding a suitable balance among them.

The selection of a sub-region starts performing a Principal Component Analysis (PCA) ([27], [13]) of the entire historical rainfall record over the spatial domain of interest. This analysis returns: **1)** the principal component loadings or EOFs (Empirical Orthogonal Functions), which can be plotted over the spatial domain; **2)** the Principal Component (PC) time-series. Each EOF can be interpreted as a standing oscillation, whose time evolution is described by the time-series of the corresponding PC. A great number of EOFs can be found, but only the first two or three, sorted by the share of explained variance, capture the dynamical behaviour of the

system. The remaining are discarded as noise.

To determine the sub-regions we are mostly interested in the EOFs since they represent the co-variability of rainfall precipitation across the spatial domain. The goal is to isolate portions of land with similar loading's values and a high fraction of explained variance: this ensures that the observed oscillation is representing a physical feature of the rainfall pattern. To check this fact we can repeat the PCA limited to the chosen sub-region. Thanks to the homogeneous precipitation field, a very strong first EOF is found, accounting for around 50% of explained variance.

An example of this procedure is shown for Ghana and parts of neighbouring countries: the PCA returns a main mode of oscillation, already quite homogeneous throughout all the domain, which represents the yearly alternation of the wet and dry regimes (Figure 2.5); the second EOF (Figure 2.6) shows another mode, which display a clear meridional structure and it represents the north-ward progression of the rainy season over the region, from the coast to the inner part of the country. We follow this second pattern to define the sub-regions presented in Figure 2.7. They will be used in the following analysis as spatial domain to compute the regional ORS.

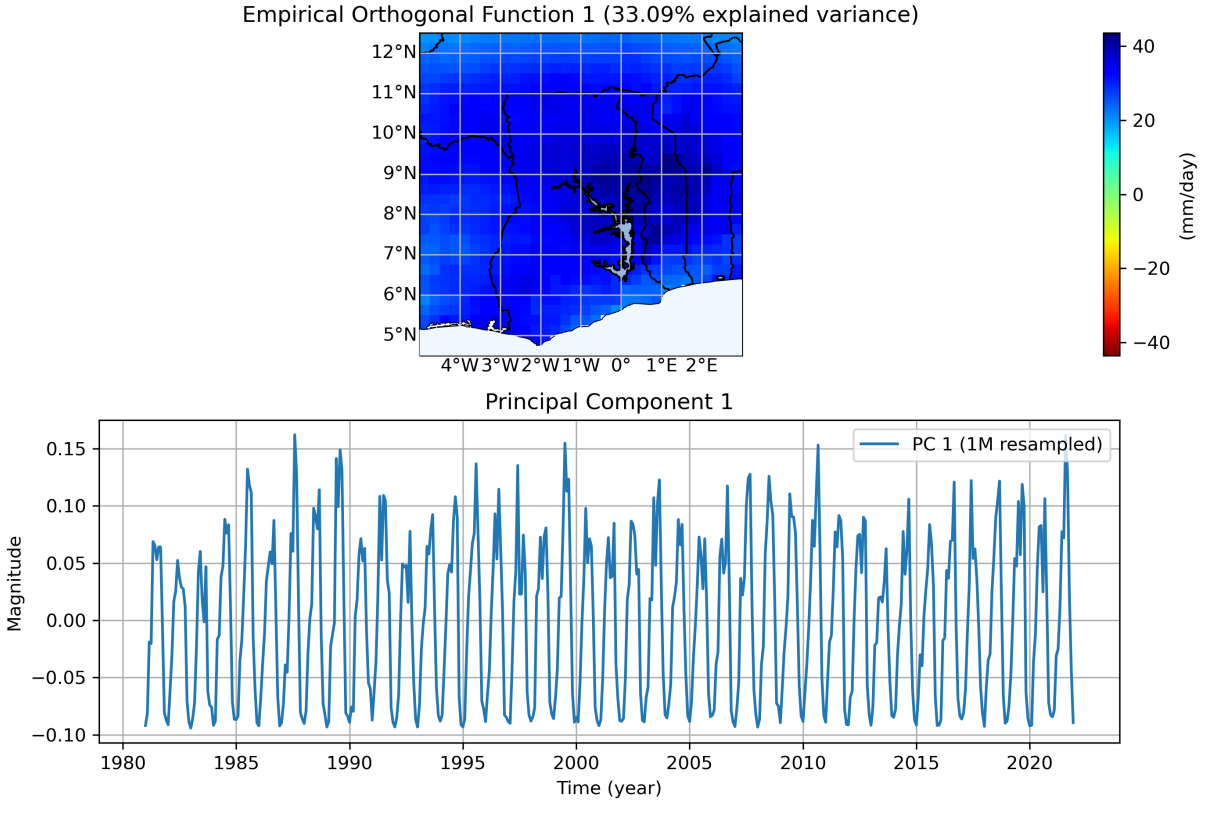

Figure 2.5: PCA of daily precipitation over Ghana - First EOF (top) and PC time-series (bottom) (CHIRPS 1981-2021). The top map shows the spatial loadings, while the bottom plot represents the PC time-series re-sampled with a time window of 30 days. The oscillation has a clear annual frequency, matching with the yearly alternation between wet and dry regimes.

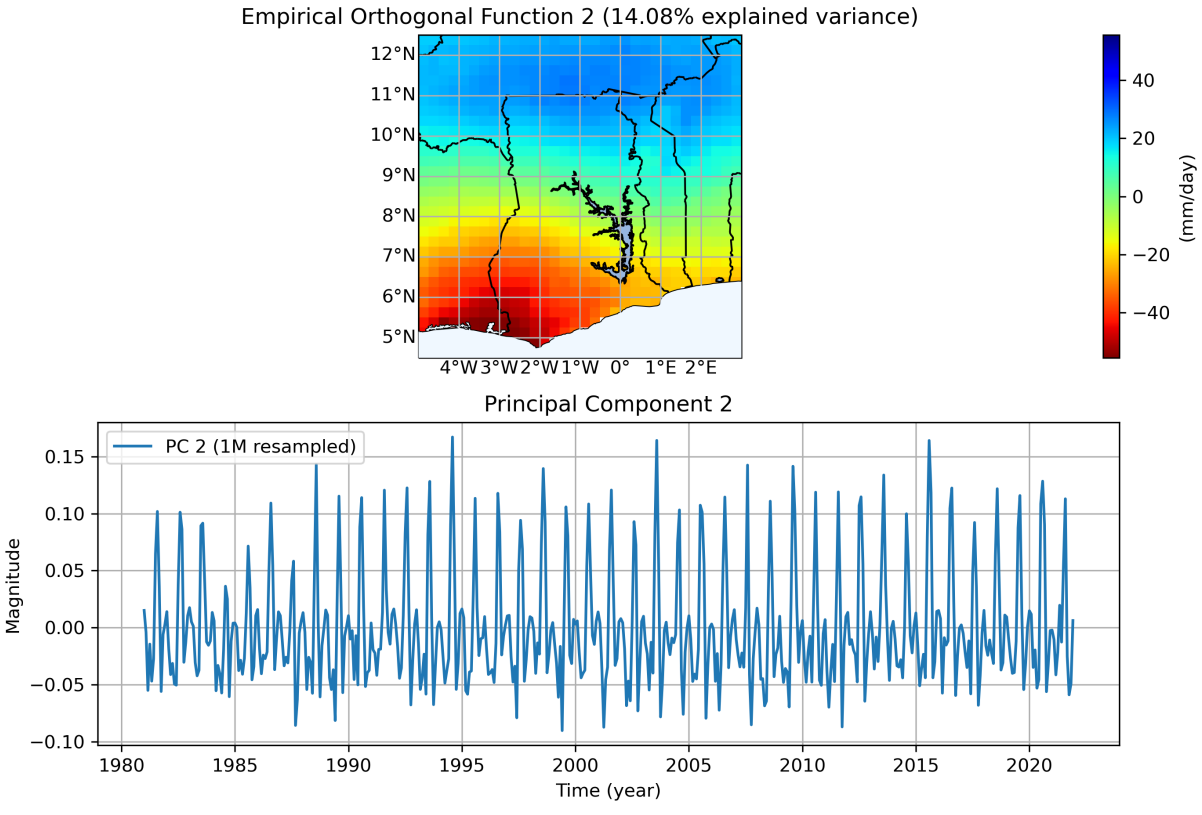

Figure 2.6: PCA of daily precipitation over Ghana - Second EOF (top) and PC time-series (bottom) (CHIRPS 1981-2021). The top map shows the spatial loadings, while the bottom plot represents the PC time-series re-sampled with a time window of 30 days

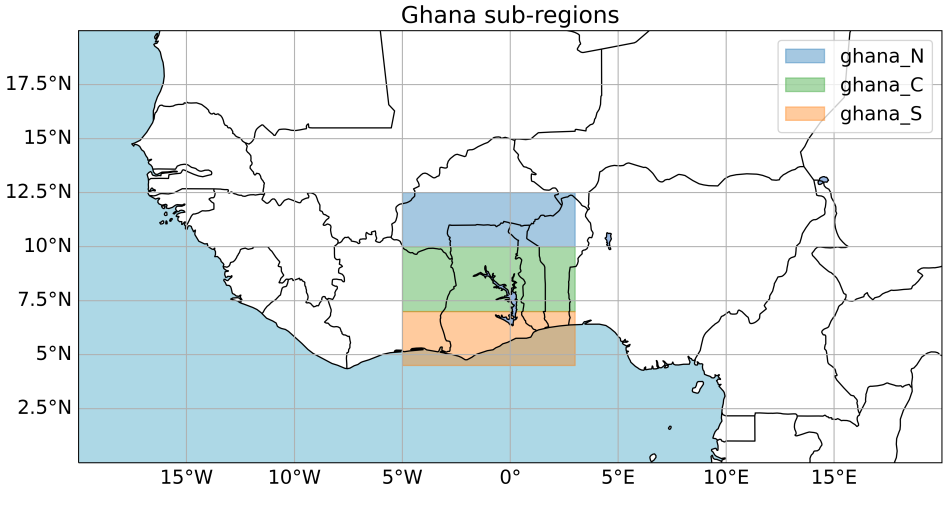

Figure 2.7: Map of West Africa depicting the sub-regions division performed over Ghana and neighbouring countries.

2.3.2 Regional Onset Detection

In this section we describe the method used to compute the 1981-2021 climatology of regional ORS. This stage of the study is crucial, as the choice of a method to identify the ORS implies also the choice of a definition of the onset itself. We follow the methodology introduced by Camberlin and Diop [11], which is based on a PCA of the precipitation time-series over the area of interest. The method can be described as follow:

1. given a sub-region as defined in Section 2.3.1, a PCA is performed on the historical time-series of

the daily precipitation anomaly (respect to the long-term daily mean). The rainfall data are square-root transformed, in order to avoid giving excessive weight to an heavy but isolated convective event. The PCA is performed in 'S-mode', viewing each point of the grid as a variable and each day as an observation;

2. the time-series of the first principal component amplitude (PC1) is isolated;
3. the PC1 time-series is split in "*detection windows*" which begin and end well within the dry season, encompassing a wet season in between;
4. for each *detection window* we compute the cumulative PC1 score value, which shows a decreasing trend during the dry season and an increasing trend during the wet period;
5. the onset of the rainy season is identified as the day when the cumulative PC1 curve reaches its absolute minimum.

Figures 2.8 shows this method applied to `Ghana Center` region for two showcase years, 1992 (**a**) and 2015 (**b**). In 1992 the rainy season has a clear onset on the very first days of April. On the opposite, in 2015 the detection of the ORS is less trivial: in such scenario, the choice of the onset's definition can shift the timing of the ORS of several days or even weeks. The PCA-based method identifies the start of the wet season only at the end of April, even if a remarkably wet period is observed already during March and even February. Unfortunately, this is not a rare situation and it is observed in several years of the analysis.

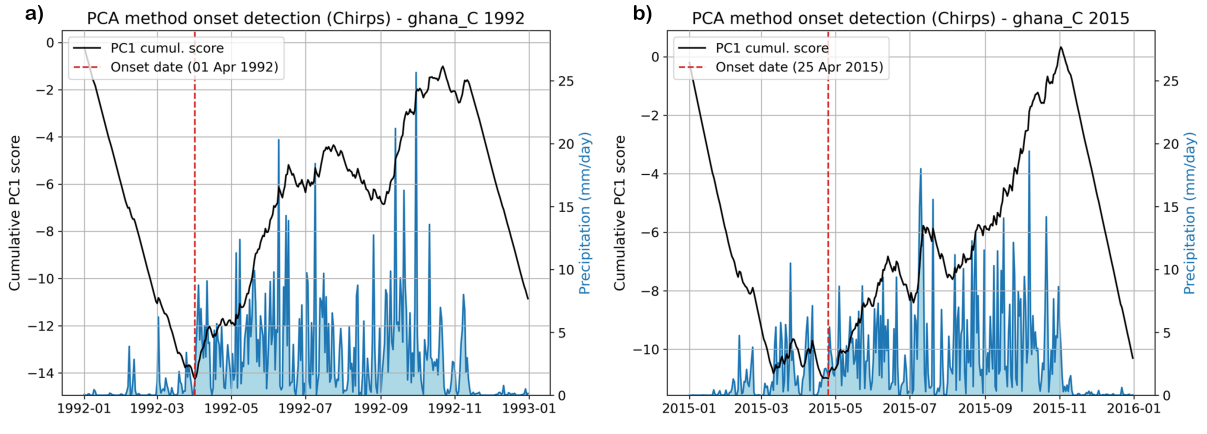

Figure 2.8: Onset detection over `Ghana Center` for 1992 and 2015. The blue lines and shadings represent the sub-region-averaged daily precipitation time-series, while the black line is the cumulative PC1 score. The regional onset date is identified by the vertical dashed red line as the global minimum of the cumulative PC1 score.

When the rainy season is lacking of clear onset conditions, the PCA-based method tends to be conservative, delaying rather than assessing an early ORS (with a subsequent dry spell). This bias matches the operational needs, as it is considered a worse prediction error to forecast the onset too early rather than too late. The method has also proven to be very reactive and precise for "sharp" onsets, which are easily detectable.

In conclusion, the PCA-based onset detection method combines two important features: it is purely based on rainfall data and it is specifically designed to be regional. In many other approaches, the options would have been either adapting a local detection method to act as regional, or selecting a regional method renouncing to the precipitation-based component. The decision to adhere to rainfall measurements is particularly desirable to obtained an historical records of the ORS resembling as much as possible the effective hydrological conditions of the soil, of primary interest for farming activities.

Visual Evaluation of the Regional Onset Detection

To assess the performance of the PCA-based ORS detection algorithm, we examined the spatial distribution of rainfall in the days preceding and following the onset. Figures 2.9, 2.10 and 2.11 were made through averaging the precipitation observed from 5 days before to 5 days after the historical onset. Consequently, each map does not represent the climatological average of precipitation recorded on a specific calendar date, but rather

the climatological mean of rainfall observed on the n^{th} day before/after each detected onset. This technique was also adopted by Camberlin and Diop [11] (who introduced the regional detection method) to show the good functioning of the PCA-based detection method, which effectively captures the shift in precipitation regime.

This maps provide additional information about the characteristics of the regional onset itself. Within each sub-region, we can recognise areas where the precipitation are most intense. This is because the regional onset derives directly from the first EOF of the PCA, which is always representative of the dominant oscillation's pattern. However, it only represents a single mode of oscillation, with the consequence that the obtain regional ORS will be more representative of the geographical area where the first EOF belongs to, and less for other parts of the sub-regions. This is an issue that will strongly emerge while evaluating the threshold-base regional forecast (Chapter 4).

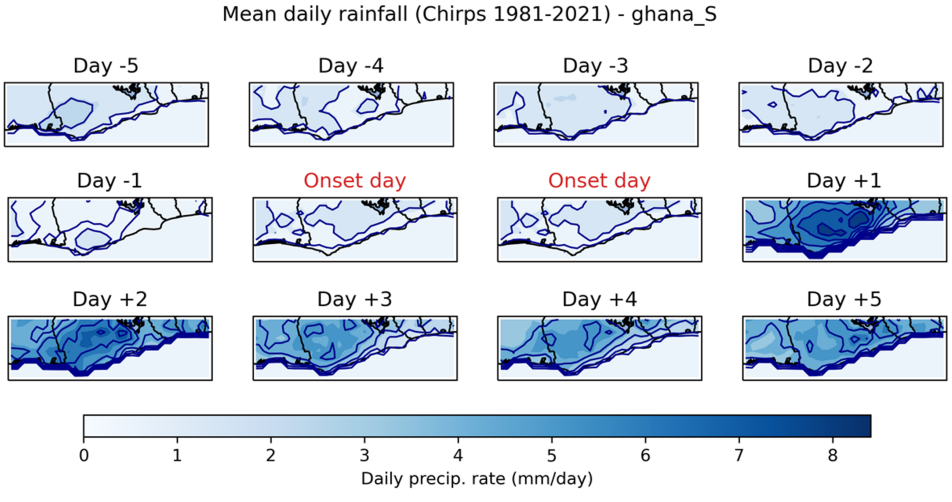

Figure 2.9: Composite map of daily precipitation in the 5 days preceding and following the onset over `Ghana South` (mean 1981-2021). Contour lines at every mm/day of precipitation, starting from 1 mm/day. To locate the sub-regions, use as reference the sub-region definition of Figure 2.7.

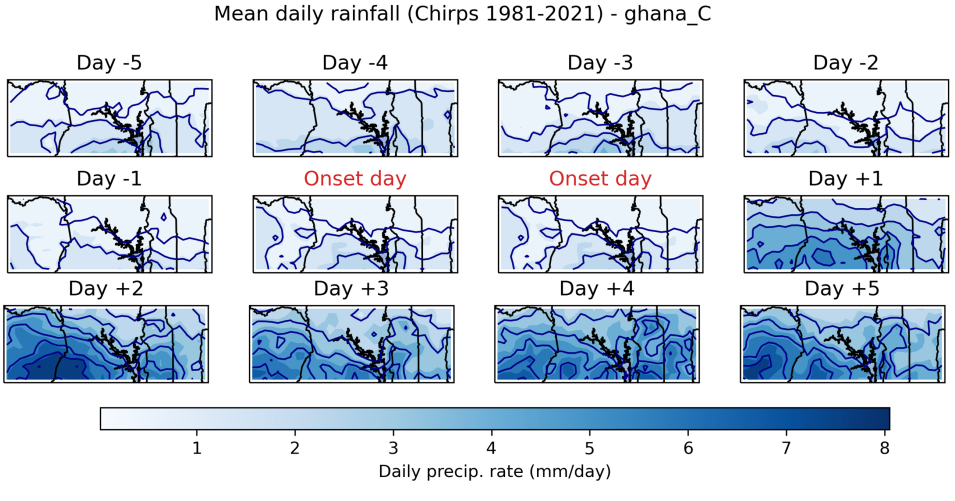

Figure 2.10: Composite map of daily precipitation in the 5 days preceding and following the onset over `Ghana Center` (mean 1981-2021). Contour lines at every mm/day of precipitation, starting from 1 mm/day. To locate the sub-regions, use as reference the sub-region definition of Figure 2.7.

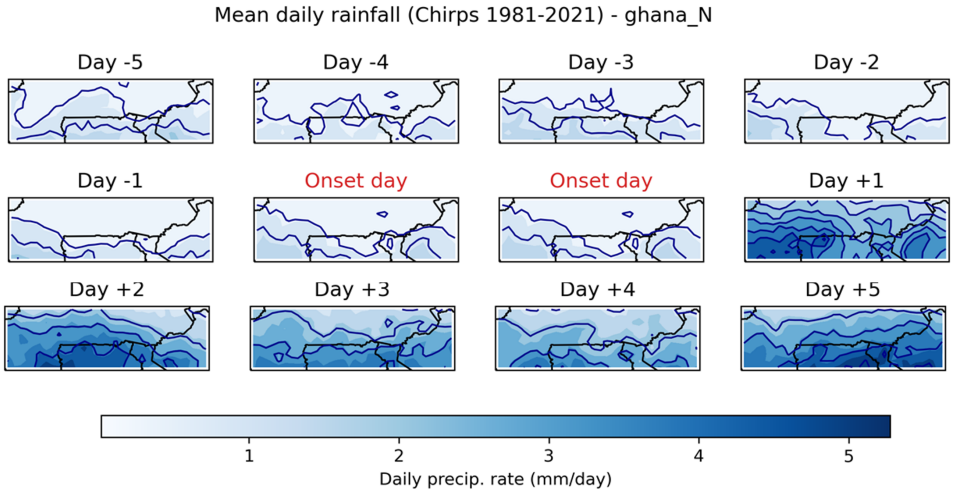

Figure 2.11: Composite map of daily precipitation in the 5 days preceding and following the onset over `Ghana North` (mean 1981-2021). Contour lines at every mm/day of precipitation, starting from 1 mm/day. To locate the sub-regions, use as reference the sub-region definition of Figure 2.7.

2.3.3 Results: Regional Rainy Season's Onset Climatology

The time-series of the regional ORS is characterised by the north-ward progression of the rainy season, matching what we could already observe with the local method (Section 2.2.2). On average, the onset of the `Ghana Center` sub-region is the one with the highest variability, with a standard deviation of almost 15 days. Despite in few consecutive years the southern and central regional ORS appear to be well correlated and close in time (e.g. 1987-1995), other periods show the ORS of `Ghana Center` behaving independently, or slightly resembling the one of `Ghana North`. However, considering the entire time records, we find low correlation among adjacent sub-regions ($r = 0.14$ for North-Center, $r = 0.28$ for Center-South), and very low among `Ghana South` and `Ghana North` ($r = 0.08$). The absence of a steady common pattern in onset's variability hints that the ORS of each sub-region is dominated by the variability of the atmospheric processes, which act differently is different sub-regions.

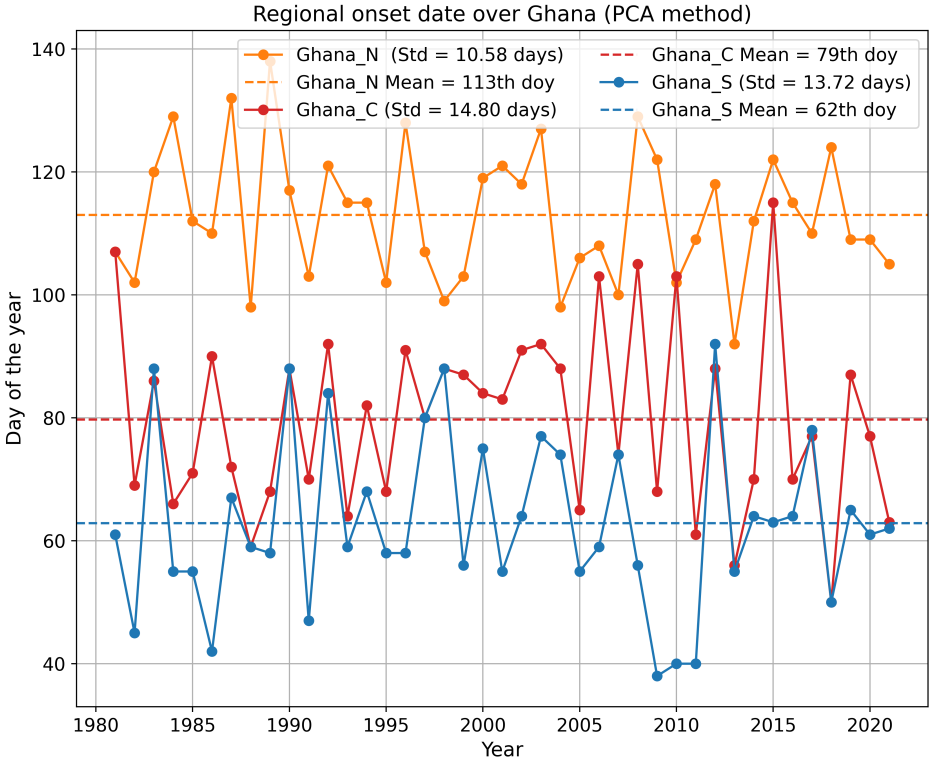

Figure 2.12: Historical time-series of regional ORS detected with the PCA-based algorithm using CHIRPS daily precipitation data (1981-2021).

Chapter 3

Atmospheric Circulation's Proxies of an Imminent Rainy Season Onset

The adoption of a regional onset definition is functional to investigate the link between the timing of the ORS and the large-scale atmospheric fields evolution. Throughout this chapter we analyse wind components in both the horizontal and vertical direction (and their divergence) at different pressure levels. In the following pages, the reader is provided first with an overview of the seasonal wind circulation over West Africa (Section 3.1), followed by the description of the method applied to find the correlations between the onset of the rainy season and the changes in atmospheric circulation (Section 3.2). Sections 3.3 and 3.4 present the findings of this research, along with an assessment on their operational application for the prediction of the ORS. The chapter ends with a brief analysis of the correlation between wind's circulation and the rain pattern behaviour at the beginning of the wet season (Section 3.5).

3.1 Climatology of Large-scale Wind Circulation

The climatological wind configuration over West Africa shows recurrent characteristics, which are typical of a monsoonal climate and which are the main drivers of the precipitation patterns of the region. Of primary importance is that both the low level (925 and 850 hPa) and the near-tropopause (200 hPa) circulation experience a reversal in wind's direction during the spring months, in conjunction with the beginning of the wet season.

Close to the ground, easterly flows interest the West African region north of 10°N with maximum intensity in the dry months (October to February). Proceeding south towards the Guinea Gulf, the winds turn progressively to the meridional direction, gaining maximum strength during the wet season (June to September). A wind convergence zone, called Inter-Tropical Discontinuity (ITD) [26], marks the encounter of the moist south-westerly winds with the hot, dry north-easterly flow blowing from the Sahara region (Harmattan flow). The strength of the low pressure region known as the Saharan Heat Low (SHL) determines the position of the ITD, which is pushed northward during the spring and summer months (Figure 3.1). Strong meridional temperature gradients and the intrusion of the south Atlantic moisture give rise to the wet season, which fully develops between June and September. At higher pressure levels the atmosphere is dominated by the zonal circulation (Figure 3.2). During most of the year the winds are directed eastward, with maximum strength over the Sahel; in the wet season (June to September), however, the pattern is reversed, with the stronger wind located south of 10°N and blowing in westward direction.

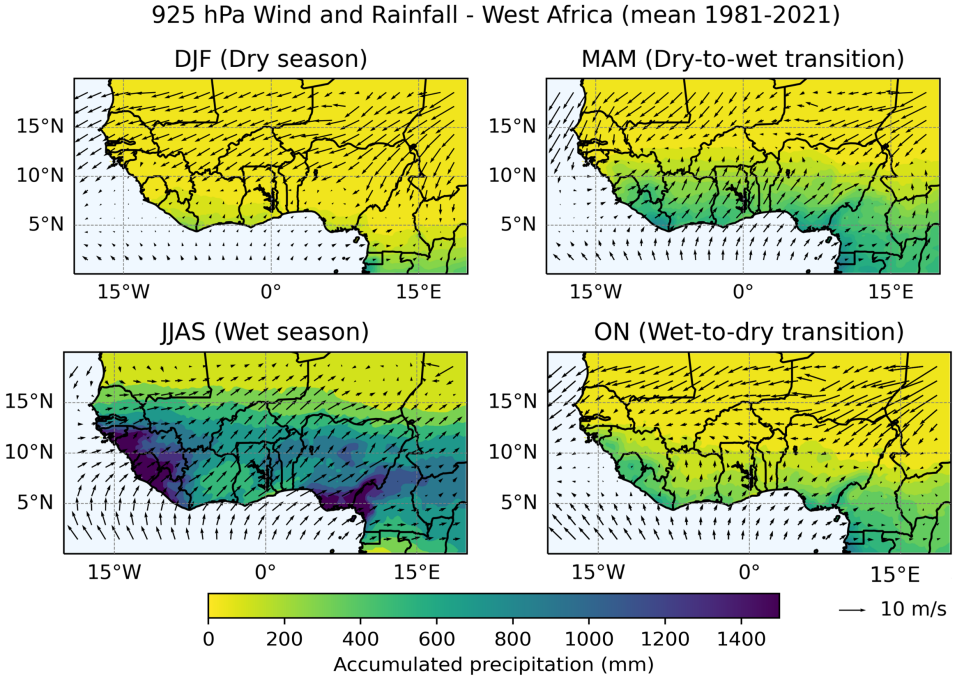

Figure 3.1: Composite maps of 925 hPa wind flow and accumulated rainfall (1981-2021 climate). ERA5 reanalysis (wind fields) and CHIRPS observation (precipitation data).

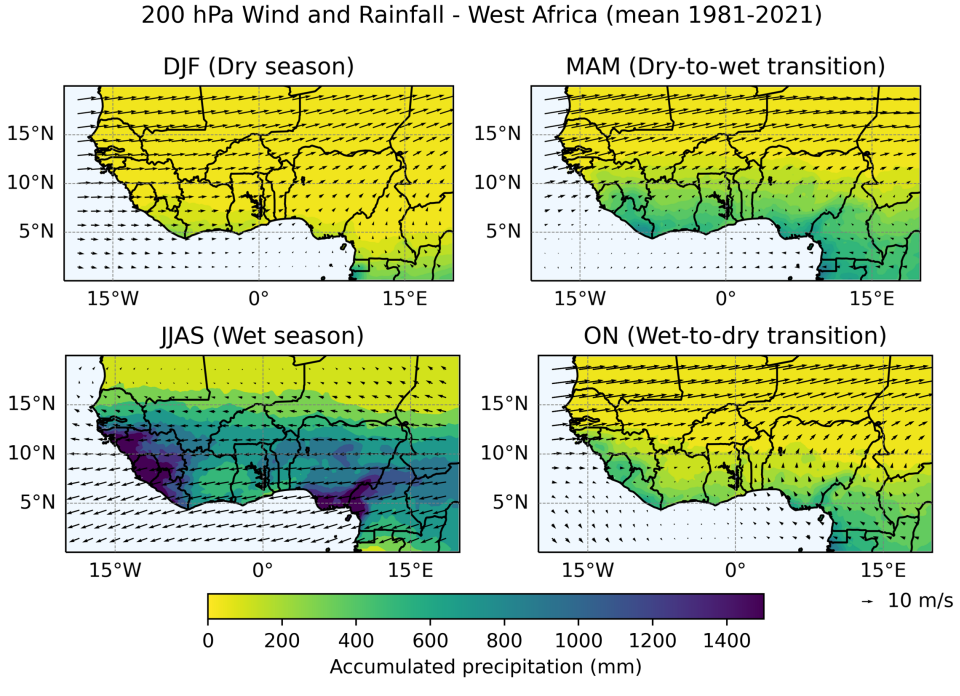

Figure 3.2: Composite maps of 200 hPa wind flow and accumulated rainfall (1981-2021 climate). ERA5 reanalysis (wind fields) and CHIRPS observation (precipitation data).

3.2 Methods - Correlating Onset and Wind Fields

Subsection 2.3.2 shows how the PCA-based onset detection method is able to capture the dry-to-wet transition in historical precipitation's data. A similar approach, described in subsection 3.2.1, is here applied to the ERA5 wind fields listed in the Data section. The composite maps are used to identify the most promising atmospheric fields, which are further investigated under a year-by-year lens (section 3.2.2). This allows to

analyse the wind's inter-annual variability, giving insight for operational purposes.

3.2.1 Composite Maps Creation

The shift of rainfall regime is previously illustrated as precipitation's composite maps, showing the daily precipitation climatology for the 5 days preceding and following the detected ORS (Figures 2.9 to 2.11). The very same procedure is repeated for the different wind components, adopting an extended time window. Given a wind field, such as the meridional wind at 850 hPa:

- for each year of the climatology and for a chosen sub-region, we select from ERA5 wind re-analyses the 15 days before and after the computed ORS;
- the selected data are averaged day-wise. The result is a single time-series, centred on an "artificial" onset date, where the n^{th} element is the average of all n^{th} days before/after the climatological onset dates.

The newly-obtained time series display the evolution of the chosen wind field over 30 days centred on the "artificial" onset date. The collection of composite maps that results for each wind component is a visual tool to identify patterns in the atmospheric circulation and their changes occurring during the ORS. The most interesting example of such products are given in the following *Results* section (3.3.1).

It is important to mark the distinction between the recently-discussed composite maps, showing atmospheric fields, from the ones depicting precipitation data. To this scope, we need to consider the origin of these maps. They derive from a PCA applied to precipitation records, from which the dates of the onset are determined. It follows that the clear change in the intensity and spread of the rainfall observed in Section 2.3.2 is a strong indication of the effectiveness of the PCA-based method in highlighting the precipitation's regime shift. On the other hand, we need to be cautious about its physical and climatological interpretation: since the precipitation composite maps show the same quantity used to identify the ORS, it is possible that the change in rainfall pattern appears "artificially" well-defined, in terms of both rainfall intensity and speed of the dry-to-wet transition.

By contrast, these remarks do not apply to the composite maps showing changes in atmospheric fields: the latter derive from a dataset (ERA5 reanalysis) that is fully independent from the precipitation's records one (CHIRPS). Therefore, the observed changes in atmospheric patterns have a high statistical significance and represent a trustworthy climatology of the large-scale circulation changes related to the onset of the rainy season.

3.2.2 Year-by-year Analysis

The composite maps generated in the previous section return a clear description of the dry-to-wet transition from the large-scale circulation point of view. Though, as we are interested in the operational usage of this information, it is necessary to abandon multi-decadal averages and to look into the inter-annual variability of such atmospheric fields, evaluating their applicability for operational purposes.

Two perspectives were used to study the relations between the wind circulation and the observed rainfall. The first is a broad seasonal view, where we analysed the wind profile respect to the precipitation records across the entire year. This was performed for single wind components, averaging the wind speed of a specific sub-region. The second perspective narrows the time window, analysing different wind's indicators at or around the time of the onset. The used quantities are again averaged over the considered sub-region, and the time-series is smoothed with a 30 days running mean. In this way, we remove short time scale variation, which are not representative of the seasonal evolution of the considered field.

3.3 Results

3.3.1 Wind Composite Maps

Through composite wind fields maps, generated in section 3.2.1, we detect precise circulation patters occurring during the onset of a rainy season. The interesting patterns are the ones showing a gradual change in the direction of the considered wind's component. This shift in wind's direction develops with a south-to-north

progression, following the longitudinal axes that has been identified in Section 2.1 studying the development of the rainy season in the climate.

The zonal wind at 200 hPa is a prominent example to illustrate the typical monsoonal wind shift. Here we illustrate the collection of composite maps shown in Figure 3.3: these plots deliver the maximum amount of information if analysed keeping in mind the chronological development of a standard rainy season. In this analysis, there are two time-lines to follow:

- the explicit ordering constitute by the number of days before and after the ORS;
- the ordering given by the climate of the rainy season across the sub-regions, which in Section 2.1 is shown to start from `Ghana South` and then to proceed towards `Ghana Center` and `Ghana North`.

The second time-line allows to capture the evolution of the wind component across all the rainy season, not only in the fifteen days around the onset of a specific sub-region.

At the beginning of the year, before that the rainy season takes place in the south of Ghana, the upper branch of the Walker circulation is positive across all the region, with the wind at 200 hPa blowing east-ward (Figure 3.3 **a**). Though, it is already visible a meridional gradient of the wind intensity, which weakens approaching the coastal area. After the onset in `Ghana South` the east-ward flow progressively weakens, eventually reversing to the west direction. Moreover, the line of wind reversal is located south of the sub-region where the ORS takes place.

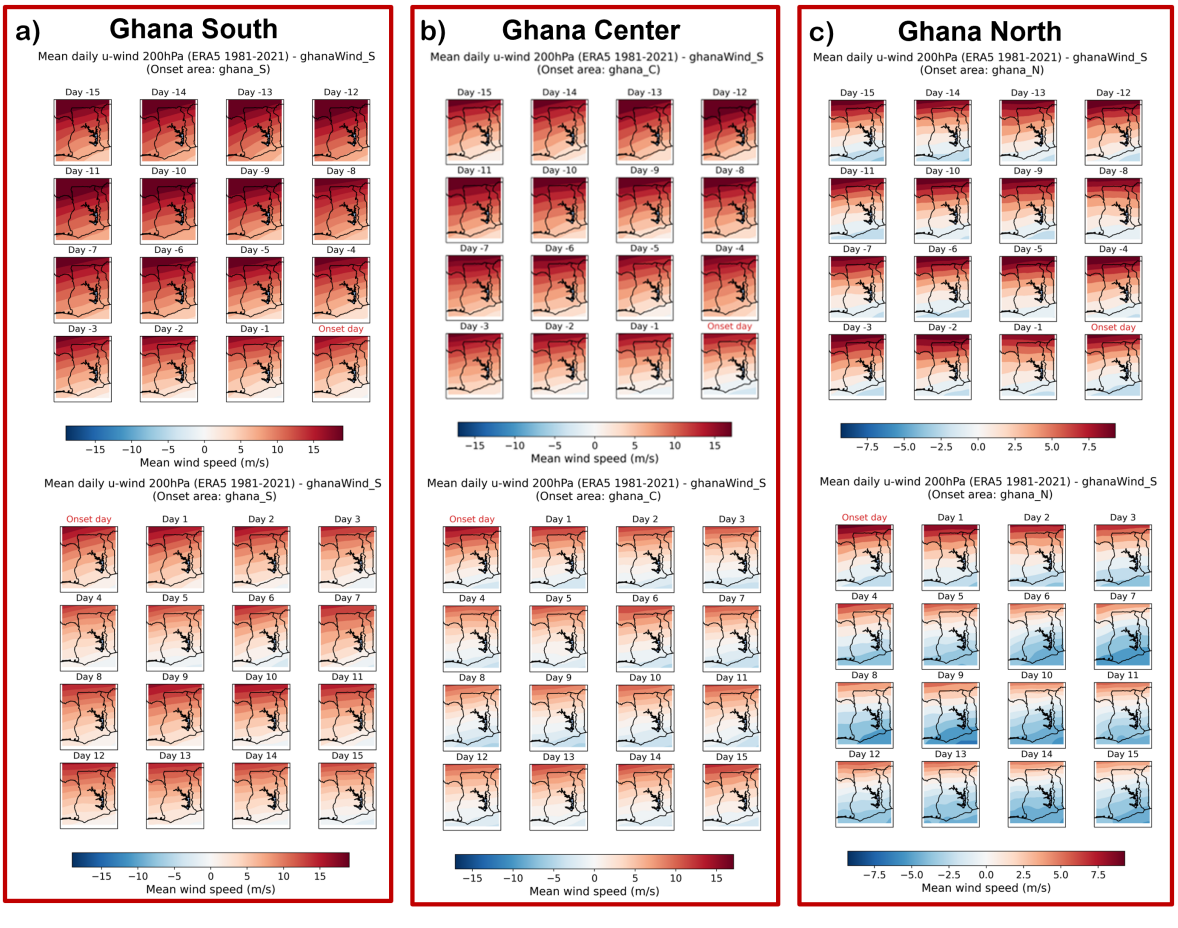

Figure 3.3: Zonal wind velocity at 200 hPa during the month of the onset (ERA5, 1981-2021 daily average). A wind reversal happens during the spring season, proceeding from the coast towards the inner part of the country.

The meridional wind at 925 hPa exhibits a behaviour similar to the zonal wind at 200 hPa. A line of wind convergence between southerly and northerly wind moves towards the interior of the region, with the difference that now the reversal of the wind anticipates the start of the wet season. This behaviour is in line with the

atmospheric circulation's theory underlying the WAM, according to which the maximum of convective precipitation is located few hundreds km south of the Inter-Tropical Front (the line of meridional wind convergence) [19]. Fig. 3.4 illustrates this phenomenon for the onset date of `Ghana South`. In the left panel (**a**) it is shown a slow north-ward progression of the ITF, which experiences a sudden shift in the days following the onset (**b**). During the ORS a maximum of north-ward wind speed expands over the south of the country, while the convergence zone is locate in the northern region. The same pattern is also observed for the ORS of `Ghana Center` and `Ghana North`. Other two wind components showing interesting evolution during the onset season are the meridional wind at 850 hPa and the zonal wind at 925 hPa. Their behaviour closely resemble the pattern of meridional wind at 925 hPa. A version of the Fig. 3.3 for these wind components has been included in the Appendix (Figures A.1, A.2, and A.3).

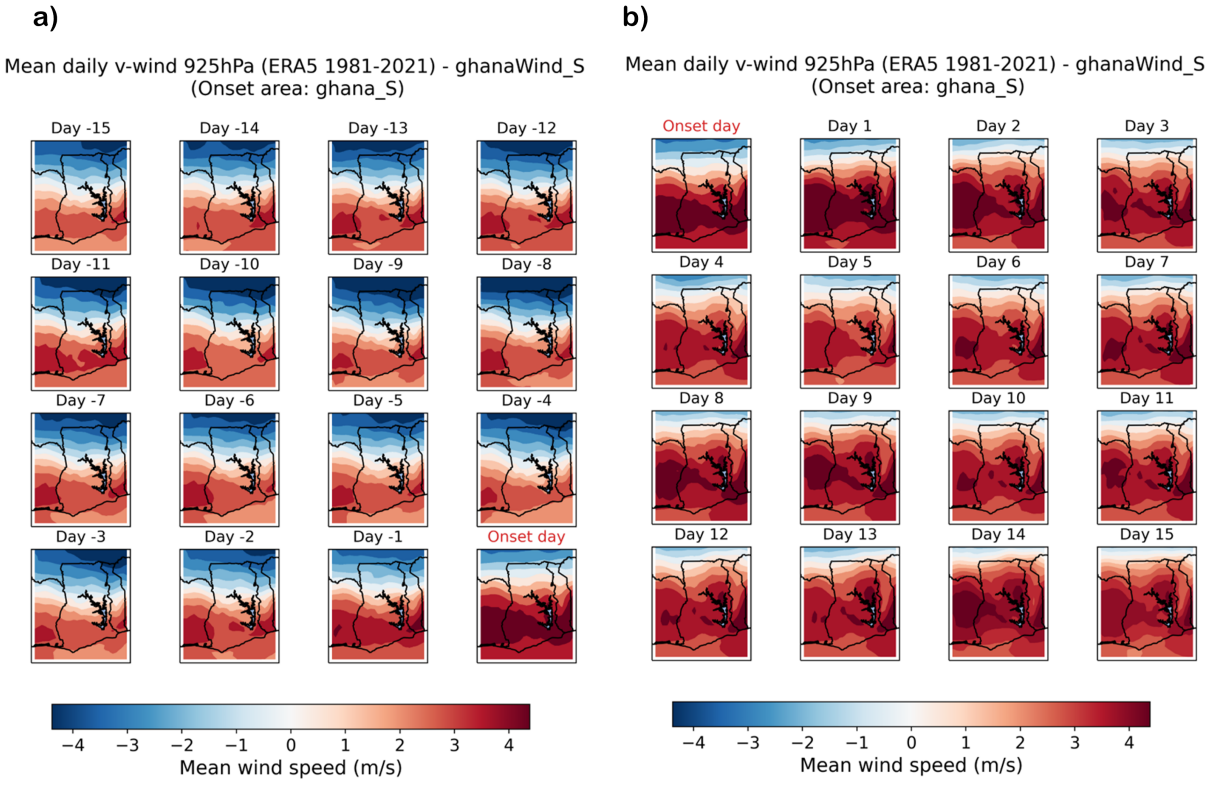

Figure 3.4: Meridional wind velocity at 925 hPa during the month of the onset (ERA5, 1981-2021 daily average). The ITF is clearly marked by the band of wind convergence, which progresses northward and makes the circulation of the regions dominated by southerly winds.

To summarize, this analysis has identified four wind components that have a strong climatological relation with the ORS. The atmospheric circulation resulting from these wind patters is depicted in Fig. 3.5 for each sub-region. The plots show the daily climatology of mean wind speed and rainfall amount, averaged over the chosen area. The following conclusions can be drawn:

- the meridional and the zonal wind components at 925 hPa are closely related and experience a marked direction reversal in `Ghana North` and `Ghana Center` (Fig. 3.5 **a and c**); the close-to-surface wind circulation in `Ghana South` is instead positive (north-ward and east-ward) throughout all the year, therefore the increase in wind speed does not cause a change in wind direction;
- the meridional wind at 850 hPa shows a similar behaviour across all the three sub-regions, switching from north-ward to south-ward direction. This reversal occurs considerably later than the one observed for the close-to-surface winds;
- the zonal wind at 200 hPa flows with opposite direction respect to the surface winds: here strong westerly winds weaken in spring to be replaced by easterlies blowing with increasing intensity during the summer. The differences between the sub-regions are less marked than what observed at 925 hPa and 850 hPa, but still present. In particular, the time when the zonal wind changes direction is shifted

of about 1 month between each sub-regions. Ghana South starts to experience easterlies around the beginning of April, while in Ghana North this happens only at the beginning of June.

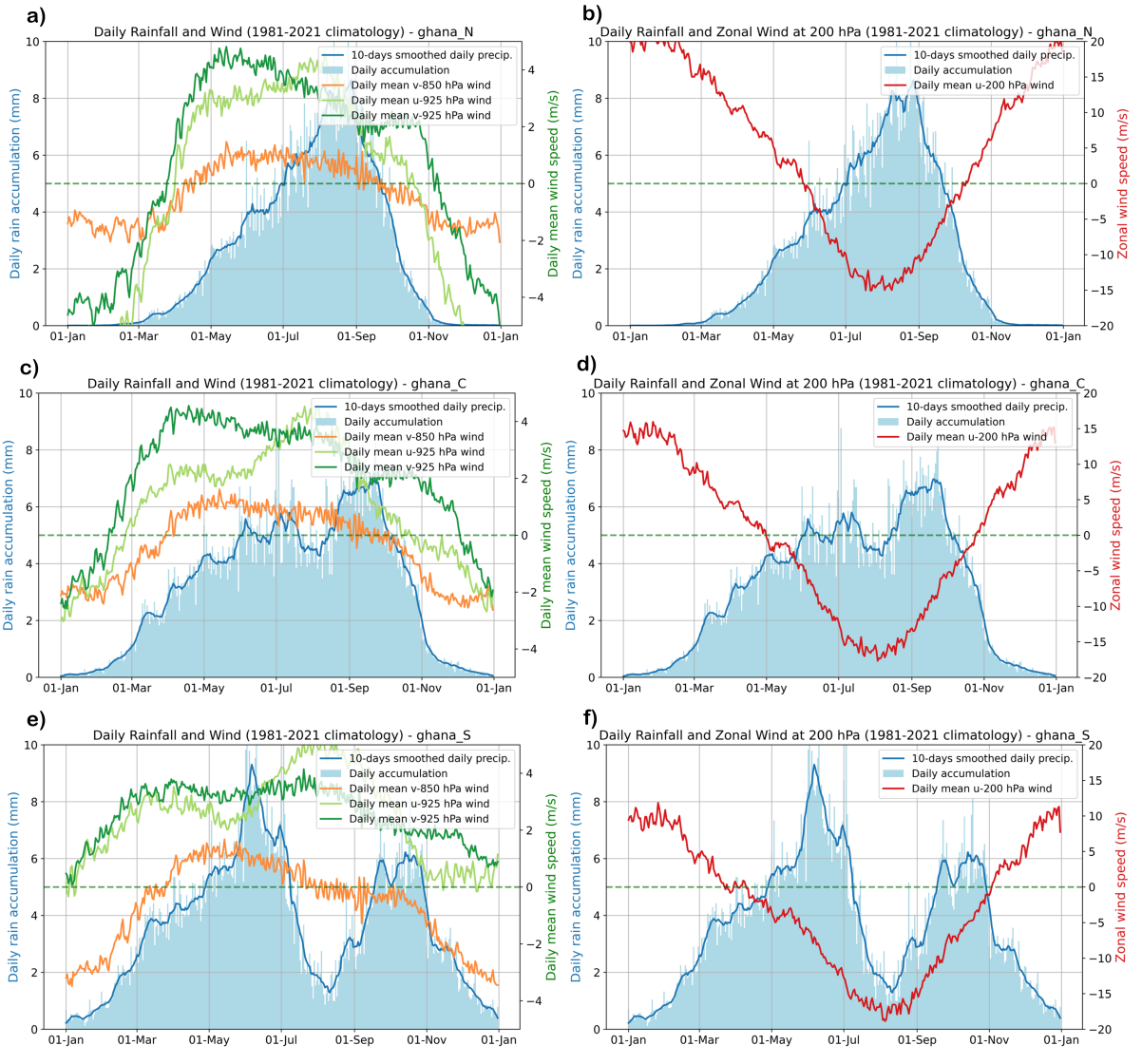

Figure 3.5: Mean daily precipitation and wind speed in the 1981-2021 climate (CHIRPS and ERA5). Plots **a, c and e** show the meridional wind at 850, 925 hPa and the zonal wind at 925 hPa; plots **b, d and f** show the zonal wind at 200 hPa. Every wind component has a similar behaviour, with a change in wind direction occurring in the spring months. The wind speed's magnitude differs between close-to-surface winds an the wind close to the tropo-pause.

The vertical wind in the mid-troposphere is a standard proxy for convection and, consequently, for convective activity. Since the rainy season is essentially constituted by "waves" of convective events, the vertical wind at 500 hPa has been analysed. Figure 3.6 shows the development of the vertical wind velocity in the 1981-2021 climatology. The unit of vertical wind intensity is Pa/s, which implies that negative values represent up-draught, while positive represent down-draught. Precipitation is expected during up-draught events, therefore the scale on the y-axis has been reversed in the sake of easier interpretability of the plots.

The vertical wind has a very close correlation with the precipitation, with Pearson coefficients reaching -0.90 for Ghana South and -0.95 for Ghana Center and Ghana North. This not only confirms that the vertical wind at 500 hPa is a strong proxy for convection, but also that the bulk of rainfall observed over Ghana during the rainy season is indeed generated by convective events. Furthermore, the seasonal behaviour resembles quite closely what observed for the horizontal wind: during spring, a progressive strengthening of the wind velocity generates a reversal of the wind direction, from down-ward to up-ward, which lasts for the entire rainy season. It is interesting that this behaviour is clearly visible even for the second rainy season of Ghana South

(Figure 3.6 **c**), while it was not distinguishable in any other wind component. To conclude, we point out that the strong correlation of the vertical wind with precipitations makes the wind profile prone to short-term variability, which reflects the highly unpredictable nature of convective events.

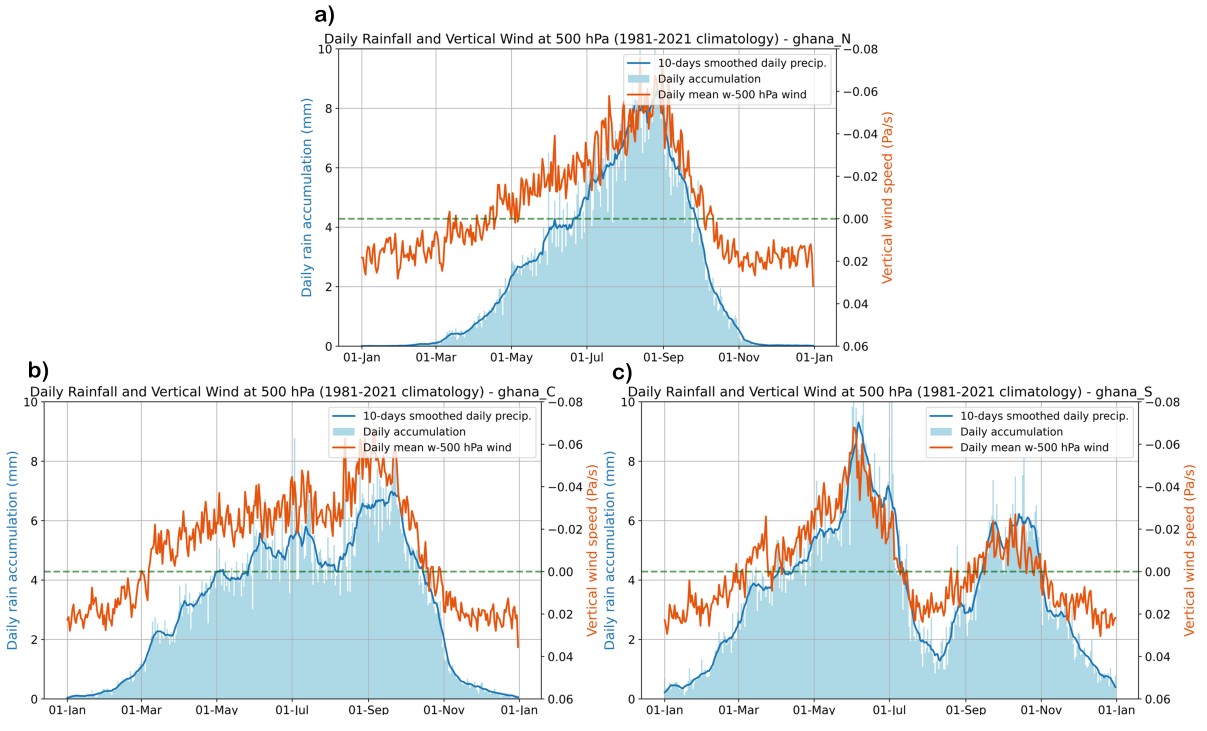

Figure 3.6: Mean daily precipitation and vertical wind speed at 500 hPa in the 1981-2021 climate (CHIRPS and ERA5). The vertical wind has an extremely high correlation with the precipitation's time series. This is due the tight link between the mid-tropospheric vertical wind and convection, which is the main source of monsoonal rainfall.

3.3.2 Inter-annual Variability

In this section we examine the wind components identified earlier (Section 3.3.1) to determine potential predictors of the ORS. As first step, we investigate the correlation between the daily wind profile and the precipitation records. The correlation coefficient of these two time series, smoothed with a 30-days running mean, has been computed for every year of the climatology. Table 3.1 contains, for each wind's component and each sub-region, the mean and the standard deviation of such coefficients.

	Ghana South		Ghana Center		Ghana North	
	r	**std**	**r**	**std**	**r**	**std**
V 850 hPa	0.72	0.08	0.80	0.06	0.78	0.05
U 200 hPa	-0.42	0.14	-0.84	0.07	-0.91	0.03
V 925 hPa	0.46	0.15	0.68	0.08	0.69	0.06
U 925 hPa	0.29	0.17	0.78	0.07	0.79	0.04
W 500 hPa	-0.87	0.06	-0.93	0.03	-0.94	0.03

Table 3.1: Correlations between the 30-days smoothed precipitation and wind time series (1981-2021 mean). `Ghana South` stands out for the lowest correlation and the highest inter-annual variability respect to every wind components.

The wind components of the previous table can be distinguished into three groups:

1. **meridional wind at 850 hPa**, which is positively correlated with precipitation across all the sub-regions. The coefficients are remarkably high, considering that they are computed on daily values;

2. **meridional wind at 925 hPa, zonal wind at 200 hPa and 925 hPa**. These wind components show high correlation in `Ghana Center` and `Ghana North`, but low in `Ghana South`. Focusing on the latter region, the relatively high standard deviation implies that the south is where we observe the highest

inter-annual variability. It suggests that the precipitations occurring along the coast are weakly related to the pattern of the wind component belonging to this group. It is likely that the ocean-land interaction is responsible for smaller scale processes leading to rainy events, which are not to be linked to changes in the seasonal wind circulation;

3. **vertical wind at 500 hPa**, which shows extremely high correlation with the precipitation records, and very low variability. This denotes that the vertical wind is an accurate and stable precipitation proxy over all sub-regions, even on a daily time scale.

Two remarks need to be made about the previous results. Firstly, we emphasize that the precipitation's time series and the wind profiles origin from two distinct and independent datasets (respectively CHIRPS and ERA5 reanalyses). Therefore, these correlations are a proof of both the reliability of CHIRPS's rainfall observations and of the ability of ERA5 re-analyses to provide an atmospheric circulation picture consistent with the observed precipitation patterns.

The second remark concerns the vertical wind at 500 hPa. This wind component exhibits the highest correlation with rainfall records, making it appear the most promising predictor of the start of the rainy season. On the other hand, the extremely high correlation makes the vertical wind profile dominated by short-time scale variability, with the result that the wind profile is as spiky as the precipitation record itself. Despite being clearly visible in the climatology (Figure 3.6), on a daily/weakly basis it is very difficult to identify stable trends of such wind component. For this reason, we need to exclude the vertical wind from the list of possible predictors.

The second step of our analysis focuses on the wind patterns observed at or in proximity of the time of the onset. Figure 3.7 displays the speed of the wind components at the day of the onset for every year of the climatology (1981-2021). Interesting results follows from the comparison of these plots with the annual wind pattern shown in Figure 3.5:

- the meridional wind components, both at 925 and 850 hPa, show a trend which follows the seasonal circulation pattern observed in Figure 3.5. As the meridional wind increases during the development of the season, so does the wind speed at the onset, which is then correlated with the timing of the onset itself. At 850 hPa (Figure 3.7 **a**) this pattern is shared by each of the sub-regions, without notable differences among them. On the other hand, at 925 hPa (Figure 3.7 **c**) `Ghana South` is less affected by this trend respect to the other sub-regions;
- 93% of the onsets occur when the zonal wind at 200 hPa is blowing east-ward, and the share rise to 99% when considering the wind at 925 hPa. While these two plots look similar (Figure 3.7 **b** and **d**), they describe two opposite patterns as the background wind's circulation at the two pressure levels is opposite. At 200 hPa, the zonal wind is switching from east-ward to west-ward direction, therefore the rainy season (almost) always starts before that the wind has turned to be west-ward. At 925 hPa, the zonal wind turns from blowing west-ward to east-ward and, consequently, the observed ORS happens always after that the wind reversal has occurred.

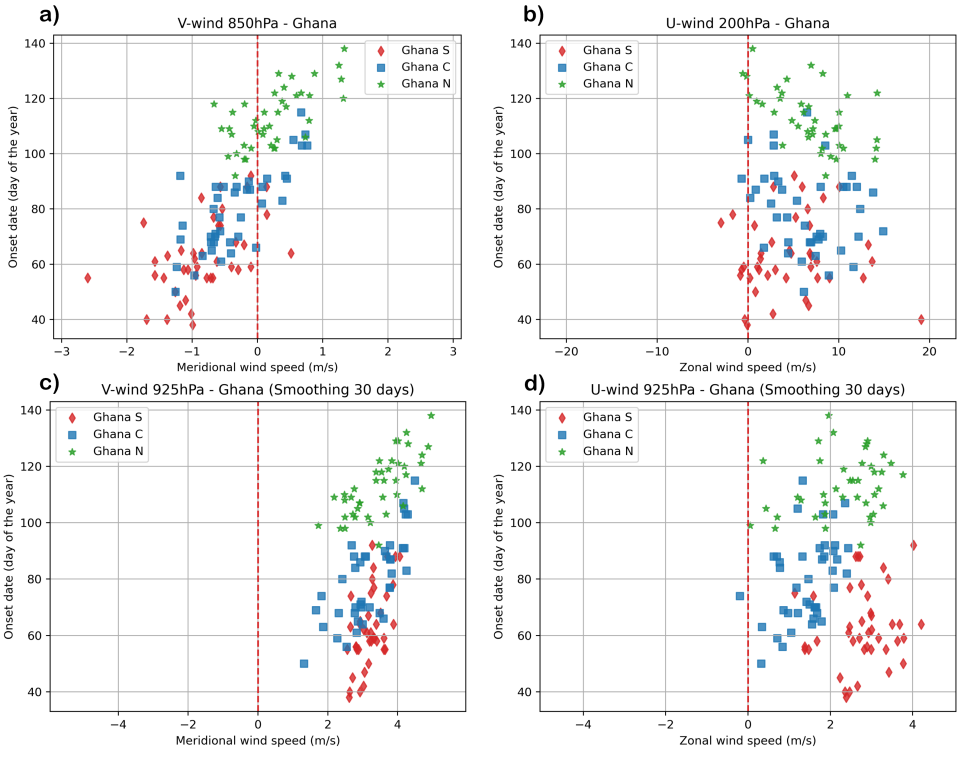

Figure 3.7: Wind speed on the day of the rainy season onset for: **a)** meridional wind at 850 hPa, **b)** zonal wind at 200 hPa, **c)** meridional and **d)** zonal wind at 925 hPa. The wind speed is always intended as the daily value of the 30-days smoothed curve.

From Figure 3.7 we can identify intervals of wind speed where the ORS has been observed, which could represent a prototype of an onset's predictor. However, it is necessary to compare the extend of such intervals with the actual wind speed values observed during the beginning of the rainy season. This is done in Figure 3.8 for the meridional wind at 850 hPa and the zonal wind at 200 hPa, while the equivalent for both components of the wind at 925 hPa is found in Appendix (Figure A.4). The plots show that the identified intervals are rather large respect to the natural wind's variability, with the result that they encompass wind values observed in a time window of at least one month. Moreover, most of the wind's profiles are not monotonic curves, meaning that the same wind speed can be recorded more than once during a rainy season. This and other issues will be extensively analysed in Section 3.4, where we draw conclusions about the feasibility of an operational usage of wind information.

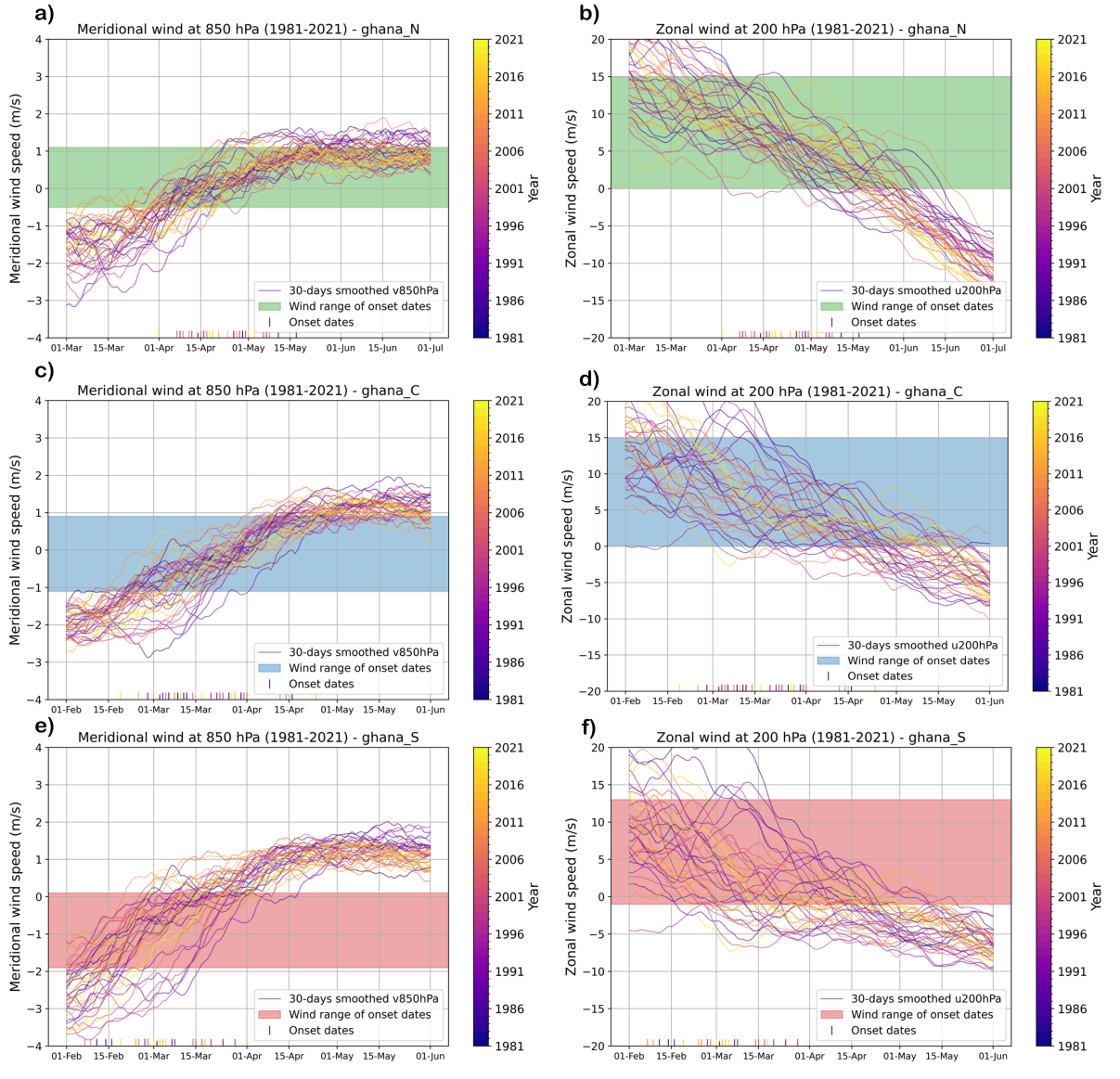

Figure 3.8: 30-days smoothed wind profiles during the start of the rainy season (ERA5 1981-2021). The shaded bands mark the wind speed intervals where the ORS is observed during the year of the climatology. The interval's boundaries have been computed according to Figure 3.7. The rainy season's onset dates corresponding to each plotted year are shown as a rug plot above the x-axis.

3.4 Discussion: Applicability of Wind-based Predictors

This section discusses the previous results through the lens of the operational forecasting needs, highlighting strengths and flaws of the wind-based predictions of the ORS.

The existence of a fixed and easily-detectable wind patter during the ORS would be ideal to generate wind-based onset's predictions. Despite the results of Section 3.3.1 demonstrate that some patters are strongly present in the climate, the same conclusion does not hold when analysing the years individually. To further display the inter-annual variability of the wind, Figure 3.9 collects the precipitation and the 850 hPa meridional wind records of six different years in `Ghana Center` sub-region. The collection of years was chosen to be representative of two different but recurring scenarios: when the rainy season has a very sharp and well-defined onset (**left panels**), and the opposite situation of a very smooth start of the wet season (**right panels**). A more in-depth analysis of the "shape" of the ORS can be found in Section 3.5.

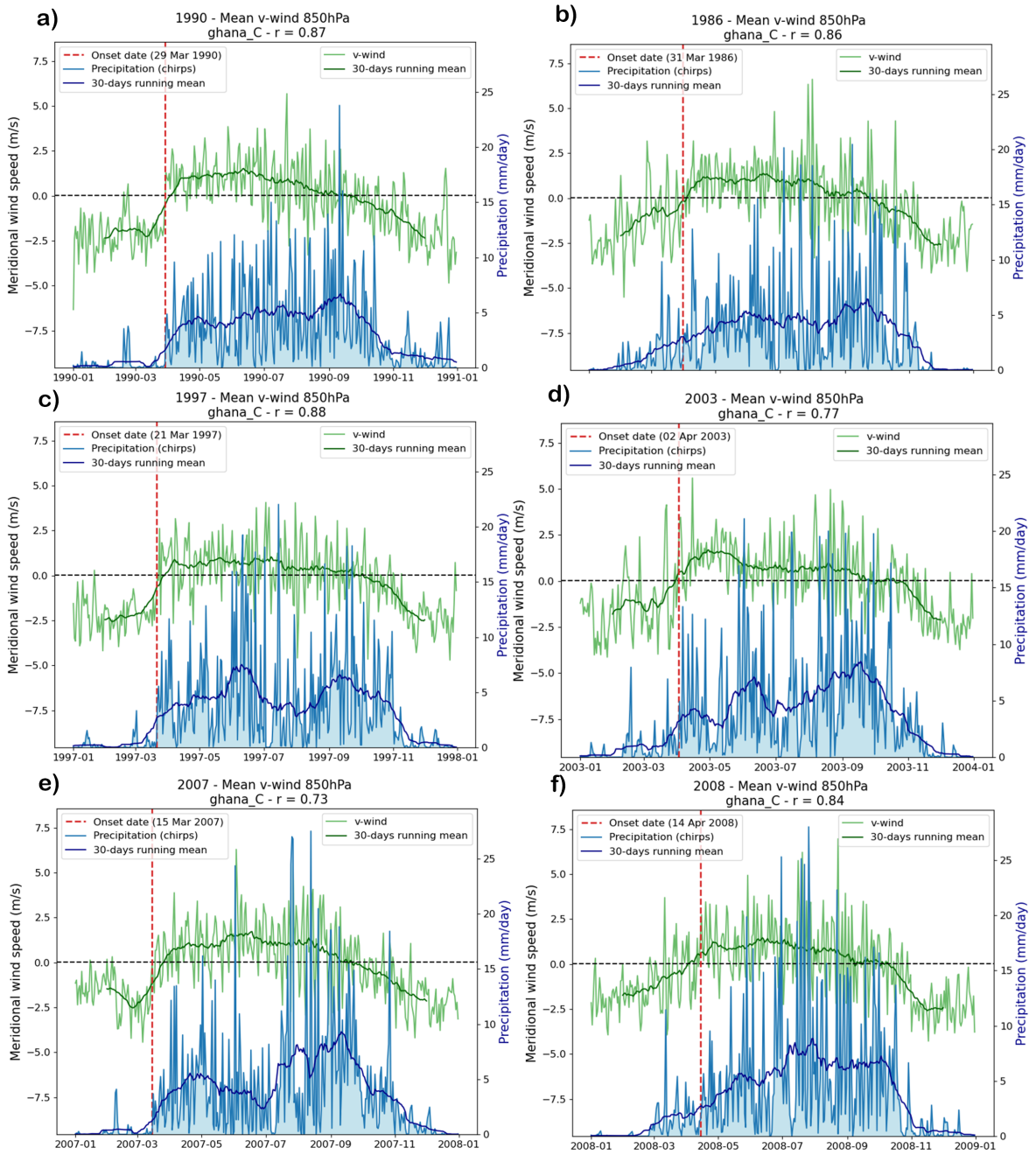

Figure 3.9: Meridional wind at 850 hPa and rainfall for six different years in `Ghana Center` (ERA5 and CHIRPS). The left hand plots (**a**, **c** and **e**) shows years when the onset of the season is well-defined, while on the right (**b**, **d** and **f**) the rainy season has a much smoother start.

While the previous plots display both the daily values (bright lines) and the 30-days smoothed values (dark lines), it is best to focus on the first ones as the 30-days running mean is never available when producing forecasts. From the daily wind profiles of the "sharp onsets" (**a**, **c** and **e**), it appears that a threshold on the 850 hPa meridional wind (around -0.5 m/s) could provide a first predictor of the ORS.

However, in the years when the onset of the season is not well-defined (**b**, **d** and **f**), the wind profile is extremely spiky. Some peaks, located well before the observed ORS, have magnitude similar to what recorded at or later the onset date. In the latter situation, implementing a wind speed threshold would certainly result in a very early detection of the ORS. This was already observed in Figure 3.8, comparing the intervals of wind speed when the historical onset are recorded with the inter-annual variability of the wind profile. Therefore, there are severe limits to what can be achieved using wind thresholds as direct predictors of the rainy season onset.

3.5 Seasonal Outlook: The Sharpness of the Rainy Season Onset

The onset of the rainy season at a certain location can be characterized by what we call here *sharpness*, to say how gradual or sudden is the transition from the dry to the wet season. This information is not particularly important for the end-users of the forecast, usually small-holder farmers, because it has weak agronomic consequences. It is though important for the forecasters themselves: the current onset's detection methods have a much higher uncertainty when the rainy season has a gradual start, while they perform better for very sharp onsets. Knowing beforehand what a rainy season will look like can therefore put the forecasting system on alert about tricky situations, providing a better estimate of the prediction's uncertainty.

3.5.1 Methods

A tailor-made index, hereby called ***SHPindex*** (SHarPness index), has been developed to investigate the sharpness of the ORS. It is defined through the amount of rain that has fallen in a time window of 60 days, centred on the observed start of the rainy season; the index returns a value between 0 and 1 representing the fraction of rainfall registered after the onset, compared to the total precipitation during the defined time window. The mathematical description is:

$$SHPindex = \frac{P_{cum}(+30days)}{P_{cum}(-30days) + P_{cum}(+30days)}, \tag{3.1}$$

where $P_{cum}(\pm\#days)$ represents the cumulative amount of rainfall recorded in a number of days preceding $(-)$ or following $(+)$ the onset.

The performance of the algorithms forecasting the start of the rainy season is closely dependent on the sharpness of the start itself, resulting in small prediction errors for very sharp onsets. Consequently, we investigate the relationship between large-scale atmospheric patterns and the onset's sharpness, which we aim to evaluate for the reason illustrated previously. Cumulative wind quantities contain information about the wind pattern of the season, up to the day when they are computed (in our analysis, the computation always starts from the 1st January of each year and ends at the onset date). Despite not being proper physical quantities, they are useful indicators as they include a "memory" of the seasonal atmospheric development during the months preceding the onset. We assume that the sharpness of the ORS is not the result of a peculiar wind configuration on the day of the onset itself. On the opposite, it may be a feature that has been building up during the dry season preceding the start of the rains. These analyses have been carried out with the different wind components listed in the section data (Section I). In the following section, we include the results regarding the meridional and zonal wind components at 925 hPa, identified as the fields of primary importance for the above-mentioned relations.

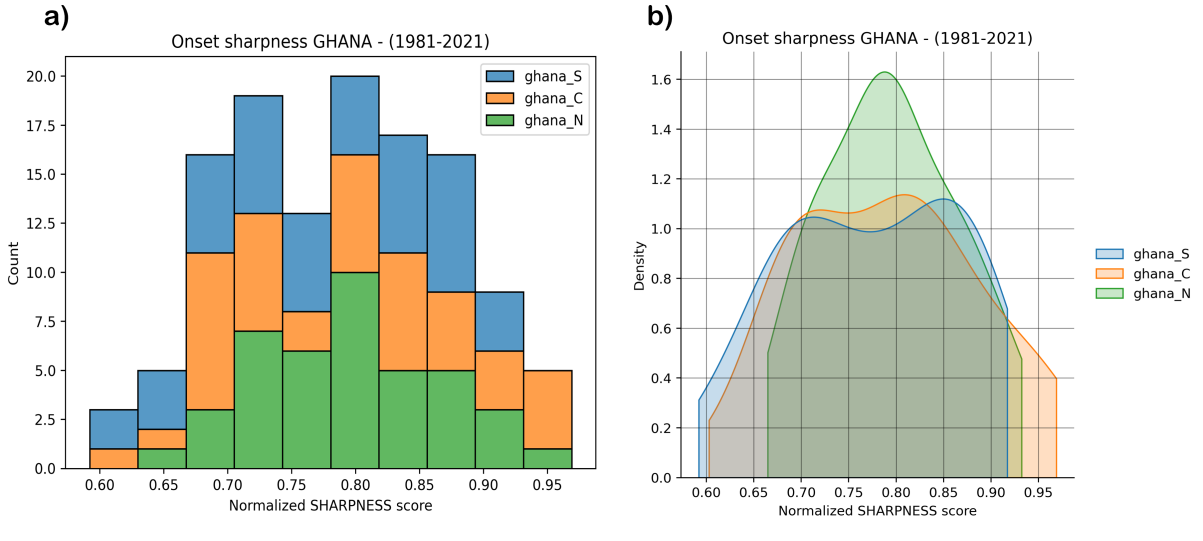

Figure 3.10: Distribution of the SHPindex for the three Ghana's sub-regions.

3.5.2 Results

Figure 3.10 illustrates the distribution of SHPindex, which provides valuable insight about the rainy season's behavior and reveals variations across the sub-regions. Firstly, Figure 3.10**a** reveals that the lowest SHPindex has a value of 0.59, registered in the `Ghana South` region during 2004. In other words, of the precipitation falling in the month preceding and in the month following the onset of 2004, 59% fell in the one succeeding the start of the season. This is an important indication that the dates that we marked as the ORS are indeed showing a clear transition between the dry and the wet season. The PCA-based method used to identify such historical onset dates was extensively described in Section 2.3.2. Therefore, even in a season with a very smooth onset as 2004, the PCA-based method was able to define an onset when the majority of precipitation is recorded after that date.

There is a large variability of the onset's sharpness, with scores spanning up to 0.97 and the most common values lying between 0.7 and 0.85. In addition, the SHPindex distribution of each specific sub-region (Figure 3.10**b**) shows some remarkable differences: `Ghana North` is characterized by a highly peaked distribution, while the other two regions are showing a much larger range of values and the absence of a unique prominent mode. In the second part of this thesis, dealing with the prediction of the onset (Part II), those differences highly impact the predictability of the onset of the rainy season. As an example, the ORS predictions of `Ghana North` with a Random Forest model have, on average, higher skills than the ones for the south and center of the Country (Section 5.3.2).

Having discussed the characteristics of the SHPindex's distribution, we describe the link between a sharp onset and the wind's circulation, pointing out the differences between the north and the center-south of the Country. Firstly, in `Ghana North` we detect a prominent accumulation of very sharp onsets in the first part of the rainy season (Figure 3.12): about 87.5% of the onsets with SHPindex higher than 0.85 happen before the 110th day of the year (20th April). This is considered a slightly early onset's timing since, according to the climate, the mean onset date for `Ghana North` is the 113th day of the year. Such a pattern is not visible in the other two sub-regions, even though a similar weak tendency is present. The second difference concerns the correlation between the sharpness of the onset and the atmospheric circulation at 925 hPa (Figure 3.11), which is relatively strong in `Ghana Center` and `Ghana South`, but appears almost absent in the north. Sharp onsets are observed to be negatively correlated with the cumulative zonal and meridional wind at 925 hPa, with correlation coefficients spanning between -0.58 and -0.63. Other wind fields, such as the vertical wind speed at 500 hPa, show some correlation but without reaching a statistical significance higher than 50%. For this reason, they have not been included in further analyses.

In the center and south of Ghana, the strong correlation between the cumulative winds and the SHPindex suggests that the onset's sharpness is influenced by the atmospheric circulation observed in a time range spanning between 3 weeks and 2 months before the onset itself. A physical interpretation of the phenomenon could follow from the fact that the shift from dry to wet regimes is always matched with a change in the direction of the wind, from positive to negative, for both the zonal and meridional wind at 925 hPa. Therefore, we can associate negative (south-ward or west-ward) wind components with suppressed precipitation and positive (north-ward or east-ward) wind components with enhanced precipitation. At the time of the onset, the more negative the cumulative wind is, the less precipitation is expected in the preceding days. This increases the chance of observing a sharp onset. On the contrary, if the cumulative wind is less negative, it is likely that some precipitation already took place before the observed start of the wet season, leading to smooth onset and to a low SHPindex. This can explain the negative correlation observed between the SHPindex and the cumulative winds.

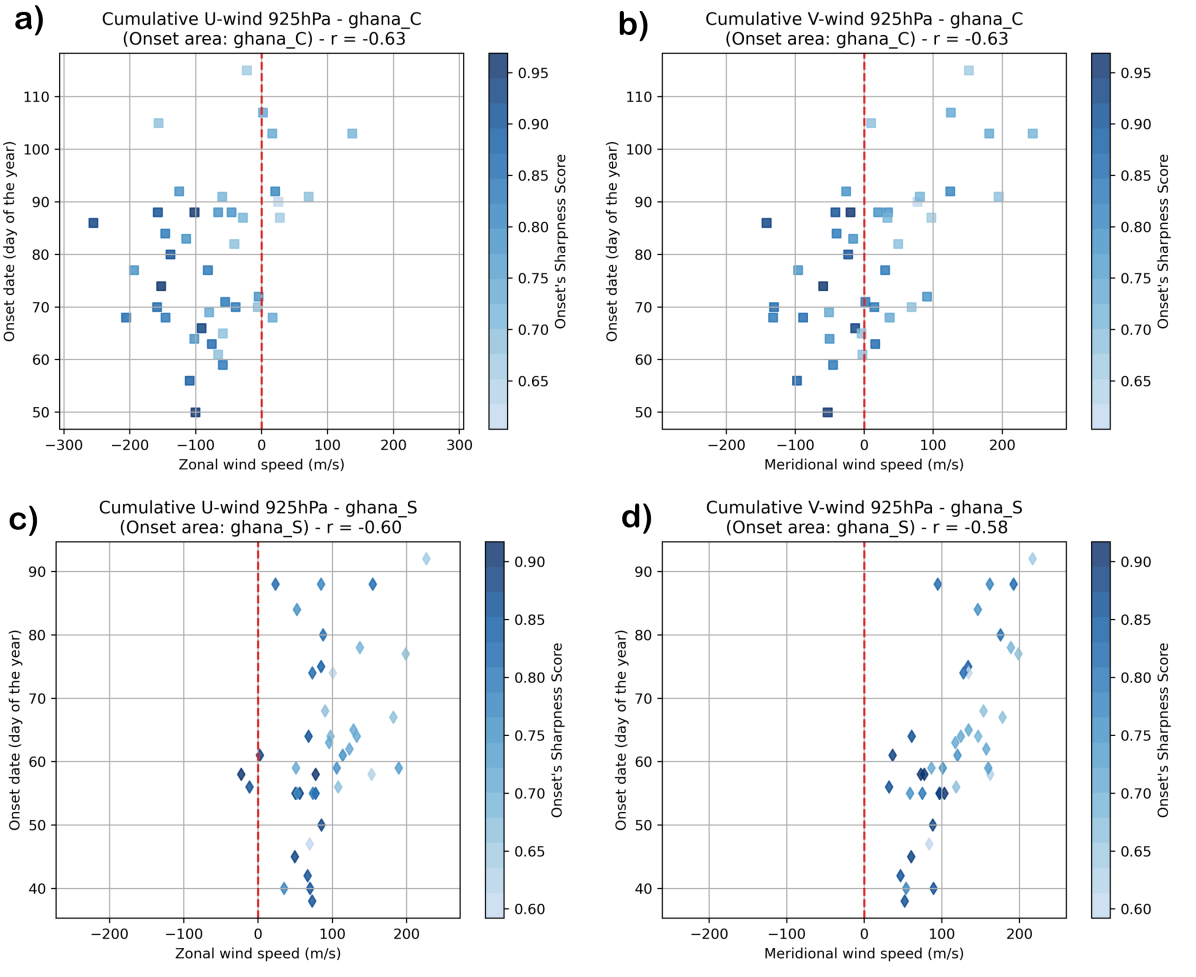

Figure 3.11: Correlation between onset timing (y-axis), cumulative wind at 925 hPa (x-axis) and onset sharpness score (colour bar). Figures a) and b) refers to `Ghana Center`, c) and d) to `Ghana South`. The Pearson coefficient **r** refers to the correlation between cumulative wind and SHPindex, which is quite significant in both regions for both wind components.

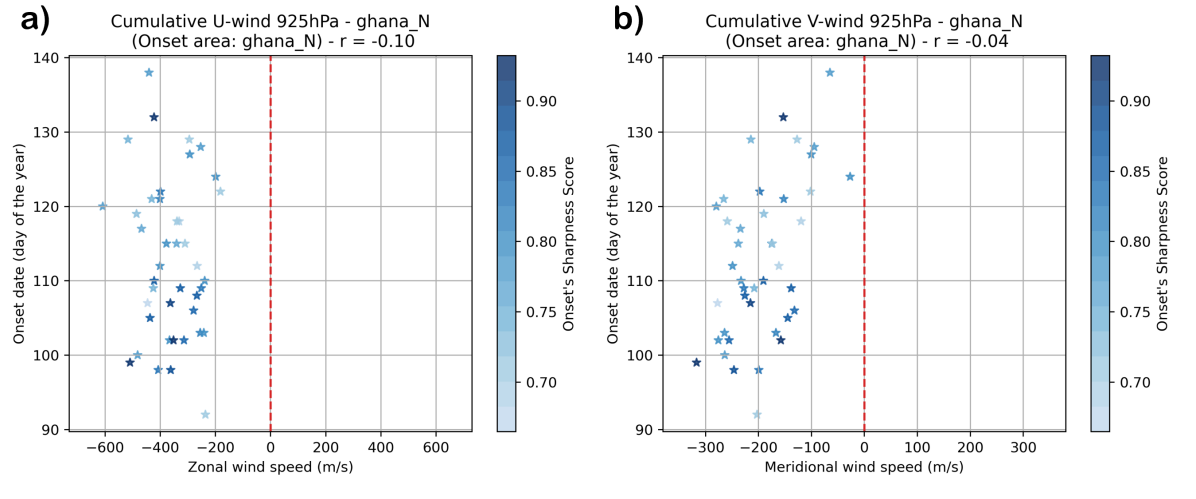

Figure 3.12: Correlation between onset timing (y-axis), cumulative wind at 925 hPa (x-axis) and onset sharpness score (colour bar) for `Ghana North`. The Pearson coefficient **r** refers to the correlation between cumulative wind and SHPindex. For `Ghana North` a trend between early onset dates and SHPindex is visible comparing the y-axis values with the color magnitude.

Part II

The Forecasting Algorithms

Here we address the challenge of operationally predicting the onset of the rainy season. As a consequence, the reader will recognise an ubiquitous effort to reproduce the forecasting conditions as close as the data's availability allows. In the first part (Chapter 4), after presenting the existing operational algorithms, we describe a threshold-based forecast of the regional onset. The latter forecasting routine is analysed and validated through the comparison between the obtained predictions and the observed onset, on both a regional and a local scale. The second part (Chapter 5) describes a completely different forecasting method, based on the supervised learning technique *Random Forest*. This method revealed to be unsuitable for operational applications, yet it showed promising outcomes when applied to atmospheric reanalyses data.

Preliminary Remarks

We want to clarify that, within this project, "ORS prediction" is meant as the detection of the onset from operational forecast of rainfall and wind speed issued by Numerical Weather Prediction (NWP) centres. The algorithms presented in the following chapters fall therefore under the category of Model Output Statistics (MOS) practises. The latter consist of statistical techniques to post-process inputs (usually forecast from a NWP model) and to extract complex but more societally useful information, such as the timing of the ORS. The lead-time is the period of time between the moment a forecast is made and the predicted event. In our ORS predictions the lead-time depends on the NWP forecast's time-steps which are provided as input. As we use indicators starting from 5 days ahead, the same time window can be assumed as the lead-time of our forecasts.

Data

This data section is important to understand the peculiarity of the following work, which aims to simulate with maximum fidelity the operational settings. In the following chapters multiple forecasting algorithms will be applied to produce both **reanalysis forecast** (Section 5.3) and **simulated forecasts** (**hindcast**) (Sections 4.3, 4.5 and 5.4). This differentiated terminology is introduced to indicate the different datasets from which the predictions are generated.

Reanalysis Forecast

With reanalysis forecast we indicate ORS predictions produced with:

- CHIRPS daily precipitation data (daily accumulation, 0.25° resolution, 1981-2021) [21];
- ERA5 reanalysis of meridional wind speed at 850 hPa (daily average, 0.25° resolution, 1981-2021) [23];
- ERA5 reanalysis of zonal wind speed at 200 hPa (daily average, 0.25° resolution, 1981-2021) [23].

The daily mean of wind speed is obtained averaging the original hourly data. Further information about the usage of those dataset in the forecasting model is provided in Section 5.3.1. Information about CHIRPS and ERA5 datasets are provided in the data section of Part I.

Simulated Forecast (Hindcast)

Simulated forecast or hindcast is referred to ORS predictions generated from operational weather forecast issued by the European Center for Medium-Range Weather Forecasts (ECMWF). Both medium-range (lead time up to 15 days) and sub-seasonal (lead time up to 60 days) forecasts were used. The collection of medium-range operational forecasts has been made available for scientific research through the TIGGE dataset, a product of the THORPEX project [6]. TIGGE contains ensemble forecasts of 13 different global Numerical Weather Prediction (NWP) center. In this research we limit to use the forecast produced by the ECMWF, available from October 2006. However, due to a major update in 2016 when the number of ensemble members and the resolution was changed, we consider only data from 2016 onward. This is because both the described threshold-based and the RF-based algorithm are extremely sensitive to the dataset used during calibration or training. The medium-range forecasts provided as input to the ORS forecasting algorithms are:

- ECMWF total precipitation, perturbed forecast (50 members, 0.2° resolution, 2016-2021);
- ECMWF total precipitation, high resolution forecast (1 member, 0.1° resolution);
- ECMWF meridional wind at 850 hPa, perturbed forecast (50 members, 0.2° resolution, 2016-2021);
- ECMWF meridional wind at 850 hPa, high resolution forecast (1 member, 0.1° resolution);
- ECMWF zonal wind at 200 hPa, perturbed forecast (50 members, 0.2° resolution, 2016-2021);
- ECMWF zonal wind at 200 hPa, high resolution forecast (1 member, 0.1° resolution).

The high resolution forecast is blended with the perturbed one, treating the first as an additional member to be add to the ensemble of the second.

Shaped on the TIGGE protocol dataset, the seasonal to sub-seasonal (S2S) forecasts collection has been archived by a joint initiative of the *World Weather Research Programme* (WWRP) and the *World Climate Research Programme* (WCRP) [43]. It delivers to the scientific community a dataset of ensemble forecasts from 11 ensemble prediction models, with a lead-time up to 64 days. Again, we consider the forecast produced by the ECMWF only. The ECMWF issues operational S2S forecast twice a week (Monday and Thursday), with a resolution of 0.4°.

The sub-seasonal forecast provided as input to the ORS forecasting algorithms is:

- ECMWF total precipitation, perturbed forecast (50 members, 0.4° resolution, 2016-2021);

The historical collection of ECMWF operational forecasts (both medium-range and S2S) contains some missing dates. However, the number of missing data-points is a minimum fraction of the analysed time range, and this issue never repeats for more than three consecutive days. We conclude that this lack of data has a minimum impact on the performance of the ORS forecasts. We tackled the problem by not producing any predictions for the days when a NWP forecast was not issued.

Post-processing

The post-processing of the raw data involved several steps, addressing the following issues:

- The ECMWF rainfall forecasts are delivered, at each time-step, as the total accumulated precipitation from the start of the forecast. All the data have been post-processed to extract the predicted daily amount of rainfall.
- The precipitation forecast were converted from kg/m^2 to mm.
- The validation of the ORS is performed against the historical local onsets (obtained in Section 2.2.2), which consist of a gridded dataset with 0.25°of resolution. Consequently, we decided to produce forecast on the same grid, meaning that the medium-range predictions have been down-scaled while the sub-seasonal ones have been up-scaled to meet the observation's resolution.
- The S2S are delivered with daily time-steps, but are issued only twice per week. In order to run the ORS forecasting algorithm daily, with both medium-range and sub-seasonal forecast as input, the dataset was modify to always use, after lead-time of 15 days, the newest version of the S2S forecast. This means that, cyclically, we use for time-steps over 15 days ahead data issued by the ECMWF up to three days before.

Chapter 4

Threshold approach

In Section 1.3 two families of onset's definition were introduced, namely local and regional ones. The predictive algorithms can be developed following these two perspectives, but algorithms based on local precipitation events are by far the most used. The majority of forecasting routines rely on a set of conditions represented by thresholds on different rainfall indicators, such as the total amount of precipitation in a certain time range or the number of consecutive wet days (1 mm or above).

In this chapter we present the development and validation of forecasting algorithms for both regional (Section 4.2) and local (Section 4.5) ORS using threshold-based methods. Hindcast for the period 2016-2021 are produced and validated in comparison with the historical local ORS (obtained in Section 2.2.2).

4.1 Existing Onset Forecasting Algorithms

Despite the extensive scientific literature about the definitions and the predictability of the rainy season, the number of publications dealing with the operational forecast of the rainy season's onset is rather limited. Rauch et al. [36] describes a threshold-based method to predict local onsets over Ghana and Burkina Faso, using 11 years (2000-2010) of ECMWF SYS4 precipitation hindcasts. The algorithm implements the ORS definition introduced by Stern, Dennett, and Garbutt [38], which sets the onset on the first day of the year when three conditions are met simultaneously. The conditions relay on precipitation's forecast only and check for the total amount of rainfall, its time distribution and the absence of dry spells in the near future. As any other threshold-based ORS definition, this suffers from a binary logic. For example, if the condition reads that the total amount of rain accumulated in the next 25 days must be at least 25 mm, an accumulation of 24.9 mm would not be identified as the ORS. To avoid such situations, Laux, Kunstmann, and Bárdossy [29] introduced a fuzzy logic approach, which smooths the conditions required for the transition from the dry to the wet season.

As mentioned, the prediction of the ORS on local scale is the approach most frequently used and explored, while few attempts were made to predict the onset on a regional scale. Laux, Kunstmann, and Bárdossy [29] implemented two complementary techniques to generate a categorical forecast of the ORS over Burkina Faso and Ghana. The "regionalisation" was achieved through a PCA in spatial mode, adopting a procedure similar to the one we applied in Section 2.3.1. The first technique consists of a Linear Discriminant Analysis (LDA) using *rainfall amount* and *number of wet days* up to 30 days before the potential onset's date. The LDA places each day in one of four categories: "dry season", "transition", "onset of the rainy season" and "wet season". The second technique is a Linear Regression Analysis (LRA) which exploits the information of the timing of the onset of regions with earlier ORS. According to the authors, while the LDA method can only assess day by day if the rainy season has already started or not, the LRA have the potential to predict the ORS a few weeks head. On the other hand, since the method is based on rain gauges measurements, quality control of the data is essential for the performance of the algorithm, which is heavily sensitive to outliers.

4.2 Building a Threshold-based Regional Onset Forecast

The development of the algorithm to forecast the regional ORS was driven by two requirements:

1. the predictions must fulfil fixed conditions on rainfall's amount and distribution;

2. the predictions must resemble the observed regional ORS, as close as the above-mentioned fixed conditions allow.

The first point led to the development of the threshold-based algorithm, which is described in full details in this Section 4.2.1. The second requirement is met tuning the thresholds of the forecasting routine and choosing the combination that best reproduces the observed regional onset. The tuning was performed both with and without a minimizing algorithm (a dual annealing method). Due to the algorithm's high sensitivity to varying thresholds, the second option proved to be the optimal one. It allows to track the algorithm's response to different threshold's combinations, maintaining thresholds with a clear physical interpretation.

4.2.1 Algorithm's Structure I: Thresholds-based Core

The fixed pre-requisites to assess the ORS are directly translated into the thresholds of the algorithm. Each day of the forecast, the following indicators are generated for every member of the NWP ensemble forecast:

Indicator	Description
PR10	Total amount of precipitation predicted in the **next 10 days**
PR25	Total amount of precipitation predicted in the **next 25 days**
PWdiff10	Number of wet days ($>$1 mm) **between day 5th and 10th** from the forecast date
PWdiff15	Number of wet days ($>$1 mm) **between day 10th and 15th** from the forecast date
PWdiff20	Number of wet days ($>$1 mm) **between day 15th and 20th** from the forecast date
PWdiff25	Number of wet days ($>$1 mm) **between day 20th and 25th** from the forecast date

Table 4.1: Indicators used in the threshold-based ORS detection. Each quantity is intended to be the spatial average over a specific sub-region.

Figure 4.1 displays the structure of the predicting algorithm, which can be seen as the aggregation of three components. **1)** The first one (light blue diamond in the flow chart) is a condition checking that the sub-region receives a minimum amount of rainfall in a 10 and 25 days time window; **2)** the second component (light green diamonds) is ensuring that no dry periods of more than 10 consecutive days are registered in the 25 days following the candidate onset dates; **3)** lastly, a counter keeps track of how many consecutive days do fulfil the former two conditions and, if that number reaches a fixed thresholds (light red diamond), the onset is detected. The last component is important as it prevents drawing a decision based on one ECMWF forecast only, which could indicate a wrong trajectory of the weather. Conversely, if the trend is confirmed by multiple forecast consecutively, the confidence in the predictions is high enough to assert the ORS. The flow chart contains various thresholds, which are marked in bold text and described in Table 4.2.

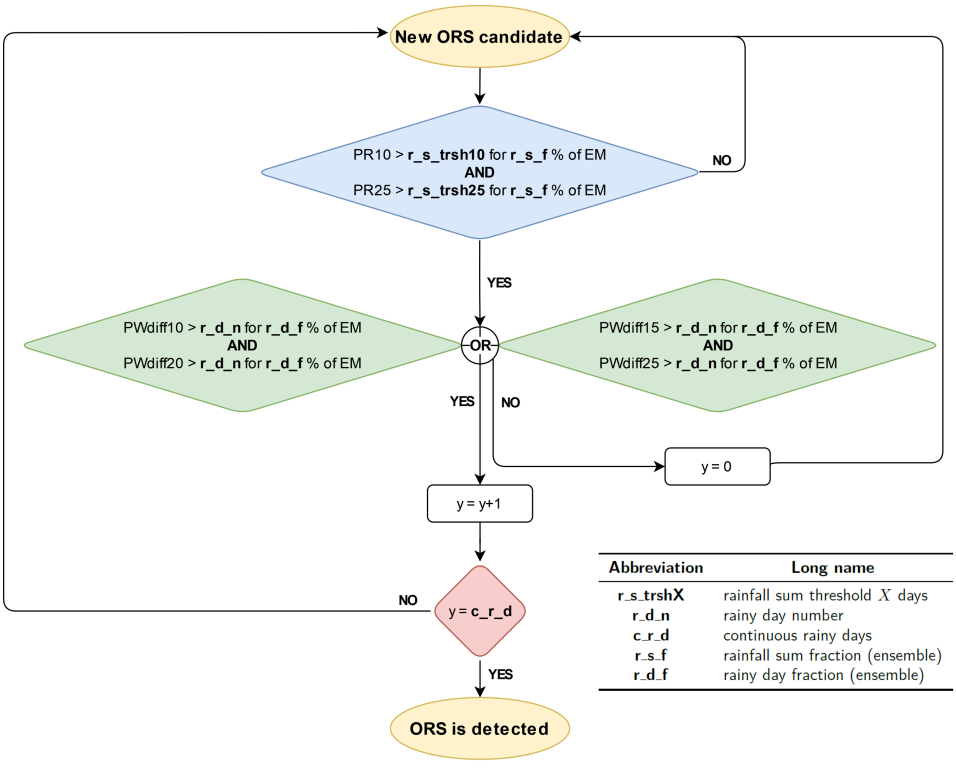

Figure 4.1: Flow chart of the regional ORS forecasting routine. The sign in bold, containing underscores, are fixed thresholds which have been calibrated for each sub-region. EM is an abbreviation for "Ensemble Members". This chart must be seen as a decision map which, given a candidate date, returns whether that date it is to be considered the actual ORS.

Abbreviation	Long name	Description	Unit
r_s_trsh10	rainfall sum threshold 10 days	Minimum cumulative precipitation in the next 10 days	mm
r_s_trsh25	rainfall sum threshold 25 days	Minimum cumulative precipitation in the next 25 days	mm
r_d_n	rainy day number	Minimum number of rainy day within a 5 days window	day
c_r_d	continuous rainy days	Number of consecutive days which do respect all the threshold's conditions	day
r_s_f	rainfall sum fraction (ensemble)	Minimum fraction of ensemble member fulfilling condition on rainfall sum	-
r_d_f	rainy day fraction (ensemble)	Minimum fraction of ensemble member fulfilling condition on rainy day	-

Table 4.2: Description of the thresholds used in the forecasting algorithm of the regional ORS.

The thresholds **r_s_trsh10** and **r_s_trsh25** represent amounts of precipitation, which can strongly vary for different locations. They are computed, for each grid cell, as follow:

$$\textbf{r\_s\_trshX} = 0.9 * raintype * X * \textbf{rain\_trsh\_fract}, \tag{4.1}$$

where X is the number of days, **rain_trsh_fract** is a thresholds representing the minimum number of rainy days in the forecast, and 0.9 is a scaling factor. To conclude, $raintype$ is the daily precipitation threshold for

a given grid-cell. It is computed for four categories, each of them representing a different percentile of the annual rainfall climate in that specific location ("Low" = dry day percentile, "Medium" = 40%, "High" = 75%, "Very High" = 90%). To take into account possible biases in the rainfall prediction, such quantities are computed on the daily rainfall average of TIGGE data, the same used in the forecast. Each of the percentile is computed excluding dry days (< 2 mm), so that it is representative of the wet season. The regional daily rainfall threshold is obtained averaging the cell-specific $raintype$ dataset over the sub-region area.

4.2.2 Algorithm's Structure II: Constraints

The algorithm's structure, as presented in Figure 4.1, is the core of the forecast, which is responsible for the "active" detection of the ORS. In addition to this framework, we incorporated other "passive" conditions that do not directly contribute to detect the onset. Instead, they constraint the range of possible dates fed into the previously described algorithm. These constraints are therefore crucial to avoid too early or too late ORS predictions.

For every candidate onset date, the following indicators are computed (ensemble member-wise):

Indicator	Description
PW10	Number of wet days (>1 mm) predicted in the **next 10 days**
PW20	Number of wet days (>1 mm) predicted in the **next 20 days**
V850F	Meridional wind speed at 850 hPa at the **10th day following** the candidate onset date*

Table 4.3: Indicators used to constraint ORS detection. Each quantity is intended to be the spatial average over a specific sub-region.
*The meridional wind speed is the 5-days mean of the speed forecast between 5 and 10 days ahead.

One constraint acts as a "no-start" condition, meaning that the ORS cannot be set if:

- V850F $>$ **V850min** in **w_b_f** % of EM* for **c_WL_d** consecutive days

Other two conditions act instead as "forced-start" ones, meaning that the ORS is automatically set if one of the two is met:

- V850F $>$ **V850max** in **w_b_f** % of EM* for **c_WU_d** consecutive days
- (PW10 $>$ **tot_rd10** AND PW20 $>$ **tot_rd20**) in **r_d_f_e** % of EM* for **c_PW_d** consecutive days

* EM is the abbreviation for 'Enseble Members'.

The thresholds marked in bold text are specific for each sub-region and are calibrated on the observed regional ORS. A description of the thresholds is present in Table 4.4.

Abbreviation	Long name	Description	Unit
V850min	v 850 hPa lower thresholds	Minimum meridional wind speed at 850 hPa	m/s
V850max	v 850 hPa upper thresholds	Maximum meridional wind speed at 850 hPa	m/s
w_b_f	wind boundary fraction (ensemble)	Minimum fraction of ensemble member fulfilling conditions on wind speed	-
c_WL_d	continuous windy days (lower boundary)	Number of consecutive days which do respect the wind speed lower condition	day
c_WU_d	continuous windy days (upper boundary)	Number of consecutive days which do respect the wind speed upper condition	day
tot_rdX	total number of rainy day (X days)	Minimum number of rainy days in the next X days	day
r_d_f_e	rainy day fraction (ensemble)	Minimum fraction of ensemble member fulfilling condition on rainy days	-
c_PW_d	continuous rainy day	Number of consecutive days which do respect all conditions on rainy day	day

Table 4.4: Description of the thresholds used as constraint in the forecasting algorithm of the regional ORS.

The reader will notice that two of these constraints are base on wind forecasts. This is the only case where it was possible to include a large-scale parameter into a threshold-based algorithm; after experimenting different wind predictors, the use of those quantities in the "active" algorithm has proven to be very challenging and to lead to a worsening of the predicting skill. The reason of the wind's unsuitability to such scope lies in the high inter-annual variability of wind speed, addressed in Section 3.3.2.

4.3 2016-2021 Hindcast

The newly-built algorithm was fed with ECMWF historical forecasts covering the years 2016-2021, as described in section Data at the beginning of Part II. This generates hindcasts which are closely comparable, if not identical, to actual operational forecasts. Section 4.3.2 contains the validation of such hindcasts.

4.3.1 Tuning of the Thresholds

Once that the structure of the algorithm is built, it is necessary to tune its thresholds. This delicate operation was performed setting as target the observed regional ORS, obtained in Section 2.3.2. An additional guidance was introduced, i.e. we privileged configurations leading to absent or small early forecast errors, even if their overall performances were not the best ones. Penalising the detection of early ORS is driven by practical applications, for which is preferable to maintain a conservative approach towards the ORS assessment, avoiding false starts predictions. The optimal thresholds configuration, used to generate the following results, is reported in Appendix B.1.1.

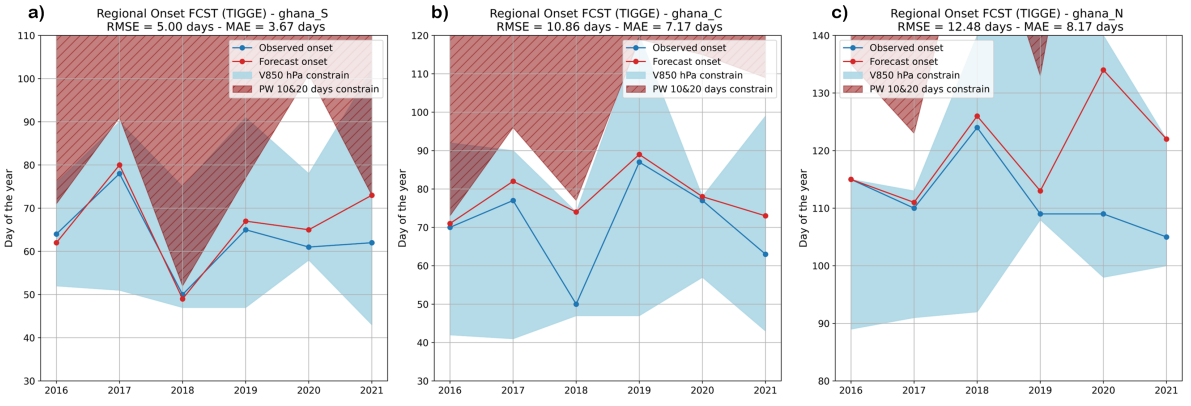

Figure 4.2: Threshold-based regional ORS predictions. The constraints on candidate onset dates, described in section 4.2.2, are indicated as shading of different colors: blue represents the range of dates fulfilling both meridional wind's constraints, red indicates the dates excluded by the constraint on the number of rainy days.

Figure 4.2 shows the regional ORS predictions, together with the observed onsets and the range of candidate onset dates allowed by the constraints introduced in Section 4.2.2. In every sub-region the constraints are at least once responsible for the detection of the ORS. However, the effectiveness of the constraints can vary substantially, suggesting that the degree of correlation between the chosen indicators and the ORS is extremely variable, even among consecutive years and within the same sub-region. It can also be noted that there is no situation where the ORS is detected earlier than the observed one, as consequence of the choice to penalise false start predictions.

Comparing the hindcasts with the regional observed ORS, two groups of years can be distinguished: **1)** the ones when the forecast error is very small or non-existent (e.g. 2016, 2017, and 2018 for `Ghana South`; 2016, 2019, and 2020 for `Ghana Center`; 2016, 2017 and 2018 for `Ghana North`) and **2)** the ones when the forecast error is larger than 5 days. Overall, the skills of the algorithm (measured as prediction-observation RMSE) decreases proceeding north-ward.

However, comparing the forecasts against the regional ORS observations carries little information about the goodness of the forecasting algorithm. This is because the observed regional ORS is not the ultimate target of the forecast, which is instead the observed local ORS (obtained in Section 2.2.2). The regional ORS is only the target of the algorithm's tuning, to ensure that the method is "trained" to forecast what we have identified as the regional onset of the wet season.

The imposition of fixed conditions on the detection of the ORS (such as a minimum accumulation of rain or number of wet days) generates unavoidable discrepancies between the predicted and the target ORS, even choosing the optimal threshold's configuration. Even when they are greater than 10 days, the occurrence of such discrepancies should not lead the reader to automatically judge the forecasts as poor in skills: the latter will be determined (in the next section) based on the comparison with the observed local ORS.

4.3.2 Validation

The validation of the regional ORS forecasts is performed as a comparison with the observed local ORS (Section 2.2.2). This choice is justified by the fact that the local perspective is the one of primary importance for the end-users of the forecasts.

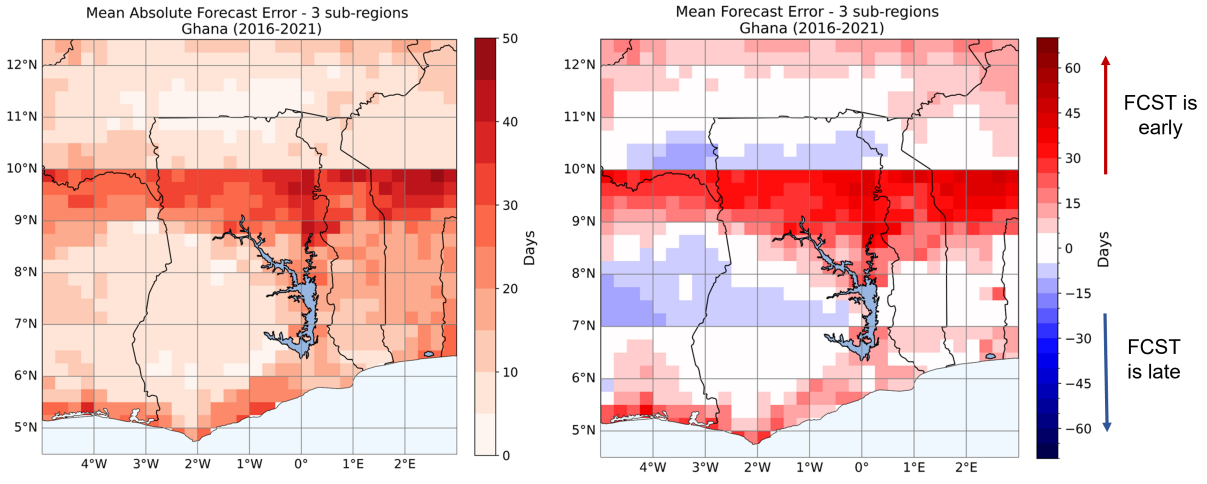

Figure 4.3: Validation of the regional ORS forecast. In each pixel, the shading represents the difference (in days) between the observed local ORS of that location and the predicted regional ORS relative to the sub-region which the pixel belongs to. **Left**: 2016-2021 mean absolute forecast error. **Right**: 2016-2021 mean forecast error. Red shades indicate that the forecast predicted the ORS too early, blue ones that the forecast predicted the ORS too late.

The prediction errors on the six analysed years (Figure 4.3, **right**) indicate that `Ghana Center` is the most problematic region. In particular, its northern border and the area around Lake Volta is where the discrepancies between observations and predictions are the largest, reaching up to 45 days. The situation is particularly concerning as the errors are mostly early predictions, which carry the worst consequences for agricultural planning. The other two regions (`Ghana South`, excluding the coast, and `Ghana North`) benefit of much lower forecast errors and there are large areas where the mean onset difference is limited to less than 5 days (white cells). Looking at the mean prediction errors but in absolute term (Figure 4.3, **left**) allows to grasp a better understanding of the algorithm's skills, since it removes possible error's compensation effects present in the simple mean. However, the situation appears to not differ substantially from the previous case, meaning that little compensation is present and that there is consistent directionality of forecast errors (either early or late) across the analysed year.

4.3.3 Algorithm's Sensitivity to Sub-regions' Definition

Analysing Figure 4.3, the first features that stand out are the artefacts generated at the border of the sub-regions. This is a natural consequence when comparing two datasets of different nature: one containing, for each year, a different data for each grid-cell; the other containing, for each year, only three values representative of all the spatial domain. Focusing on the area [1-5°W, 5-9°N], a region of very high predictive skills in the south is abruptly interrupted by an area with too late forecasts (blues shadings) when crossing the `Ghana South-Ghana Center` border. This led the author to explore the possibilities of changing sub-regions' definition, in order to smooth out such transitions and to increase the area of high skill "hotspots".

A new set of sub-regions (Figure 4.4) was defined according to the observed ORS climatology (Figure 2.4) and to the performance of the ORS forecast applied to the original regional division (Figure 4.3). In the climate, the original `Ghana North` region shows a latitudinal displacement of the ORS, reason for which it was split into two longitudinal bands, `Ghana North-North` and `Ghana South-South`, aiming to better catch the meridional progression of the rains. `Ghana South` and `Ghana Center` were split into a West and and East sub-regions, with the `Ghana South-West` region aiming to cover the homogeneous rainfall pattern observed at [5-1°W, 6-8°N].

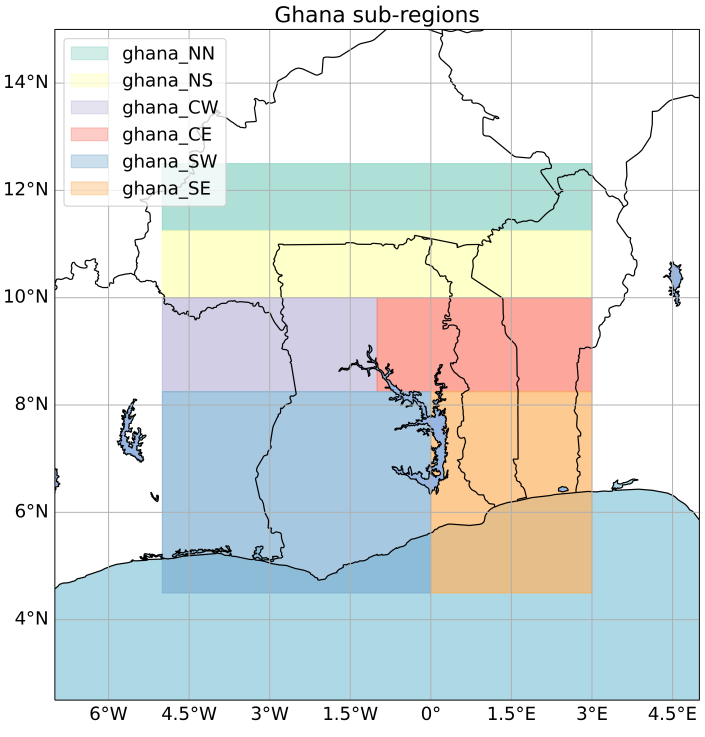

Figure 4.4: Alternative definition of the sub-regions'.

The algorithm to generate the new sub-regions' hindcast and the calibration procedure of its thresholds were identical to what described in Section 4.2. The only difference is that is was not possible to apply the upper constrain on the meridional wind speed at 850 hPa, therefore the "forced-start" condition (described in 4.2.2) is represented by the thresholds of the number of wet days only. The calibration results, both as optimal thresholds values and as comparison with the observed regional ORS (as in Figure 4.2) are attached to the Appendix (Table B.1.1 and Figure B.1). Here we limit to show the results of the validation against the observed local ORS (Figure 4.5).

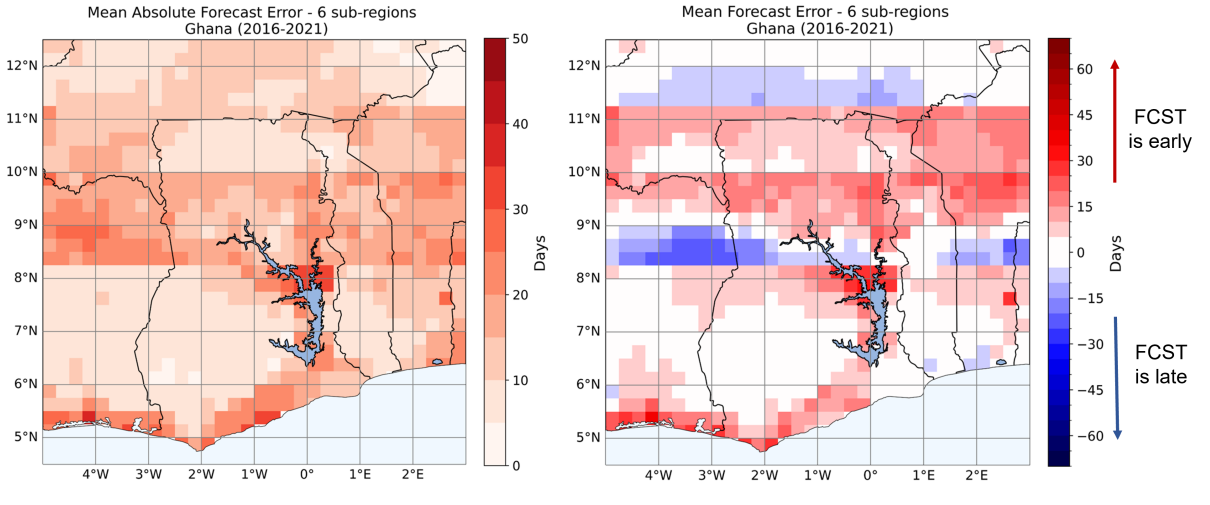

Figure 4.5: Validation of the regional ORS forecast. In each pixel, the shading represents the difference (in days) between the observed ORS for that grid cell and the predicted ORS relative to the sub-regions which the cell belongs to. **Left**: 2016-2021 mean absolute onset difference. **Right**: 2016-2021 mean prediction error. Red shades indicate that the forecast predicted the ORS too early, blue ones that the forecast predicted the ORS too late.

The 6 sub-regions forecast (6-SRF) has, overall, lower forecast errors if compared with the 3 sub-regions equivalent (3-SRF). The maximum absolute discrepancy between predictions and observation is lowered from 45 days for 3-SRF to 35 days for 6-SFR. Moreover, the fraction of grid cells experiencing mean absolute

forecast error larger than 20 days drops from 21% for 3-SRF to 12.5% for 6-SRF. However, while the improvements are well tangible in the areas where the 3-SRF forecast's skills were very poor, little progress or even deterioration of the forecast performance is seen in the high skills "hotspot". As an example, the prediction in `Ghana South-West` region appears to be less biased towards late forecast error adopting the 6-SRF (comparing Figure 4.3 and 4.5 right plots); however, taking into account the absolute forecast error (Figure 4.3 and 4.5 left plots) we come to the opposite conclusion: the 3-SRF shows a larger area where the absolute discrepancies are lower than 10 days. This suggests that 6-SRF are less biased towards a early or late prediction errors, which produces compensation effects in the mean forecast error, but the latter is not lower than the 3-SRF case in absolute terms. In `Ghana North` is even more clear that the new sub-regions division does not improve the forecast skills.

4.3.4 Results

The validation's results shown in Figures 4.3 and 4.5 refer to mean quantities, which outline the overall performance of the forecasting algorithm during the analysed years. However, in operational settings extreme caution is required before drawing conclusions from mean quantities. Ultimately, the local and time-punctual information is the only that really matters for the forecast's end-users, who will never experience the mean conditions. For this reason, the ORS prediction's performance has been analysed individually for each years between 2016-2021, comparing the 3 and 6 sub-regions division. The individual year forecast's error maps are attached to the Appendix (Figures B.2 and B.3).

The forecast's performances at a fixed location can vary considerably among the single years. In most cases the forecast error's direction (early or late prediction) is constant in the time series, but the magnitude of the discrepancies can fluctuate from few days up to a month, even between consecutive years. However, there are areas where the forecast errors are relatively small (less than 10 days) during all the time series, while others suffers from huge offsets systematically. Such pattern was already highlighted by the time-averaged forecast errors (Figures 4.3 and 4.5), but the single-year analyses return a picture where those areas of high and low forecast's skills are surprisingly well-defined.

The final assessment on the performance of the threshold-based algorithm is built computing, for every grid-cell, how many years of the analysed ones (2016-2021) show an absolute forecast error lower or equal to 10 days (Figure 4.7). This threshold was chosen assuming that the end-users' benefit from a inaccurate forecast can outweigh its damages if the forecast's error stays within 10 days.

The result is that there are some hotspots, notably in `Ghana South-West` and in `Ghana North`, where the forecast meets the defined "quality standard" for all the six years of the analysis. Figure 4.6 summarized the performance of the algorithm is the two different sub-regions division. The conclusion is that the 3-SRF has the highest forecast's skill in terms of fraction of grid-cells meeting the above-mentioned conditions for 5 or 6 years, but it also contains the largest disparity between high- and poor-performing regions. 6-SRF has high skill's hotspot of smaller spatial extension, but it shows better performances in the low-skill areas, such as north and east of Lake Volta.

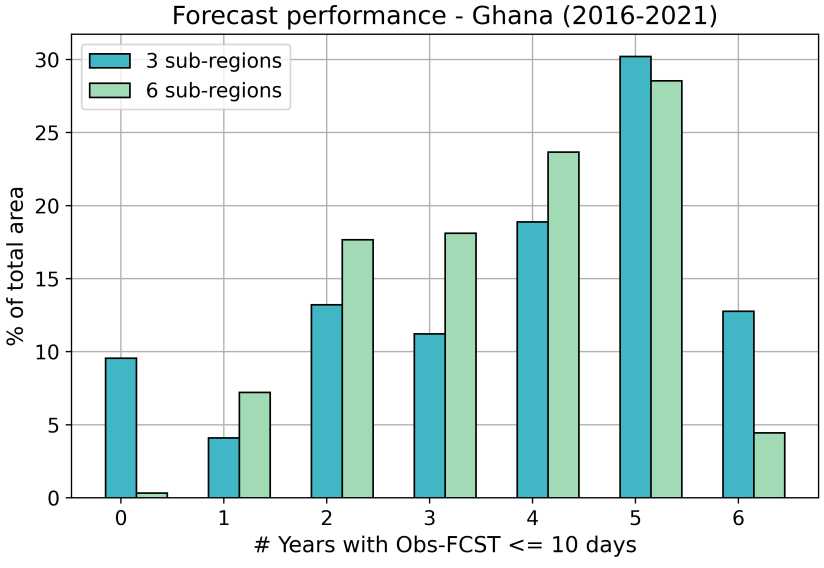

Figure 4.6: Fraction of the spatial domain falling in different performance's categories. The latter are defined by the number of years (between 2016-2021) when the absolute forecast error is equal or lower than 10 days, as shown in Figure 4.7.

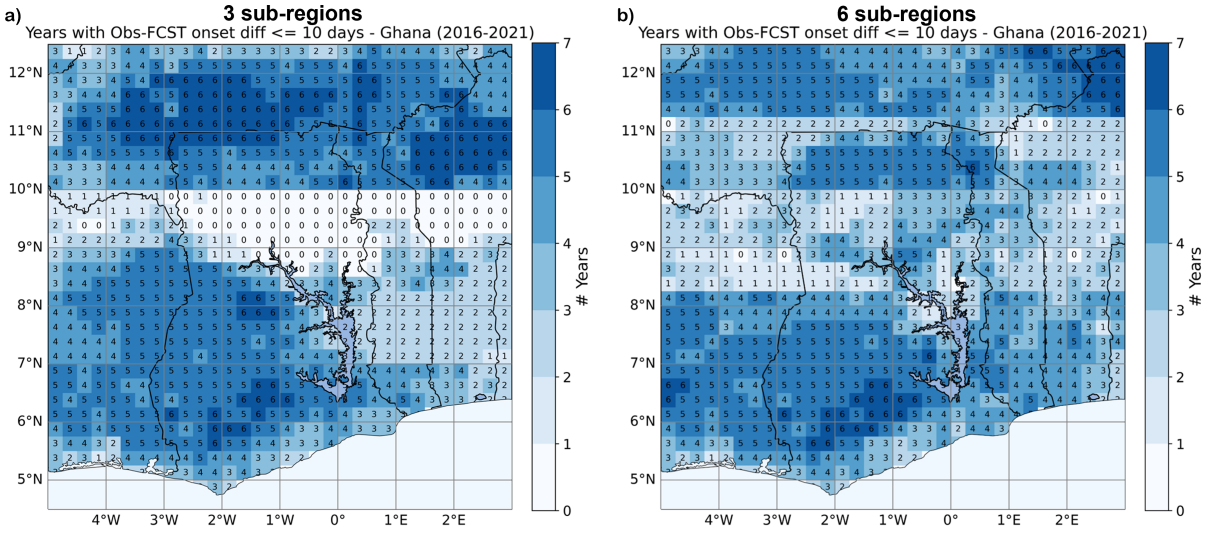

Figure 4.7: Number of years when the difference between the observed local ORS and the regional ORS forecast is equal or lower than 10 days. The plots refer to both the 3 sub-regions experiment (**left**) and the 6 sub-regions one (**right**). The small figures in each grid-cell report the exact number of years meeting the above-mentioned condition. It is remarkable that, in the 3 sub-regions case, the border between `Ghana South` and `Ghana Center` is indistinguishable.

The previous analysis leads to conclude that the threshold-based algorithm displays performances that can justify its operational deployment, but limited to specific areas of the spatial domain. The division of the domain in 3 sub-regions yields the best performances and the larger extension of such regions, which include:

- south-west Ghana, [4-1°W, 5.5-9°N]
- north Ghana-Burkina Faso border, [4-0°W, 10.5-12°N]
- north Benin, [1-3°E, 10-12°N]

On the other hand, the 6 sub-regions division yields a performance's improvement in areas outside the ones listed above. However, such improvements are not sufficient to justify the operational deployment of the algorithm in those areas.

4.4 Discussion: PCA Influence on the Regional Forecast

The performance of the regional ORS forecast opens two questions: why the skills of the algorithm can differ so much within the same sub-region? Is there a margin for improvements in the areas currently experiencing low performances? The answer to both questions is found retracing the entire process leading to the regional predictions, which origins from the initial PCA performed on the precipitation time records, described in Section 2.3.2. For the sake of clarity and conciseness, the following considerations will be focused on understanding the high performance hotspot of the south-west Ghana region (hotspot GSW), between [4-1°W, 5.5-9°N]. We will consider the 3 sub-regions division only, in specific `Ghana South` and `Ghana Center`.

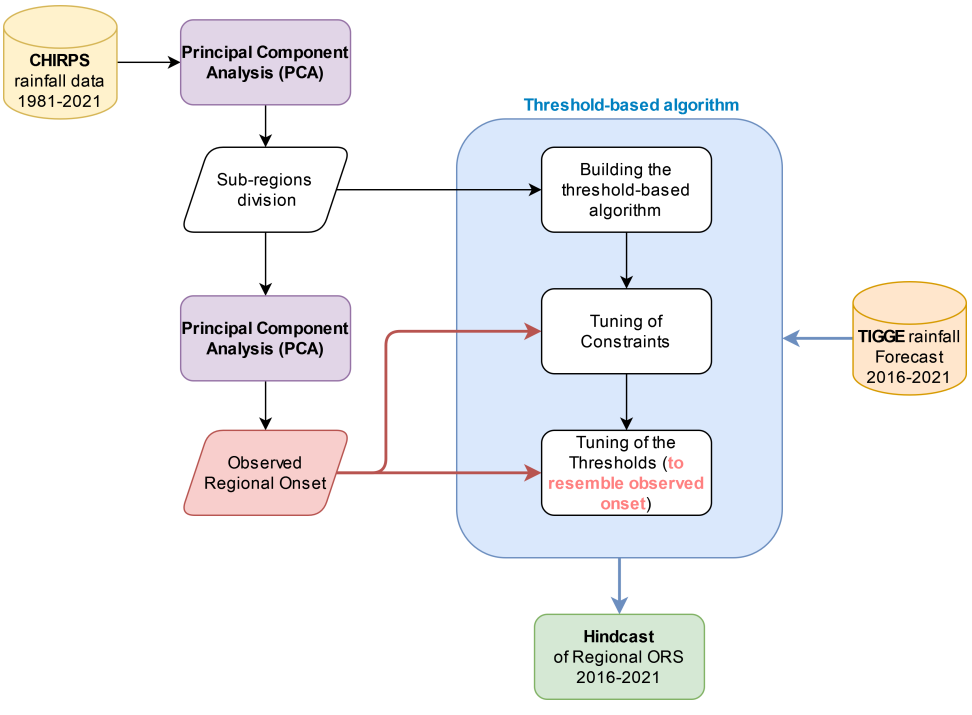

Figure 4.8: Flow chart of the entire process leading to the regional ORS predictions, from the definition of the sub-regions and the determination of the observed regional ORS, to the building of the forecasting algorithm and the tuning of its thresholds.

As shown by the flow chart of Figure 4.8, the hindcast of regional ORS originates from a first PCA performed on the entire spatial domain to defined the sub-regions (Section 2.3.1). A second PCA (restrict to a single sub-region) follows to determine the observed regional ORS (Section 2.3.2). The structure of the regional ORS algorithm emulate a decision diagram based on fixed conditions and does not differ from what would be done to develop a local ORS prediction routine. The core of the regional approach lies instead in the following threshold's tuning procedure (Section 4.3.1), which is finalized to generate forecast resembling the observed regional ORS.

To reply our initial question it is necessary to understand what information is carried by the observed regional ORS. Analysing the Camberlin and Diop [11] method, used to identified the observed onset (Section 2.3.2), it follows that only the information belonging to the fist Empirical Orthogonal Function (EOF1) is transferred to the detected regional ORS. It means that the observed regional ORS represents only the strongest mode of oscillation, which usually can be found in a specific sub-region of the PCA domain. By definition, such region will display a precipitation pattern characterised by strong spatial coherence, which makes the local mode of oscillation to be dominant over the others. Being tuned to reproduce the observed ORS, the forecast will be truly representative only of the regions which the EOF1 belongs to. To conclude, the forecast skill can differ quite strongly within the same sub-region because the algorithm is not calibrated to be representative of every location enclosed by the sub-region itself.

To provide a concrete example of what described above, we demonstrate the influence of the initial PCA on the forecast's performance of the GSW region. There are two clues suggesting that this region is strongly

represented by the EOF1 of both `Ghana South` and `Ghana Center`. Firstly, an analysis on the rainfall's spatial correlation reveals a significant coherence in the area between 4-1°W and 5.5-9°N (Figure 4.9, **left**). This spatial coherence is a necessary condition to find strong EOF. The second hint has been already shown in Section 2.3.2, when we analysed the precipitation records few days before and after the observed onset. The initial rainfall appears to be focused in a well-defined area, corresponding closely with the hotspot GSW. This demonstrates that the EOF1 is representative of the precipitation pattern of the GSW area, while it carries little or no information for other areas within the same sub-region, such as the one around Lake Volta. We can now fully explain the different regions of high and poor forecast skills of Figure 4.7 (3 sub-regions division). Moreover, since the hotspot GSW is not influenced by the `Ghana South-Ghana Center` border, we conclude that the rainfall's common oscillation of GSW drives the dominant EOF in both `Ghana South` and `Ghana Center`.

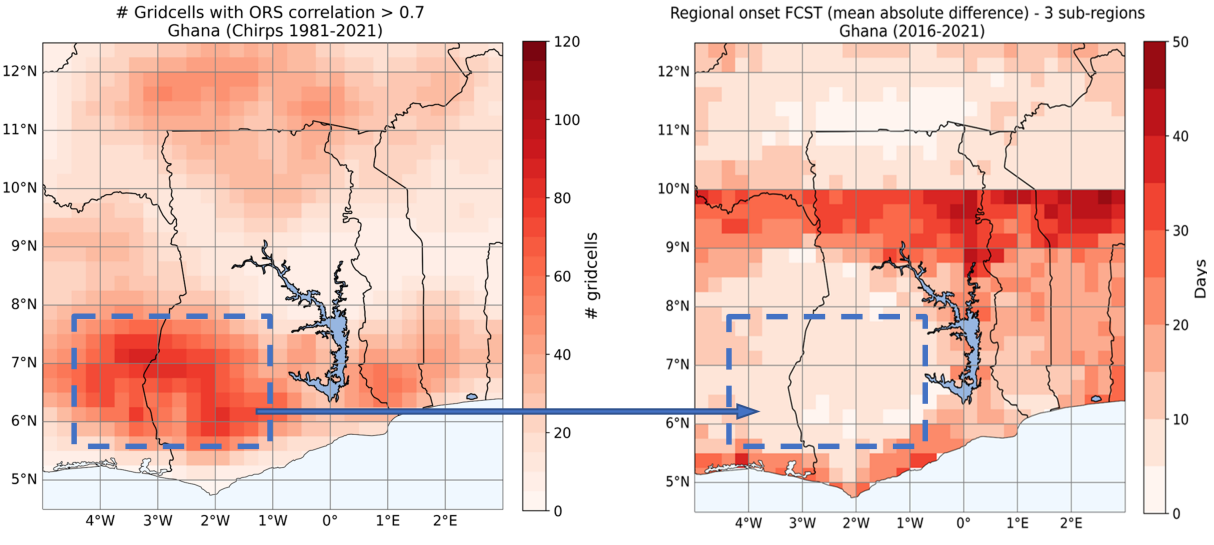

Figure 4.9: Connection between rainfall's spatial coherence and forecast skill. **Left**: given the local ORS observation time series, the map compares the ORS time series of different locations and shows, for each grid-cell, how many other cells have a correlation coefficient higher than 0.7. A high number of correlated cells signifies that there is a common oscillation of the ORS dates, which is most commonly shared by several adjacent locations. The area north and east of Lake Volta is poorly correlated, meaning that the spatial variability is very large. **Right**: mean absolute forecast errors of hindcast performed with the 3 sub-regions division. There is a clear connection between the high spatial correlation of south-west Ghana and the high performance of the forecast in the same area.

Lastly, we answer the question whether it is possible to find a different sub-regions' division representative of areas where the forecast is now poorly performing, as the Lake Volta region. The experiment conducted doubling the sub-regions' number (Section 4.3.3) had precisely this aim, and it demonstrated that only small improvement can be reached. The explanation is to be found in the precipitation's spatial correlation shown in Figure 4.9: by definition, areas where the ORS varies independently between adjacent grid-cells cannot return strong EOFs, making impossible to create a forecast representative of those regions. This leads to conclude that forecasting the ORS regionally is possible only in regions with high spatial coherence of rainfall, and little can be done for areas which do not show such quality.

4.5 Threshold-based Local Onset Forecast

This section reports few experiments in the direction of predicting the ORS with a local perspective. As the latter has not been the focus of this thesis, the following work must be intended as an exploratory research on further applications of the algorithm developed in Section 4.2.

The building of a threshold-based forecast of the regional ORS brought to the development of an algorithm which can be easily applied to the local ORS detection. As discussed previously, the essence of the regional prediction does not lie in the algorithm's structure itself, but rather in the tuning of the thresholds. Therefore, we modified the tuning procedure to generate local ORS hindcasts for the 2016-2021 period, which are shown below. The steps to produce local ORS predictions can be summarized in the following:

- the functioning of the "active" thresholds structure is identical to the regional case (Section 4.2.1), but the condition to detect the ORS are modified (Figure 4.10) and the used indicators (Table 4.5) now refer to a single grid-cell (instead of to a sub-region's mean);
- the constraint's module is removed;
- the tuning procedure is carried out as described in Section 4.3.1, but the target consisting of the observed local ORS determined in Section 2.2.1.

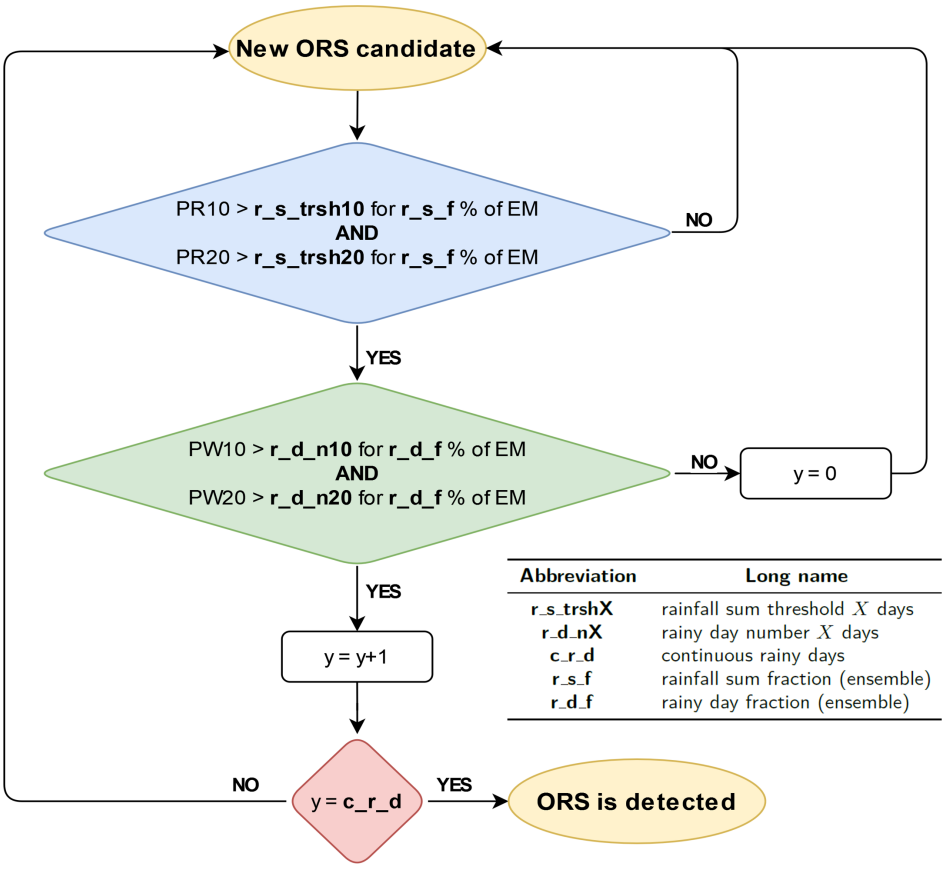

Figure 4.10: Flow chart of the local ORS forecasting routine. The sign in bold, containing underscores, are fixed thresholds calibrated on the observed local ORS. EM is an abbreviation for "Ensemble Members".

Indicator	Description
PR10	Total amount of precipitation predicted in the **next 10 days**
PR20	Total amount of precipitation predicted in the **next 20 days**
PW10	Number of wet days ($>$1 mm) predicted in the **next 10 days**
PW20	Number of wet days ($>$1 mm) predicted in the **next 20 days**

Table 4.5: Indicators used in the threshold-based local ORS detection.

A table containing the full description of the thresholds is present in the Appendix (Table B.5). Due to their importance, we limit here to describe the thresholds denoted as **r_s_trsh10**, **r_s_trsh20** and **r_d_n10**, **r_d_n20**. The first two thresholds represent the minimum accumulation of rain (in respectively 10 and 20 days) necessary to assess the ORS. Their mathematical formulation is:

$$\textbf{r\_s\_trshX} = 0.9 * raintype * X * \textbf{rain\_trsh\_fractX}, \tag{4.2}$$

where X is the number of days (10 or 20), **rain_trsh_fractX** is a threshold representing the minimum number of rainy days in the forecast, and 0.9 is a scaling factor. $raintype$ is a matrix of daily rainfall values specific for each location, which has already been introduced in Equation 4.1. The formula returns therefore a different value for every cell of the grid.

The second two thresholds refers to the minimum number of wet days that must be observed before asserting the start of the rains. Such value is given by

$$\textbf{r_d_nX} = X * \textbf{rain_trsh_fractX}, \tag{4.3}$$

where X is the number of days, **rain_trsh_fract** is a threshold representing the minimum number of rainy days in the forecast and it is the same threshold present also in the formulation of **r_s_trshX** thresholds. All the thresholds share the common "sub-threshold" **rain_trsh_fractX**.

As with the regional predictions, the local algorithm was tuned on the 2016-2021 ECMWF operational forecast. During the tuning it was found that a single threshold configuration covering all the spatial domain was performing poorly, leading to extremely late forecast in the north of the region. Consequently, a mask was created splitting the region according to the forecast's performance of the first attempt. The optimal thresholds for the two sub-regions are reported in Appendix (Table B.6). The validation of the local forecast was performed following the same steps as Section 4.3.2. The same information shown in Figure 4.11, but for each individual year, can be found in the Appendix (Figures B.4 and B.5).

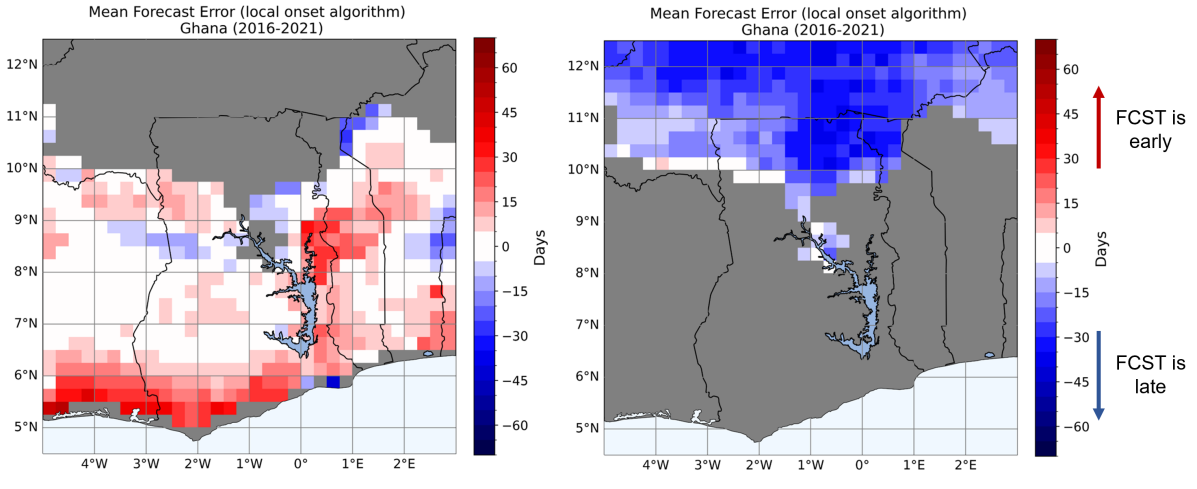

Figure 4.11: Local ORS predictions' validation. 2016-2021 mean forecast error (local ORS observation minus local ORS predictions). Reds shades indicate that the forecast predicted the ORS too early, blue shades the opposite. The top right shows that late forecast errors are widespread in the north of the region.

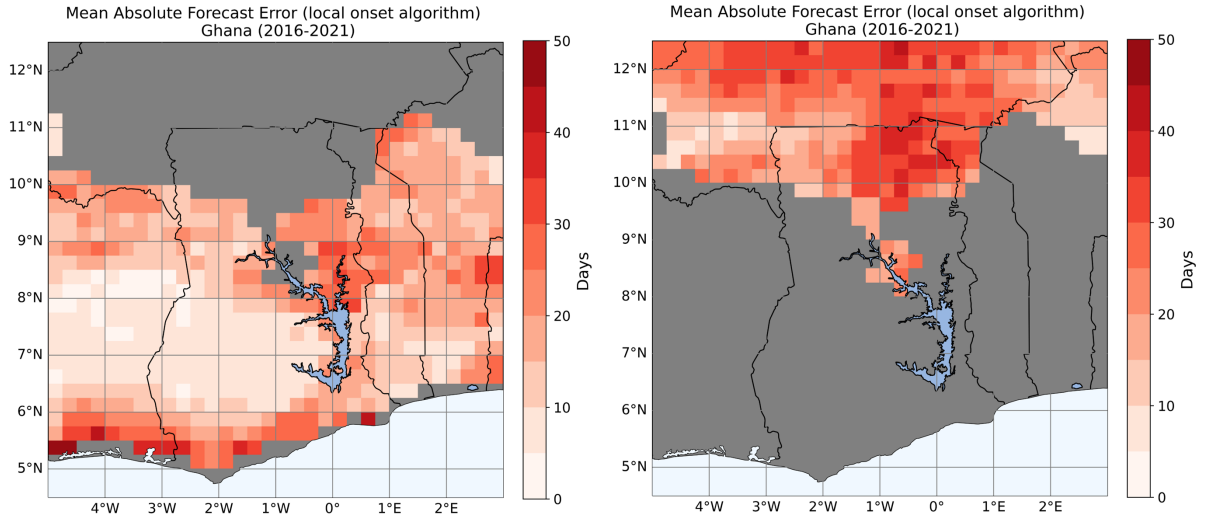

Figure 4.12: Local ORS predictions' validation. 2016-2021 mean absolute forecast error. Opposite to the mean errors, the absolute version ensures that no compensation effect are displayed.

The imposed conditions to detect the local ORS (Figure 4.10) check that a minimum amount of precipitation is forecast in the 10 and 20 days following the candidate onset date, and that such rainfall amount is spread over a minimum number of days. The mean forecast errors show that the algorithm is too conservative for the north of the region, predicting the ORS systematically late, with peaks of 45 days of delay (Figure 4.11, right). An analysis of the algorithm's sensitivity, comparing different thresholds' configurations, concludes that this is mainly caused by the condition on the number of wet days (in particular 20 days ahead).

In the center and south of the region the forecast errors are generally lower and we observe both early and late detection areas. The area close to the coast, between 5-0°W, is affected by the strongest discrepancies, showing large early detection errors (up to 50 days). The absolute mean forecast error (Figure 4.12) resembles closely the previous plots, meaning that little compensation is present in the mean forecast errors plot. This also implies that the directionality of the discrepancies between observations and predictions is constant during the entire time series.

Turning to the assessment of this local ORS prediction experiment, we produce a plot showing the number of years when the forecast error is limited to 10 days, with the same intent of what described in Section 4.3.4. Compared with the results obtained with the regional ORS predictions (with 3 sub-regions), it is clear that the local algorithm has lower skill. The fraction of area meeting the required "quality standard" for each of the 6 years is just 4%, rising to 11.6 % for 5 years and 18.2 % for 4 years. In comparison, the fractions for the regional predictions were respectively 12.5%, 30.1% and 19%. However, it is remarkable that the local algorithm shows a pattern of high forecast performance very similar to the regional predictions (in this case we consider only the center and south of the region): the high skill hotspot of south-west Ghana is again clearly recognizable (especially comparing left Figure 4.3 with left Figure 4.12).

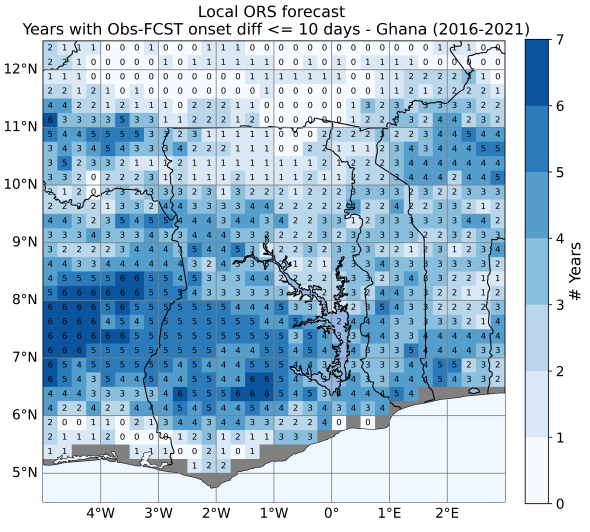

Figure 4.13: Number of years when the difference between the observed local ORS and the local ORS forecast are equal or lower than 10 days. The plots shows the results for both the south and the north of the regions, but the reader has to remember that a different thresholds configuration was adopted for the two sub-regions . The small figures in each grid cell report the exact number of years meeting the above-mentioned condition.

To conclude, it is necessary to stress that the outcomes shown above are the result of a limited number of experiments, where different conditions and thresholds configuration were matched and tested. Therefore, it is likely that a better combination of indicators and conditions can be found, improving the local ORS forecast. To the author's opinion, this is almost certain for the north of the region, where the forecast errors suggest the presence of a systematic bias.

Chapter 5

A Supervised Learning Approach

This chapter describes the attempt to predict the regional ORS with a Random Forest (RF) algorithm. The aim is to demonstrate that the ORS prediction is an issue that could greatly benefit from a supervised learning approach. The obtained results suggest though that further research is needed to see an operational application of such technique.

5.1 Usage and Benefit of Supervised Learning in Predicting the ORS

Supervised Learning (SVL) gained increasingly higher attention among the tools to analyse and predict large amount of data, thanks to remarkable performance achieved in a variety of applications, from healthcare to the financial sector. However, the operational prediction of the ORS has been rarely addressed with such technique. Dodd and Jolliffe [15] made an early attempt using Linear Discriminant Analysis (LDA) on precipitation data to distinguish if an observed increase of rainfall is leading to the ORS. This method was not predicting the onset's date itself, but it was among the firsts to treat the ORS as a classification problem. This perspective will be adopted in the Random Forest algorithm presented in Section 5.2. The study used 22 station-based precipitation records in Mali and Burkina Faso, covering up to 77 years. The method was able to correctly classify up to 85% of Mali's and 78% of Burkina Faso's potential onsets.

Lala, Tilahun, and Block [28] implemented both a Partial Least Squares regression (PLS) and a RF classification to predict the ORS in the Ethiopian Highlands, using CHIRPS daily precipitation (1981-2019) and a set of atmospheric variables as seasonal predictors. Regarding the categorical forecast produced by the RF algorithm, predictions showed higher skills over climatology and had similar performance with dynamical models. It is worth noting that this study performed reanalysis forecasts, which may return significantly better performances compared to operational scenarios. To conclude, a recent work by Nielsen et al. [34] has implemented a RF algorithm to forecast the onset in a single location of Tanzania. It is notable that the model is trained with a mix of both local (precipitation records) and regional indicators (IOD and ENSO3.4), which emphasizes the importance of including large-scale indicators to improve local ORS predictions, as advocated by Nicholson [33].

Compared to the threshold-based methods described in Chapter 4, applying SVL to the prediction of the ORS offers three a priori advantages:

1. SVL allows to blend local predictors with large-scale indices, with a degree of flexibility that could never be achieved using thresholds;
2. a SVL classification algorithm can return a direct probability estimate of the ORS, overcoming the limitation imposed by the binary logic underlying thresholds methods;
3. any SVL method can be easily scaled-up to work on different regions other than the one for which it was created, as long as the training procedure is repeated.

These advantages follow the order of their relative relevance. Further comparisons between SVL and threshold-based approach is provided in the evaluation of the Random Forest algorithm.

5.2 The Random Forest algorithm

The Random Forest (RF) method was introduced by Breiman [9]. It consists of a collection of distinct decision trees, which are weakly correlated with each other. Each decision tree makes splits at each node based on the value of a randomly selected predictor variable. These splits aim to best differentiate the target events from non-event cases, using the training data provided for the model. The splitting continues recursively until a termination condition is met, or because all the remaining training examples belong to a single class (event or non-event) or because there are too few examples to continue the process. In this stage, a leaf node is created, which provides a forecast based on the proportion of training examples associated with each event class. When an unseen data is provided to the RF, the information is filtered through each decision tree until a leaf is reached. The forecast will then be the mean of the probabilistic forecast produces by each tree of the forest.

As mentioned above, our attempt to produce a forecast of the ORS adopting a RF algorithm draws inspiration from the work of Dodd and Jolliffe [15]. However, our study differs from this seminal paper in two key aspects. Firstly in its scope, as we aim to predict the onset date of the rainy season and not to distinguish between true and false start (given a set of possible onset dates). Secondly in the method, since Dodd et al. used a Linear Discriminant Analysis instead of a RF. Despite this differences, Dodd's work and our algorithm share the same essence, i.e. treating the prediction of the ORS as a classification problem.

In the following lines we describe our RF-based forecasts of the ORS. We aim to predict the regional ORS of the three sub-regions defined in Section 2.3.1. For each calendar date, the set of predictors described in Table 5.1 is fed to the RF algorithm. It is important to remember that each variable is averaged over the entire sub-region's area. During the training procedure, each day is associated with 27 pieces of information, describing the meteorological situation of precipitation and meridional wind field from 30 days in the past to 15 days in the future. The labelling of the onset dates should therefore allows the RF model to recognised patterns leading to the ORS.

Unlike other SVL techniques, RF does not have an internal memory; it is therefore pivotal to construct an appropriate set of predictors to effectively capture the temporal development of the atmosphere. Given these predictors, the RF algorithm processes the data as described earlier and returns a probability for each date, representing the likelihood of it being the ORS. Even if the method is not handling actual time series, the task of the RF can be deemed as a time series classification. Following the terminology introduced in Section II, we performed both reanalysis forecasts and hindcasts. Their results differ quite substantially and are therefore presented separately in Section 5.3 and Section 5.4 respectively.

The RF models have been implemented using `scikit-learn` (version 1.2.1) [35], an open-source machine learning library for `python`.

Predictor	Description
Doy	Day of the year of the candidate onset date (1 is 1st Jan, 366 is 31st Dec)
PW[n]	Number of wet days (>1 mm) in the n-days window **following** the candidate onset date (n = 5 or 10 days)
PR[n]	Amount of precipitation in the n-days window **following** the candidate onset date (n = 5 or 10 days)
W[n]	Number of wet days (>1 mm) in the n-days window **preceding** the candidate onset date (n = 5, 10, 15, 20, 25 or 30 days)
R[n]	Amount of precipitation in the n-days window **preceding** the candidate onset date (n = 5, 10, 15, 20, 25 or 30 days)
V850P[n]	Meridional wind speed at 850 hPa n-days **preceding** the candidate onset date (n = 5, 10, 15, 20, 25 or 30 days)*
V850F[n]	Meridional wind speed at 850 hPa n-days **following** the candidate onset date (n = 5, 10 or 15 days)*

Table 5.1: Predictors used in RF algorithm. Each quantity is intended to be the spatial average over a specific sub-region.
*The meridional wind speed is the 5-days mean of the speed recorded/forecast between the day in square brackets and 5 days before. For example, the predictors V850P[10] is the mean meridional wind speed observed between 5 and 10 days before the candidate onset date.

5.3 Reanalysis forecast

5.3.1 Method

Here we apply the RF algorithm to ERA5 reanalysis data. We simulate the forecast of the ORS of the three sub-regions for the 2010-2021 period, using as training set the data of 1981-2009. We built the test and the training set of predictors as described in Table 5.1. To resemble operational settings, for each candidate onset date we used only ERA5 data from 30 days before up to 15 days after that day. Due to the large amount of data and the considerable computing time of generating the test and training dataset, the detection window of each year is optimised as follow:

- `Ghana South` and `Ghana Center` sub-regions: from the 1st February up to 30 days after the observed ORS of the specific year (and sub-region);
- `Ghana North` sub-region: from the 1st March up to 30 days after the observed ORS of the specific year.

The target fed into the RF consists of an array of binary values corresponding to the days of the training time window. The target is set to 0 if the day is not the ORS, to 1 for a range of days spanning from 3 days before to 5 days after the observed ORS. The main features and the optimal hyper-parameters of the RF algorithm are reported in Table 5.2.

RF model	Ghana South	Ghana Center	Ghana North
Training period	1981-2009 (29 years)	1981-2009 (29 years)	1981-2009 (29 years)
Test period	2010-2021 (12 years)	2010-2021 (12 years)	2010-2021 (12 years)
# trees	100	100	200
Min_sample_leaf	7	18	5
Min_sample_split	5	5	3

Table 5.2: RF configuration for the reanalysis forecast. The optimal hyper-parameters has been selected through a 10-fold cross validation of the training set.

5.3.2 Results

For each day of the test years, the RF algorithm provides a probability indicating the likelihood of the ORS occurring in that particular date. The ORS is then forecast on the second consecutive day when the RF probability is higher than a fixed thresholds (0.6 in this study). In Figure 5.1 we show both the obtained predictions (left hand side) and the RF model performance (right hand side).

The RF algorithm is evaluated through the coefficient of determination (R2), while the ORS forecasts through the Root Mean Squared Error (RMSE) and the correlation coefficient (r). It appears that the method is performing better for the northern region, while the center and south of the country share lower performances. Focusing on the ORS predictions (Figure 5.1 **a, b** and **c**), in 2013 for `Ghana North` and 2011 for `Ghana South` no onset is detected, since the RF probability never reaches the fixed threshold during those years. Overall, we can distinguish the test years into two groups:

1. when the difference between the observations and the predictions does not exceed 5 days, which account for 75% of the test set (27 out of 36 analysed years);
2. when the difference between the observations and the predictions exceed 5 days, which account for the remaining 25% of the sample (9 out of the 36 analysed years).

The second group contains years in which the ORS is forecast from 10 days too late up to 42 days too early. In most of the cases (5 out of 7 years, excluding when no ORS was detected at all), the large discrepancies happens on the early detection side. This tendency is a common feature of all the test years, even for the one belonging to the first group.

We analysed the precipitation records of the rainy seasons falling in the second group, focusing on the 5 years of very early ORS detection. Two common patterns are likely responsible for the important forecast errors: seasons with a false start/dry spell in the first month of detection (as 2010 and 2015 in `Ghana Center`) and seasons when the ORS is very smooth, with precipitation gradually increasing during the spring months (2013 and 2016 in `Ghana North`, 2010 and 2015 in `Ghana center`, 2012 and 2017 in `Ghana South`). The "smoothness" of the ORS is known to play a remarkable influence on forecast's skills, as discussed in Section 3.5.2. Though, it is also visible that in every year when the ORS prediction is extremely early, a peak of RF probability is detectable later, often in correspondence of the observed ORS. This indicates that the RF algorithm is able to detected the observed ORS in most of the years with the before-mentioned patterns, however the forecasting routine will always assign the onset to the first peak of probability reaching the fixed threshold.

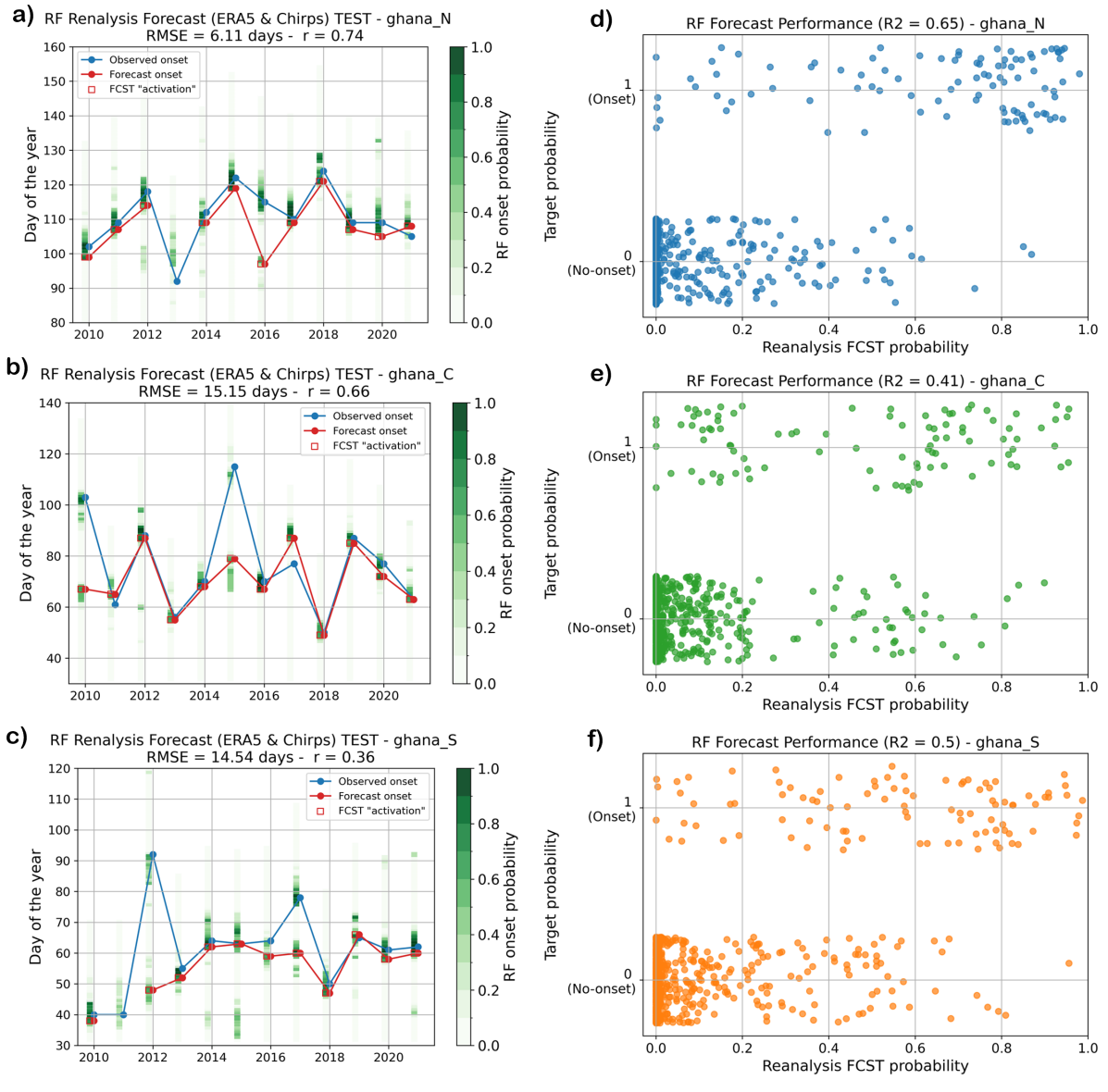

Figure 5.1: RF reanalysis forecast. **Left**: predicted and observed ORS for the set of test year. For each day of the year (y-axis), the probability generated by the RF algorithm is shown as green boxes. The day in which the threshold is reached is marked with red contours. **Right**: scatter plots comparing the target ORS and the corresponding RF predictions. It is recognisable a marked class imbalance, as the majority of days within a rainy season are classified as "no-onset" ones.

5.4 Hindcast: Simulating Operational Settings

5.4.1 Method

In this section the RF method is applied to operational weather forecasts issued by the ECMWF between 2016 and 2021. The goal is to generate a prediction of the ORS for each of the three Ghana's sub-regions.

The framework used to build an train the RF algorithm is identical to the one described for the reanalysis forecast (Section 5.3.1). However, the operational setting introduced two main complications: firstly, the creation of a uniform dataset containing precipitation and wind information from 30 days before to 25 days after a certain date; secondly, the scarcity of available data (only 6 years compared to the 41 available in the reanalysis).

The first task presented challenges due to the time lag with which the observations are made available. For

near-present data, a preliminary version of both CHIRPS (*CHIRPSprelim*) and ERA5 (*ERA5T*) data is available, but with 5-day delay respect to real-time. The data gap between the target date and the preceding 5 days was filled up with ECMWF medium-range precipitation and wind forecast at lead time of 1 days. We assumed this forecast to be the closest available approximation to an observation. A schematic of such datasets "glueing" is present in Figure 5.2. We checked that this "glueing" practise produces time series of precipitation and meridional wind close to the historical observation. Obviously, the days after the considered date, which are actual operational forecast, can show considerable deviation in absolute values from the observation. On the other hand, the seasonal trends are preserved and no signed of excessive divergence from the observation has been detected.

The second complication, i.e. the scarcity of data, has proven to be more difficult to solve and a definitive solution has not yet been found. On one hand, the performance of a RF model is dramatically sensitive to the size of the training set. On the other hand, using four or even five of the six available years for the training would leave an extremely small test sample to judge the performance of the model, especially if the prediction results in a single value per year. Therefore, two strategies were adopted: first, to maintain a decent size of the test set, multiple RF models sharing the same configuration of hyper-parameters were built, each of them being trained on 5 year and tested on the remaining one. Using all the possible combinations of training and test sets, 6 RF model were trained and tested. Second, the predictors (Table 5.1) belonging to the training set were build from each ensemble member of the forecast separately, increasing the training size of a factor 50. During the test phase, the set of predictors is instead built from the ensemble mean of the operational forecast.

RF model	Ghana South	Ghana Center	Ghana North
Training period	5 years (among 2016-2021)	5 years (among 2016-2021)	5 years (among 2016-2021)
Test period	1 year (among 2016-2021)	1 year (among 2016-2021)	1 year (among 2016-2021)
# trees	100	100	100
Min_sample_leaf	2	1	3
Min_sample_split	5	5	5

Table 5.3: RF configuration for the hindcasts. The optimal hyper-parameters has been selected through a 10-fold cross validation of the training set (on the test set of year 2016).

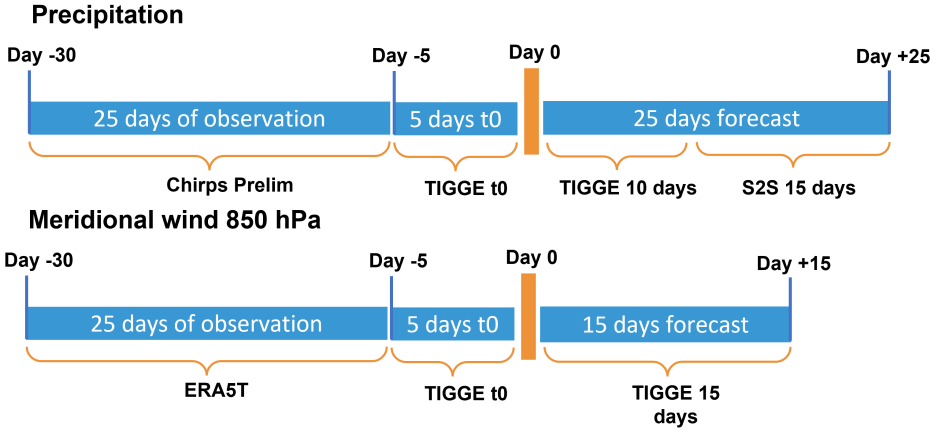

Figure 5.2: Visualisation of the dataset's generation for the operational RF forecast.

5.4.2 Results

The ORS predictions shown in Figure 5.3 are obtained from RF probabilities with the same algorithm described at the beginning of Section 5.3.2, but using a lower threshold (0.4 instead of 0.6). We remind that such predictions were generated with the very same data which would be used in operational settings. These results depict therefore a realistic picture of what would be the RF performance in operational usage.

Comparing the discrepancies between observations and predictions, it is clear that the switch to the operational scenario caused a marked deterioration of the forecast's performance. As in the reanalysis case, we can distinguish two groups of test years:

1. years when the forecast error stays within 5 days (50%, or 9 out of 18 test years);
2. years when the forecast error exceed 5 days (the remaining 50%).

72% of the analysed years the hindcast predicts earlier ORS than the observed ones. We observe that the RF algorithm returns smaller probabilities compared to the reanalysis forecast, reason why it was necessary to lower the detection threshold to 0.4. Moreover, the RF probabilities appears to be much less concentrated around the forecast or observed onset, in contrast with the reanalysis forecast. These elements suggest that not only the RF method has weak skill in detecting the observed ORS, but it also struggles to distinguish clear onset's patterns in general. However, the biggest difference between the hindcast (Figure 5.3) and the prediction on the reanalysis data (Figure 5.1) is that in the latter case, even in years where the forecast error was large, potential causes could be identified in the precipitation records of that season. For the hincasts, instead, no clear or common patter emerge among years with high prediction's errors.

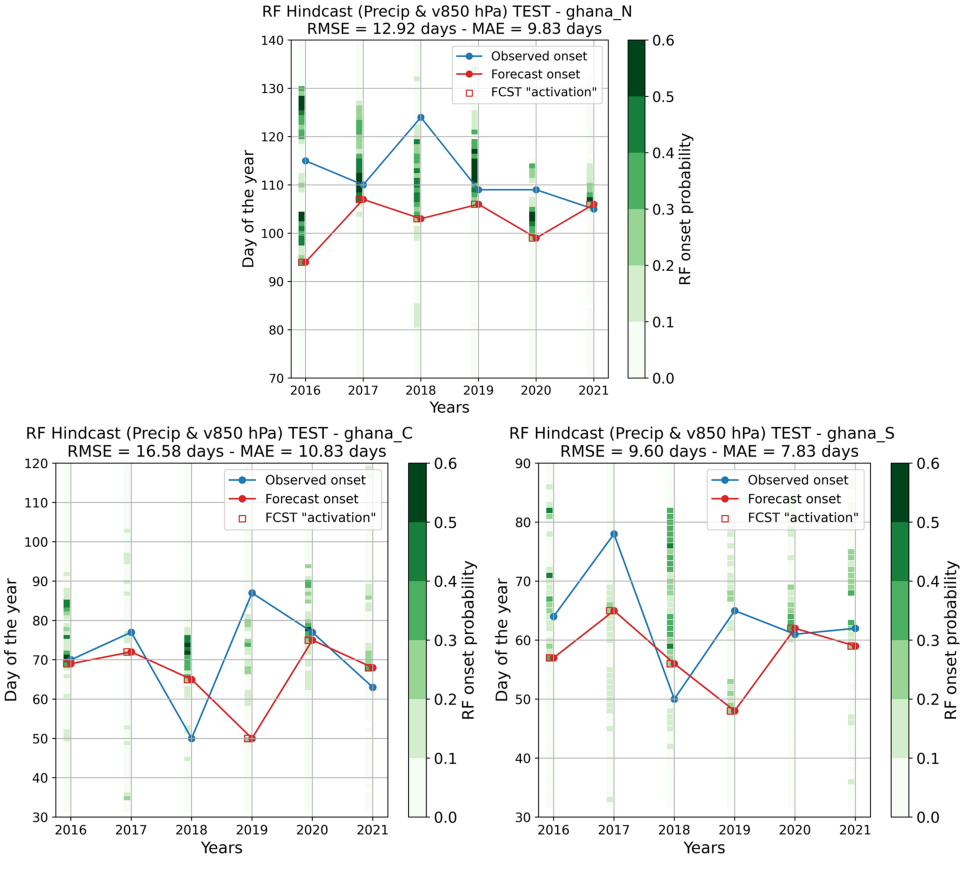

Figure 5.3: Predicted and observed ORS for the set of test year. For each day of the year (y-axis), the probability generated by the RF algorithm is shown as green boxes. The day in which the threshold is reached is marked with red contours. The above plots are obtained merging the prediction results from 6 different RF model, each of them was trained of the 5 years left after removing the one chosen to play the test role.

5.5 Discussion and Future Developments

Monsoonal precipitation patterns are dominated by non-linear relationships among the meteorological variables, making very challenging to predict the ORS with traditional methods. The potential of Machine Learning (ML) techniques lies in the possibility to uncover such complex patterns and to find the optimal predictors of the start of the rains. Moreover, it allows to overcome the binary logic at the core of most currently-used predicting routines. Such elements lead to conclude that the prediction of the ORS is a problem that could greatly benefit from the adoption of ML methods. However, the results of our experiments

using a RF algorithm showed that this method, or the framework we used, can not deliver precise and reliable ORS predictions in operational settings.

While the results obtained in the reanalysis forecast (Figure 5.1) can be considered promising, the large prediction errors of the hindcast (Figure 5.3) are a clear sign that the RF method is not working properly in realistic conditions. In reanalysis forecasts the RF method is prone to mis-predict some false onsets but it is still able to recognise the observed ORS, meaning that the forecast's accuracy could be enhanced by improving the algorithm which transform the RF probabilities time series in the punctual prediction of the ORS. For the hindcasts, instead, little improvement is possible as the RF probabilities often do not respond at the observed ORS. Moreover, a forecasting algorithm is good not only if it can correctly predict every instance, but if it is possible to assign a cause to the mis-predictions. In the previous results section we showed that this is the case for the reanalysis forecasts (Section 5.3.2), but not for the hindcast.

We can hypothesize three causes to explain the different performances between the reanalysis and the operational forecast:

1. the use of a different dataset;
2. the smaller amount of available data in the training phase;
3. the hindcast training procedure, which uses each member of the ensemble forecast separately.

Of these causes, the first one has the most obvious implications: the reanalysis dataset, even if used to simulate operational forecasts, represents the scenario of a "perfect" forecast, which is surely far from being realistic. The use of real operational forecast introduces therefore a layer of intrinsic uncertainty, which is not due to the RF model, but due to the ECMWF prediction's errors.

The second and third causes mentioned before are instead tightly related: the creation of the predictors set using each member of the perturbed forecast separately can greatly increase the amount of training samples, but it has limited effect on increasing the actual amount of information fed in the RF algorithm. Ultimately, this strategy creates, for a certain date, 50 different trajectories which shares the same outcome, i.e. the target value that date. It is likely that such situation is not optimal, as it might undermine the ability of the RF in identifying clear pattern among the data. Notwithstanding this shortcomings, the described training procedure was the only to guarantee enough training samples to train the RF algorithm with the available operational data.

Despite our experiment of using SVL to predict the ORS was not successful, we have shown the potentiality of these techniques and the remarkable advantages that they carry. As these methods are not the focus of the study, we could explore only one algorithm (RF) with few possible configurations. The author is therefore optimistic that further research, using more powerful ML algorithms and different settings, will bring improvements in accuracy and reliability of ORS operational forecasts. As an example, our RF algorithm was fed with meteorological information covering only the temporal domain; the implementation of Neural Networks techniques could instead produce local ORS forecasts endowed with both an intrinsic spatial and temporal information of the precipitation records, bringing unseen improvements in the forecasting skills. Moreover, thanks to their ability in highlighting patterns among different data sources, SVL methods can be apply to directly assess the consequences of weather pattern on farming activities [7].

Closing Remarks

This study concerns the detection and the prediction of the rainy season onset (ORS) in West African countries, in specific Ghana, Togo, large part of Benin and southern Burkina Faso. The guiding lodestar of the analysis has been the desire to prioritise the practical applications of the collected results. This aspect has gained particular weight in the second part of the thesis, where different forecasting algorithm have been tested in operational-like conditions.

We conclude that detecting and, consequently, forecasting the ORS is far from being a trivial task, even adopting fundamental tools such as precipitation's thresholds. Two causes underlie this issue: firstly, the absence of a unique, easily applicable definition of the start of the rainy season; secondly, the necessity to deal with precipitation records, which are affected by high inter-annual variability and low spatial coherence. The combination of these two elements poses challenges to the development of ORS predicting algorithm able to work properly not only at a single location, but over an entire region. However, the latter must be the objective of any study aiming to reach operational applications, as this work does.

The thesis is characterised by a regional point of view, both in the initial climatological research and in the following development of a forecasting algorithm. The rationale of this choice is that a regional perspective can mitigate the issues related to the meteo-climatic specificities of each individual location, allowing to generate more reliable ORS predictions. In particular, the main advantage would be the possibility to adopt the large-scale atmospheric circulation's components as predictors of the start of the rains, (partially) disentangling the ORS detection from rainfall data. This initial hypothesis revealed to be true only to a limited extent: while in the climate we found a clear connection between the ORS and the wind circulation, both at high and low pressure levels (925, 850 and 200 hPa), the huge inter-annual variability of the wind fields prevents their operational usage as proxies of the rainy season onset. The only operational application of wind's speed forecast was the generation of temporal boundaries of the ORS. The conclusion is that precipitation's data remain the primary and almost unique ingredient for reliable ORS forecast over the analysed area.

The prediction of the regional ORS was achieved through thresholds-based algorithms, built to detect the onset from operational rainfall forecast (both at medium-range and sub-seasonal lead-times, up to 25 days ahead). A thresholds' tuning procedure was established to ensure that the forecasts aim at the regional ORS. The obtained regional predictions were compared to observed local ORS. The ORS forecast performances showed large spatial disparities considering a single-year prediction, but fairly low temporal fluctuation if looking at an individual location.

The spatial disparities, in particular among locations belonging to the same sub-region, unveiled the crucial role of the observed regional ORS (the target of the thresholds' tuning), which descend from a Principal Component Analysis on rainfall records. We discovered that the regional forecasts are representative only for regions where the first principal component is strongest. The low temporal fluctuation of forecast errors, eventually, confirm the previous findings: the disparities of forecast's performances are not accountable to random events but, on the opposite, to systematic biases. In light of these findings, we are confident that it is possible to forecast the ORS on a regional scale only for areas which benefit from a high spatial coherence in the precipitation pattern. Of the region we analysed, such areas were found to be the south-west Ghana (between 4-1°W, 5.5-9°N), southern Burkina Faso (between 4-0°W, 10.5-12°N) and northern Benin (between 1-3°E, 10-12°N).

Two other experiments contributed to gain a more complete overview of alternative forecasting strategies. First, the thresholds-based algorithm was modified to generate local ORS forecast. We achieved performances similar to the regional case for the southern part of the analysed region (below 10°N), while a strong late

forecast error dominated the northern part. However, due to the exploratory nature of the work, the author refrains from making definitive conclusions about the possibility to improve the forecast performances. Second, a completely different approach was addressed, using a Random Forest model to generate regional ORS predictions. Both observational (up to 30 days in the past) and forecast (up to 25 days ahead) data were fed into the model, exploring the possibility of combining rainfall and wind information. Promising results were found in forecast of reanalysis data, but the same model performed very poorly when trained on operational data. We conclude that the framework in which we built this supervised learning approach, even if promising on paper, is not ready to be used in operational settings.

Our finding suggests that a thresholds-based algorithm, processing precipitation's forecast, is the approach able to predict the onset of the rainy season with the highest degree of reliability. However, only specific regions of the entire study area show forecast's skill justifying their operational deployment (Figure 5.4). This happens adopting both a regional and a local predicting algorithm, even if the locations of the high skill areas can differ between the two approaches. The validation of the forecast returns better performances adopting the regional method. However, the author is confident that blending the two approaches could bring improvements, such as an increased extend of the area where operational use is applicable and a higher degree of confidence in the predictions.

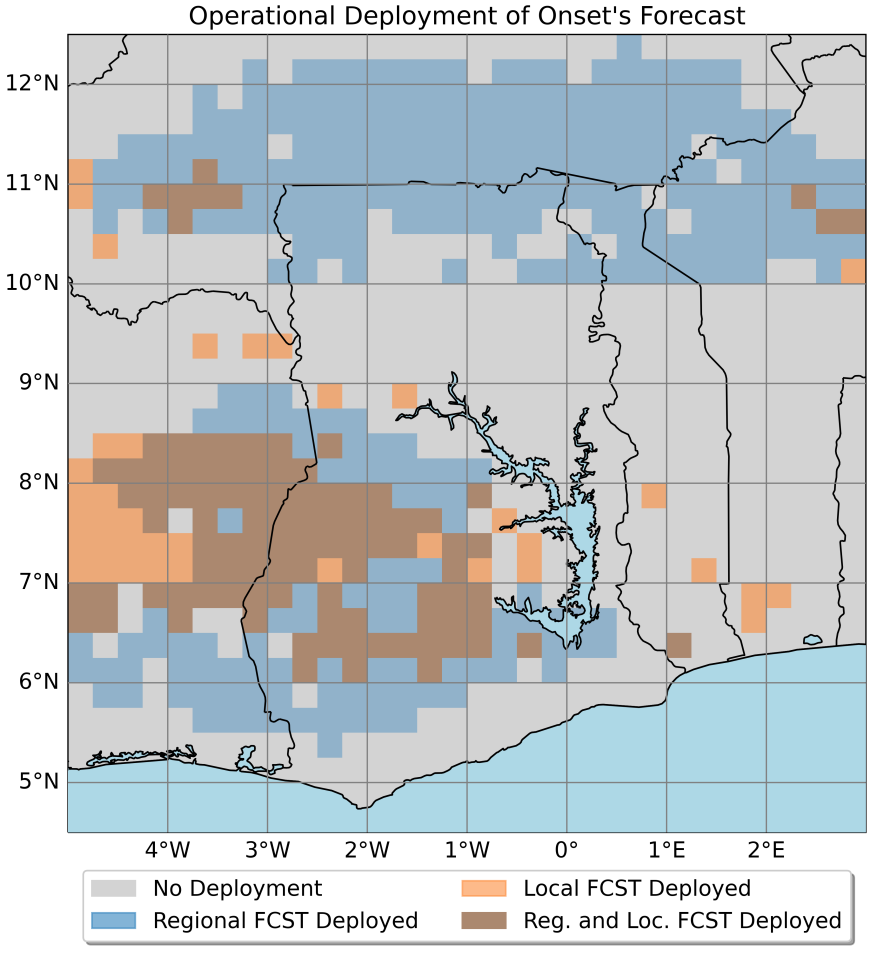

Figure 5.4: Areas where the threshold-based forecasting algorithm of the ORS can be operationally deployed. These areas are defined as the cells where the forecast error is lower than 10 days for at least 5 out of the 6 analysed years, identified from Figure 4.6 and 4.13). The map divides area where the onset can be forecast with a regional and with a local approach. The two areas overlap considerably, meaning that the hotspots of high performance tend to interest the same spatial domain.

Lastly, it is imperative to introduce some *caveats* to these conclusions. Every result of this thesis originates from a variety of choices, which the author has endeavoured to make as non-arbitrary as possible. However, a certain degree of arbitrariness is often intrinsic in the rainy season's study area, especially when reproducing

operational conditions. This holds, first and foremost, for the chosen definition of the onset of the rainy season, both in local and region frameworks. The Bombardi et al. [5] (for the local ORS) and the Camberlin and Diop [11] (for the regional ORS) were employed in quality of detection's method endowed with a solid physical meaning and a very limited arbitrary component. However, the end-user's needs might require the adoption of different onset's definitions; this is often the case in agronomic circumstances, e.g. when the onset must fulfil the water requirements of specific crop species. In this scenario, it is necessary to modify the target of the forecast calibration, operation which can be very difficult if using a regional approach. Consequently, we conclude that the prediction of the rainy season onset benefits from the utmost flexibility when carried out in local settings.

Bibliography

[1] Leonard Amekudzi et al. "Variabilities in Rainfall Onset, Cessation and Length of Rainy Season for the Various Agro-Ecological Zones of Ghana". In: *Climate* 3.2 (June 2015), pp. 416–434. DOI: 10.3390/cli3020416. URL: https://doi.org/10.3390/cli3020416.

[2] Winifred Ayinpogbilla Atiah et al. "Validation of Satellite and Merged Rainfall Data over Ghana, West Africa". In: *Atmosphere* 11.8 (Aug. 2020), p. 859. DOI: 10.3390/atmos11080859. URL: https://doi.org/10.3390/atmos11080859.

[3] Frank Baffour-Ata et al. "Effect of climate variability on yields of selected staple food crops in northern Ghana". In: *Journal of Agriculture and Food Research* 6 (Dec. 2021), p. 100205. DOI: 10.1016/j.jafr.2021.100205. URL: https://doi.org/10.1016/j.jafr.2021.100205.

[4] Vibeke Bjornlund, Henning Bjornlund, and Andre F. Van Rooyen. "Why agricultural production in sub-Saharan Africa remains low compared to the rest of the world – a historical perspective". In: *International Journal of Water Resources Development* 36.sup1 (May 2020), S20–S53. DOI: 10.1080/07900627.2020.1739512. URL: https://doi.org/10.1080/07900627.2020.1739512.

[5] Rodrigo J. Bombardi et al. "Sub-seasonal Predictability of the Onset and Demise of the Rainy Season over Monsoonal Regions". In: *Frontiers in Earth Science* 5 (Feb. 2017). DOI: 10.3389/feart.2017.00014. URL: https://doi.org/10.3389/feart.2017.00014.

[6] Philippe Bougeault et al. "The THORPEX Interactive Grand Global Ensemble". In: *Bulletin of the American Meteorological Society* 91.8 (Aug. 2010), pp. 1059–1072. DOI: 10.1175/2010bams2853.1. URL: https://doi.org/10.1175/2010bams2853.1.

[7] Christopher Bowden, Timothy Foster, and Ben Parkes. "Identifying links between monsoon variability and rice production in India through machine learning". In: *Scientific Reports* 13.1 (Feb. 2023). DOI: 10.1038/s41598-023-27752-8. URL: https://doi.org/10.1038/s41598-023-27752-8.

[8] Joseph Boyard-Micheau et al. "Regional-Scale Rainy Season Onset Detection: A New Approach Based on Multivariate Analysis". In: *Journal of Climate* 26.22 (Nov. 2013), pp. 8916–8928. DOI: 10.1175/jcli-d-12-00730.1. URL: https://doi.org/10.1175/jcli-d-12-00730.1.

[9] Leo Breiman. In: *Machine Learning* 45.1 (2001), pp. 5–32. DOI: 10.1023/a:1010933404324. URL: https://doi.org/10.1023/a:1010933404324.

[10] C. Cabot Venton. "Economics of resilience to droughts: Ethiopia analysis." In: (2018), p. 44. URL: https://2017-2020.usaid.gov/sites/default/files/documents/1867/Ethiopia_Economics_of_Resilience_Final_Jan_4_2018_-_BRANDED.pdf.

[11] P Camberlin and M Diop. "Application of daily rainfall principal component analysis to the assessment of the rainy season characteristics in Senegal". In: *Climate Research* 23 (2003), pp. 159–169. DOI: 10.3354/cr023159. URL: https://doi.org/10.3354/cr023159.

[12] Thierry Coulibaly, Moinul Islam, and Shunsuke Managi. "The Impacts of Climate Change and Natural Disasters on Agriculture in African Countries". In: *Economics of Disasters and Climate Change* 4.2 (Jan. 2020), pp. 347–364. DOI: 10.1007/s41885-019-00057-9. URL: https://doi.org/10.1007/s41885-019-00057-9.

[13] Andrew Dawson. "eofs: A library for EOF analysis of meteorological, oceanographic, and climate data". In: *J. Open Res. Softw.* 4.1 (2016). ISSN: 2049-9647.

[14] Moctar Dembélé and Sander J. Zwart. "Evaluation and comparison of satellite-based rainfall products in Burkina Faso, West Africa". In: *International Journal of Remote Sensing* 37.17 (July 2016), pp. 3995–4014. DOI: 10.1080/01431161.2016.1207258. URL: https://doi.org/10.1080/01431161.2016.1207258.

[15] Doris E.S. Dodd and Ian T. Jolliffe. "Early detection of the start of the wet season in semiarid tropical climates of western Africa". In: *International Journal of Climatology* 21.10 (2001), pp. 1251–1262. DOI: 10.1002/joc.640. URL: https://doi.org/10.1002/joc.640.

[16] Thulani Dube et al. "The Impact of Climate Change on Agro-Ecological Based Livelihoods in Africa: A Review". In: *Journal of Sustainable Development* 9.1 (Jan. 2016), p. 256. DOI: 10.5539/jsd.v9n1p256. URL: https://doi.org/10.5539/jsd.v9n1p256.

[17] Harriet Achiaa Dwamena, Kassim Tawiah, and Amanda Serwaa Akuoko Kodua. "The Effect of Rainfall, Temperature, and Relative Humidity on the Yield of Cassava, Yam, and Maize in the Ashanti Region of Ghana". In: *International Journal of Agronomy* 2022 (Jan. 2022). Ed. by Magdi Abdelhamid, pp. 1–12. DOI: 10.1155/2022/9077383. URL: https://doi.org/10.1155/2022/9077383.

[18] FAO. *Land use indicators (FAOSTAT)*. (Accessed on 14-May-2023). 2020. URL: https://www.fao.org/faostat/en/#data/.

[19] Rory G. J. Fitzpatrick et al. "The West African Monsoon Onset: A Concise Comparison of Definitions". In: *Journal of Climate* 28.22 (2015), pp. 8673–8694. DOI: 10.1175/JCLI-D-15-0265.1. URL: https://journals.ametsoc.org/view/journals/clim/28/22/jcli-d-15-0265.1.xml.

[20] Bernard Fontaine, Samuel Louvet, and Pascal Roucou. "Definition and predictability of an OLR-based West African monsoon onset". In: *International Journal of Climatology* 28.13 (Nov. 2008), pp. 1787–1798. DOI: 10.1002/joc.1674. URL: https://doi.org/10.1002/joc.1674.

[21] Chris Funk et al. "The climate hazards infrared precipitation with stations—a new environmental record for monitoring extremes". In: *Scientific Data* 2.1 (Dec. 2015). DOI: 10.1038/sdata.2015.66. URL: https://doi.org/10.1038/sdata.2015.66.

[22] Talardia Gbangou et al. "Rainfall and dry spell occurrence in Ghana: trends and seasonal predictions with a dynamical and a statistical model". In: *Theoretical and Applied Climatology* 141.1-2 (Apr. 2020), pp. 371–387. DOI: 10.1007/s00704-020-03212-5. URL: https://doi.org/10.1007/s00704-020-03212-5.

[23] Hersbach H. et al. *ERA5 hourly data on pressure levels from 1940 to present*. Copernicus Climate Change Service (C3S) Climate Data Store (CDS) (Accessed on 13-Dec-2022). 2023. DOI: 10.24381/cds.bd0915c6.

[24] James W. Hansen et al. "Potential value of GCM-based seasonal rainfall forecasts for maize management in semi-arid Kenya". In: *Agricultural Systems* 101.1 (2009), pp. 80–90. ISSN: 0308-521X. DOI: https://doi.org/10.1016/j.agsy.2009.03.005. URL: https://www.sciencedirect.com/science/article/pii/S0308521X09000468.

[25] ILO. *ILO modelled estimates database (ILOSTAT)*. Available from https://ilostat.ilo.org/data/. (Accessed on 14-May-2023). 2020.

[26] Serge Janicot, JEAN-PHILIPPE LAFORE, and CHRIS THORNCROFT. "THE WEST AFRICAN MONSOON". In: *The Global Monsoon System*. WORLD SCIENTIFIC, Apr. 2011, pp. 111–135. DOI: 10.1142/9789814343411_0008. URL: https://doi.org/10.1142/9789814343411_0008.

[27] Ian T Jolliffe. *Principal component analysis*. Vol. 29. Springer series in statistics. Springer, New York, NY, 2002, p. 488. DOI: 10.1007/b98835.

[28] Jonathan Lala, Seifu Tilahun, and Paul Block. "Predicting Rainy Season Onset in the Ethiopian Highlands for Agricultural Planning". In: *Journal of Hydrometeorology* 21.7 (July 2020), pp. 1675–1688. DOI: 10.1175/jhm-d-20-0058.1. URL: https://doi.org/10.1175/jhm-d-20-0058.1.

[29] P. Laux, H. Kunstmann, and A. Bárdossy. "Predicting the regional onset of the rainy season in West Africa". In: *International Journal of Climatology* 28.3 (Mar. 2008), pp. 329–342. DOI: 10.1002/joc.1542. URL: https://doi.org/10.1002/joc.1542.

[30] Brant Liebmann and JoséA. Marengo. "Interannual Variability of the Rainy Season and Rainfall in the Brazilian Amazon Basin". In: *Journal of Climate* 14.22 (Nov. 2001), pp. 4308–4318. DOI: 10.1175/1520-0442(2001)014<4308:ivotrs>2.0.co;2. URL: https://doi.org/10.1175/1520-0442(2001)014%3C4308:ivotrs%3E2.0.co;2.

[31] Roberto Mera, Arlene G. Laing, and Frederick Semazzi. "Moisture Variability and Multiscale Interactions during Spring in West Africa". In: *Monthly Weather Review* 142.9 (Sept. 2014), pp. 3178–3198. DOI: 10.1175/mwr-d-13-00175.1. URL: https://doi.org/10.1175/mwr-d-13-00175.1.

[32] Dang-Quang Nguyen, James Renwick, and James McGregor. "Variations of monsoon rainfall: A simple unified index". In: *Geophysical Research Letters* 41.2 (Jan. 2014), pp. 575–581. DOI: 10.1002/2013gl058155. URL: https://doi.org/10.1002/2013gl058155.

[33] Sharon E. Nicholson. "Climate and climatic variability of rainfall over eastern Africa". In: *Reviews of Geophysics* 55.3 (July 2017), pp. 590–635. DOI: 10.1002/2016rg000544. URL: https://doi.org/10.1002/2016rg000544.

[34] Kristian Nielsen et al. "Random Forest approach to forecast onset date and duration of rainy season in Tanzania". In: (Feb. 2023). DOI: 10.5194/egusphere-egu23-16200. URL: https://doi.org/10.5194/egusphere-egu23-16200.

[35] F. Pedregosa et al. "Scikit-learn: Machine Learning in Python". In: *Journal of Machine Learning Research* 12 (2011), pp. 2825–2830.

[36] Manuel Rauch et al. "Seasonal Forecasting of the Onset of the Rainy Season in West Africa". In: *Atmosphere* 10.9 (Sept. 2019), p. 528. DOI: 10.3390/atmos10090528. URL: https://doi.org/10.3390/atmos10090528.

[37] Seneviratne S.I. et al. "Weather and Climate Extreme Events in a Changing Climate. In Climate Change 2021: The Physical Science Basis. Contribution of Working Group I to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change". In: *IPCC Sixth Assessment Report* (2021), pp. 1513–1766. DOI: 10.1017/9781009157896.013. URL: https://www.ipcc.ch/report/ar6/wg1/chapter/chapter-11/.

[38] R. D. Stern, M. D. Dennett, and D. J. Garbutt. "The start of the rains in West Africa". In: *Journal of Climatology* 1.1 (Jan. 1981), pp. 59–68. DOI: 10.1002/joc.3370010107. URL: https://doi.org/10.1002/joc.3370010107.

[39] Justin Stoler et al. "Deconstructing "malaria": West Africa as the next front for dengue fever surveillance and control". In: *Acta Tropica* 134 (June 2014), pp. 58–65. DOI: 10.1016/j.actatropica.2014.02.017. URL: https://doi.org/10.1016/j.actatropica.2014.02.017.

[40] Benjamin Sultan and Serge Janicot. "The West African Monsoon Dynamics. Part II: The "Preonset" and "Onset" of the Summer Monsoon". In: *Journal of Climate* 16.21 (Nov. 2003), pp. 3407–3427. DOI: 10.1175/1520-0442(2003)016<3407:twamdp>2.0.co;2. URL: https://doi.org/10.1175/1520-0442(2003)016%3C3407:twamdp%3E2.0.co;2.

[41] Benjamin Sultan et al. "Climate Drives the Meningitis Epidemics Onset in West Africa". In: *PLoS Medicine* 2.1 (Jan. 2005). Ed. by Simon Hales, e6. DOI: 10.1371/journal.pmed.0020006. URL: https://doi.org/10.1371/journal.pmed.0020006.

[42] UNDP Africa and UNDP Africa. "Africa Human Development Report 2012 Towards a Food Secure Future". en. In: (2012). DOI: 10.22004/AG.ECON.267636. URL: https://ageconsearch.umn.edu/record/267636.

[43] F. Vitart et al. "The Subseasonal to Seasonal (S2S) Prediction Project Database". In: *Bulletin of the American Meteorological Society* 98.1 (Jan. 2017), pp. 163–173. DOI: 10.1175/bams-d-16-0017.1. URL: https://doi.org/10.1175/bams-d-16-0017.1.

[44] Liangzhi You et al. "What is the irrigation potential for Africa? A combined biophysical and socioeconomic approach". In: *Food Policy* 36.6 (Dec. 2011), pp. 770–782. DOI: 10.1016/j.foodpol.2011.09.001. URL: https://doi.org/10.1016/j.foodpol.2011.09.001.

Appendix A

Atmospheric Circulation's Proxies of ORS

A.1 Wind Composite Maps

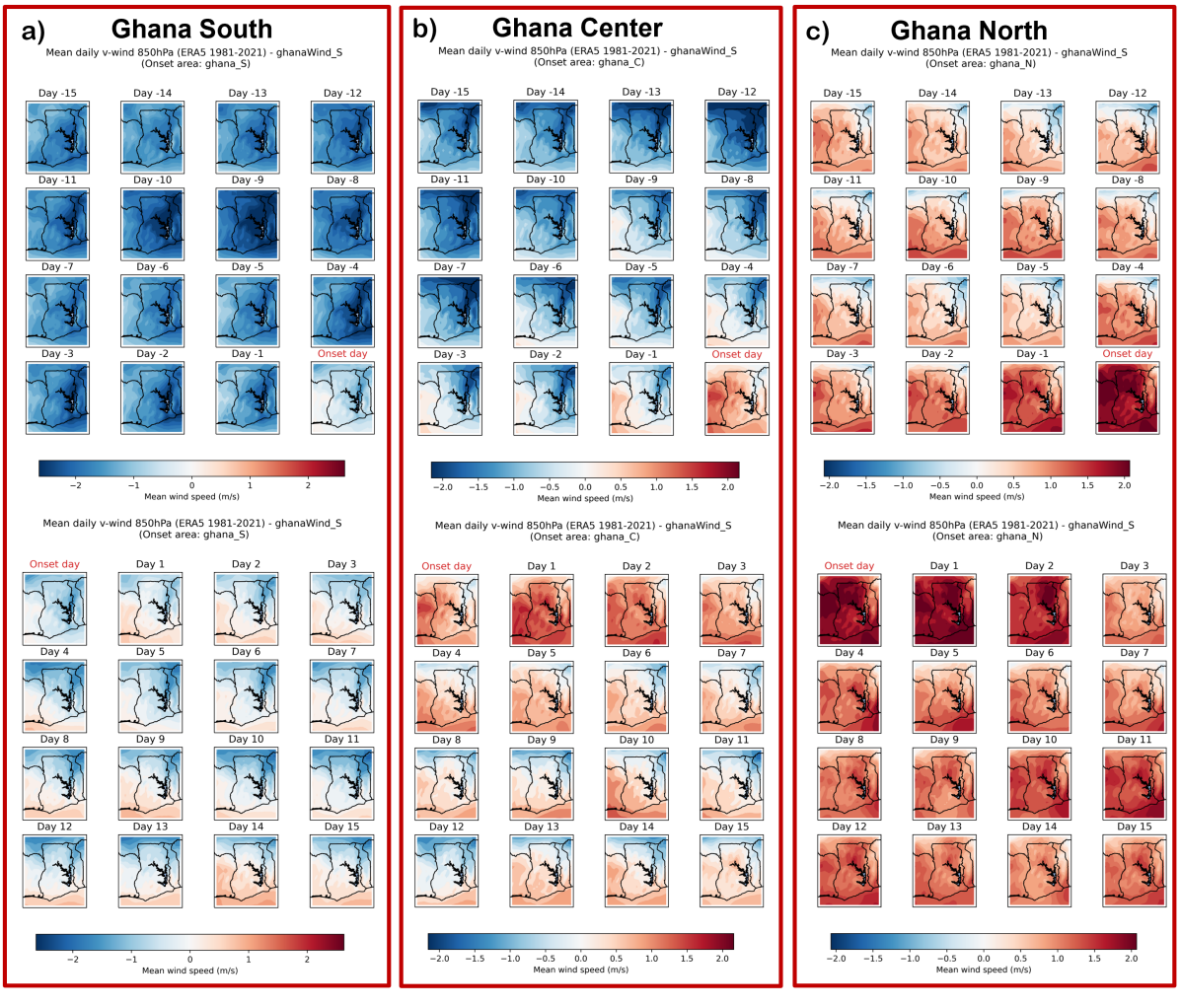

Figure A.1: Meridional wind velocity at 850 hPa during the month of the onset (ERA5, 1981-2021 daily average).

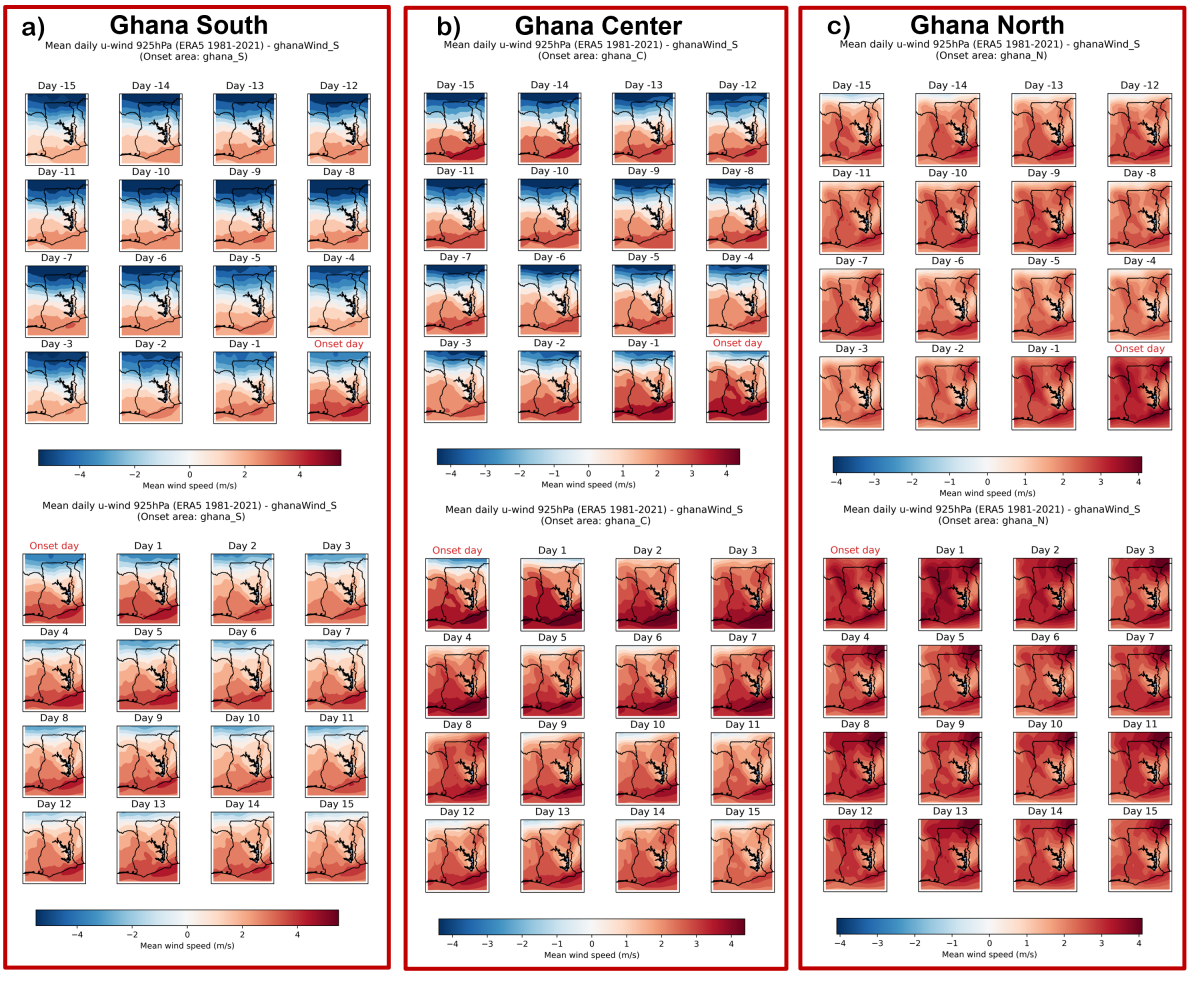

Figure A.2: Zonal wind velocity at 925 hPa during the month of the onset (ERA5, 1981-2021 daily average).

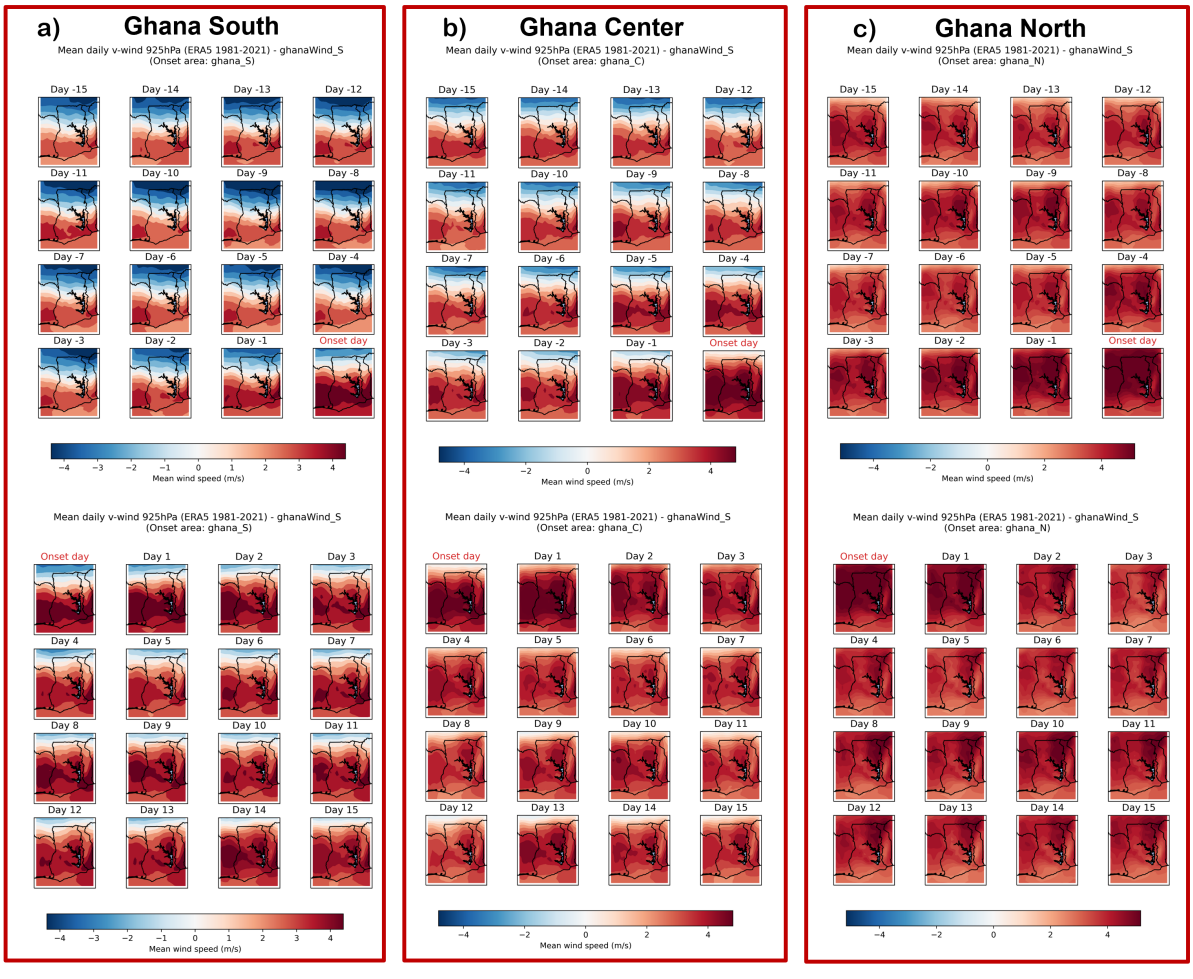

Figure A.3: Meridional wind velocity at 925 hPa during the month of the onset (ERA5, 1981-2021 daily average).

A.2 Wind Inter-annual Variability

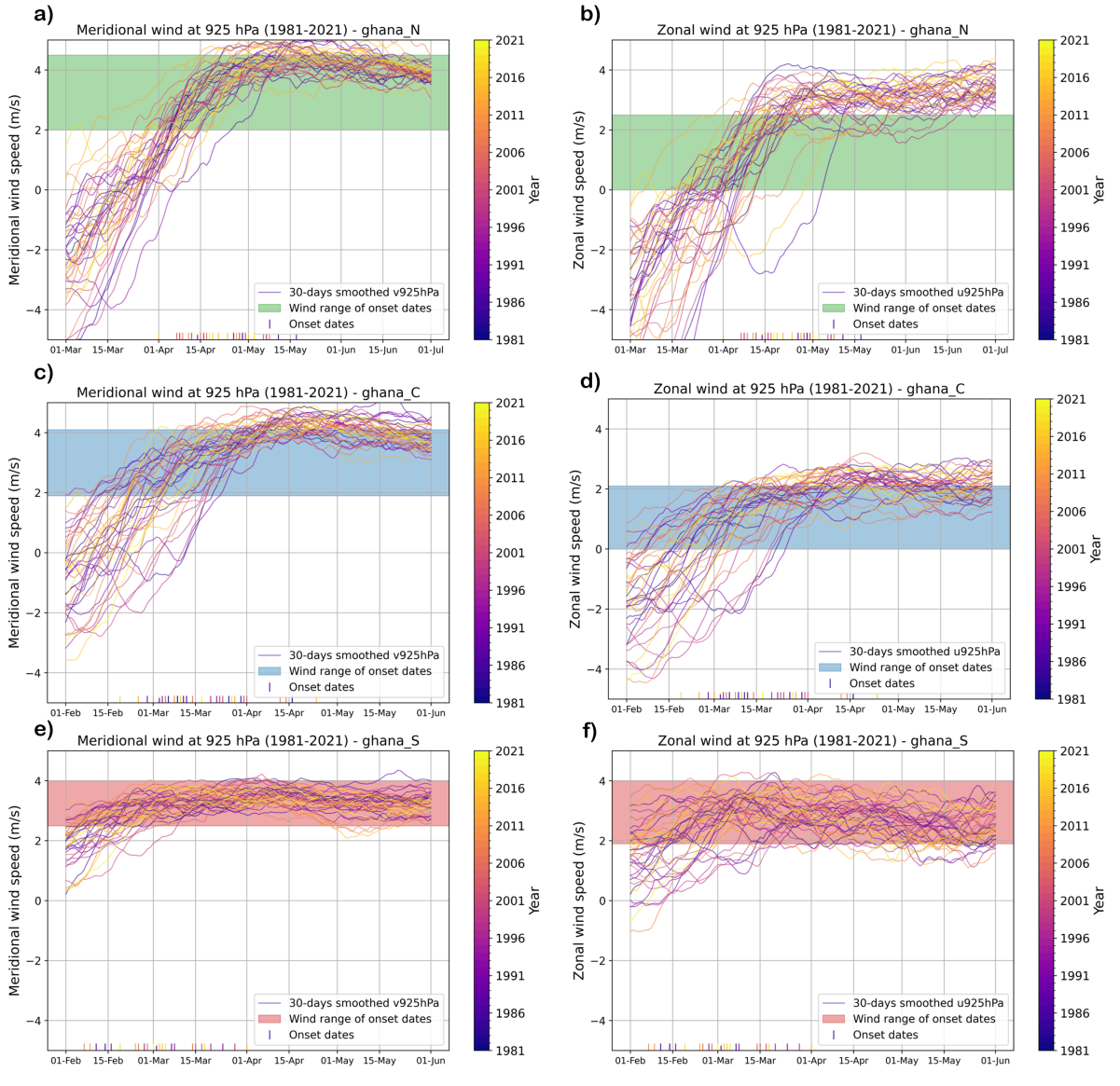

Figure A.4: 30-days smoothed wind profiles during the start of the rainy season (ERA5 1981-2021). The shaded bands mark the wind speed intervals where the ORS is observed during the year of the climatology. The interval's boundaries have been computed according to Figure 3.7. The rainy season's onset dates corresponding to each plotted year are shown as a rug plot above the x-axis.

Appendix B

Thresholds-based Algorithms

B.1 Regional ORS forecast

B.1.1 Optimal Optimal Threshold's Configuration

Threshold	Ghana South	Ghana Center	Ghana North
r_d_n (days)	2	1	1
c_r_d (days)	3	5	3
r_s_f	0.8	0.7	0.6
r_d_f	0.7	0.6	0.6
rain_trsh_fract	0.6	0.7	0.6
raintype	'Medium'	'Low'	'Low'

Table B.1: Optimal thresholds' values for the regional ORS predicting algorithm (3 sub-regions).

Threshold	Ghana South	Ghana Center	Ghana North
V850min	-2.1	-2	0.5
V850max	-0.6	-0.5	1
w_b_f	0.6	0.6	0.7
c_WL_d	5	5	5
c_WU_d	10	10	10
tot_rd10	7	7	7
tot_rd20	12	12	12
r_d_f_e	0.6	0.6	0.85
c_PW_d	3	3	3

Table B.2: Optimal thresholds' values of constrain on candidate ORS dates (3 sub-regions).

Threshold	Ghana SE	Ghana SW	Ghana CE	Ghana CW	Ghana NS	Ghana NN
r_d_n (days)	1	2	1	1	1	1
c_r_d (days)	5	5	3	5	3	3
r_s_f	0.8	0.7	0.7	0.6	0.6	0.6
r_d_f	0.6	0.6	0.7	0.6	0.6	0.7
rain_trsh_fract	0.7	0.9	0.9	0.7	0.6	0.6
raintype	'Low'	'Low'	'Low'	'Low'	'Low'	'Low'

Table B.3: Optimal thresholds' values for the regional ORS predicting algorithm (6 sub-regions).

Threshold	Ghana SE	Ghana SW	Ghana CE	Ghana CW	Ghana NS	Ghana NN
V850min	-3	-3	-0.5	-0.6	-0.5	-0.5
w_b_f	0.7	0.7	0.6	0.6	0.6	0.7
c_WL_d	5	5	5	5	5	5
c_WU_d	10	10	10	10	10	10
tot_rd10	6	6	6	6	5	5
tot_rd20	11	11	11	11	11	10
r_d_f_e	0.7	0.8	0.6	0.6	0.6	0.6
c_PW_d	5	5	5	5	5	5

Table B.4: Optimal thresholds' values of constrain on candidate ORS dates (6 sub-regions).

B.1.2 6 Sub-regions Calibration

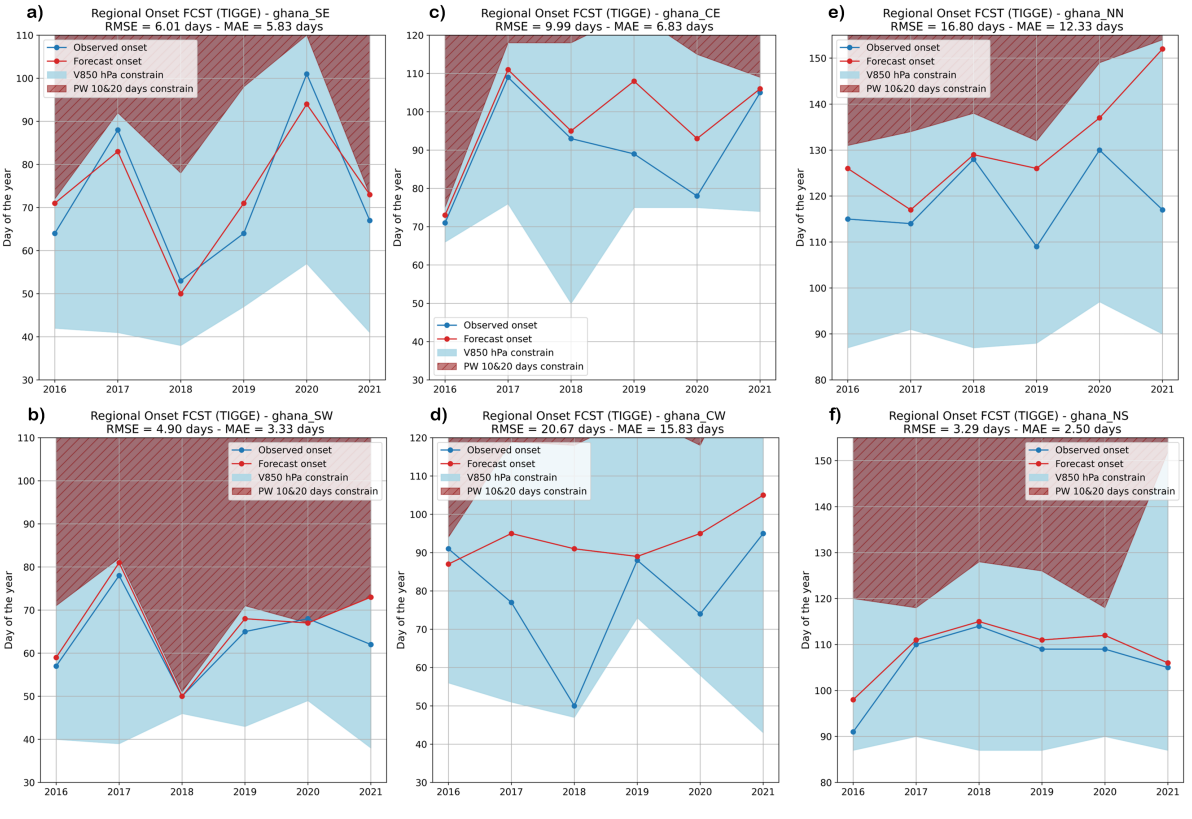

Figure B.1: Threshold-based regional ORS predictions. The constraints on candidate onset date, described in section 4.2.2, are indicated as shading of different colors: blue represent the range of date fulfilling both meridional wind constraints, red indicates the dates excluded by the constraint of the number of rainy days.

B.1.3 Hindcast Validation (2016-2021)

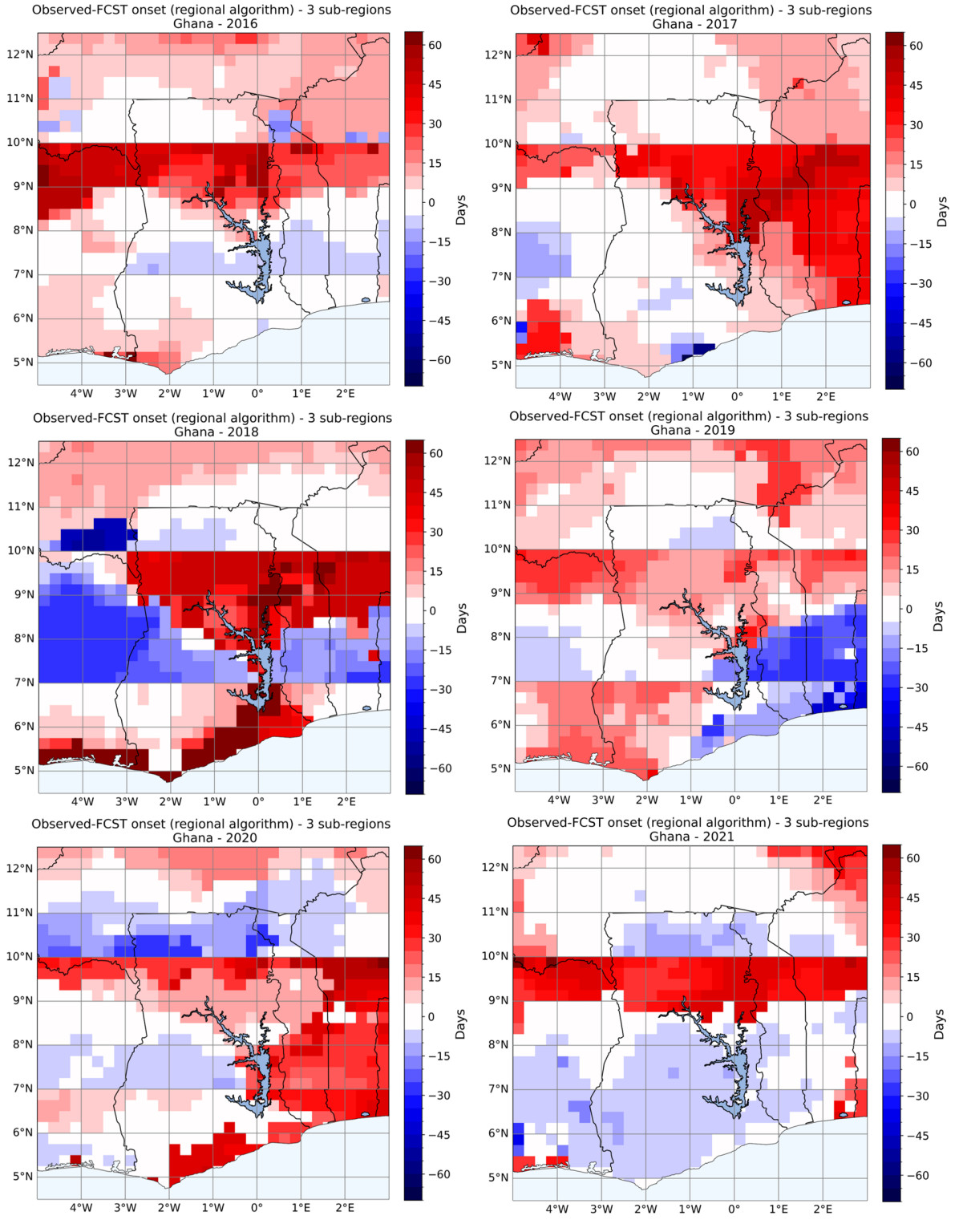

Figure B.2: Regional ORS predictions' validation (3 sub-regions division). Forecast error (local ORS observation minus local ORS predictions). Reds shades indicate that the forecast predicted the ORS too early, blue ones the opposite. The top right shows that late forecast errors are widespread in the north of the region.

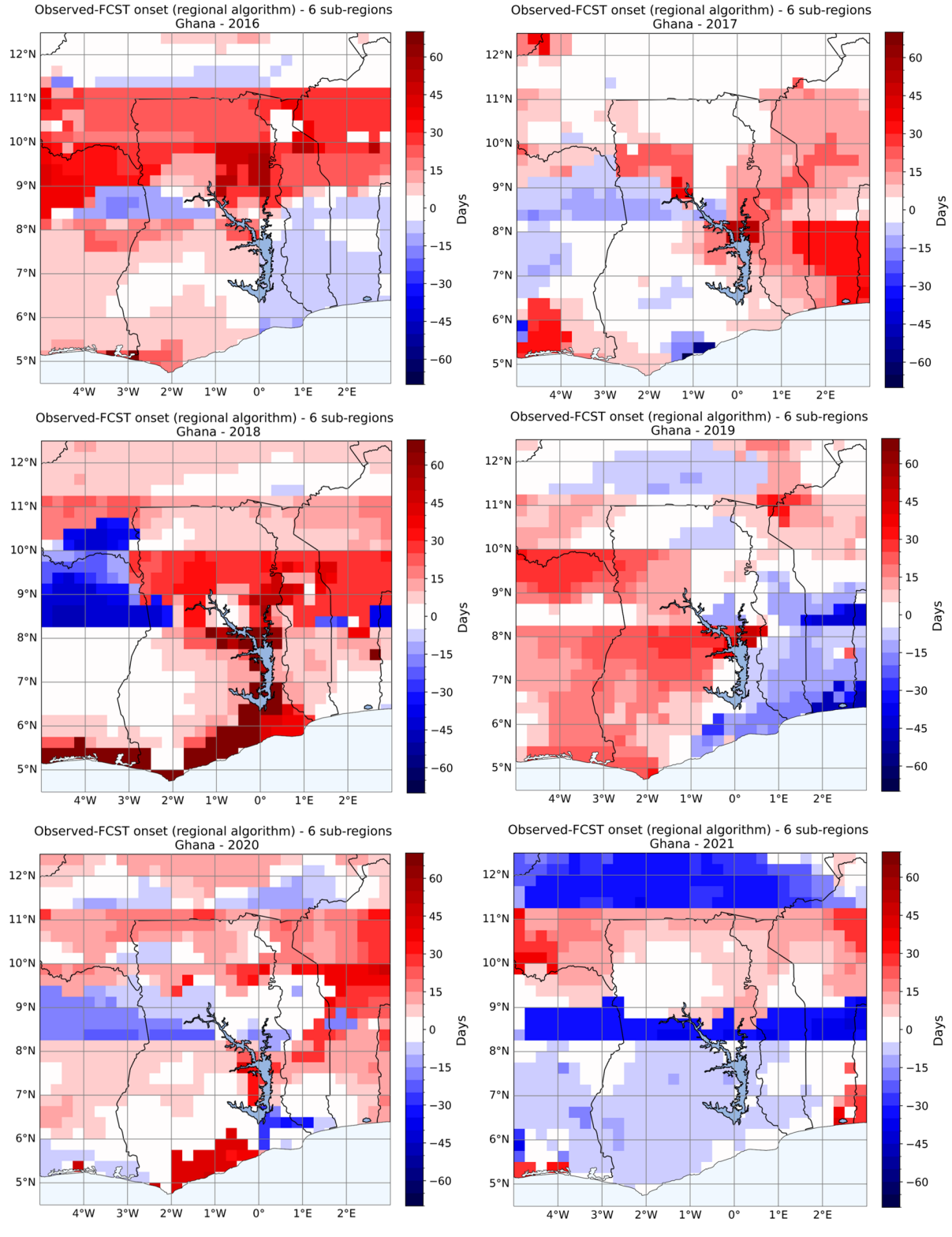

Figure B.3: Regional ORS predictions' validation (6 sub-regions division). Forecast error (local ORS observation minus local ORS predictions). Reds shades indicate that the forecast predicted the ORS too early, blue ones the opposite. The top right shows that late forecast errors are widespread in the north of the region.

B.2 Local ORS forecast

B.2.1 Optimal Threshold's Configuration - Local ORS

Abbreviation	Long name	Description	Unit
r_s_trsh10	rainfall sum threshold 10 days	Minimum cumulative precipitation in the next 10 days	mm
r_s_trsh20	rainfall sum threshold 20 days	Minimum cumulative precipitation in the next 20 days	mm
r_d_n10	rainy day number 10 days	Minimum number of wet day in the next 10 days	day
r_d_n20	rainy day number 20 days	Minimum number of wet day in the next 20 days	day
c_r_d	continuous rainy days	Number of consecutive days which do respect all the threshold's conditions	day
r_s_f	rainfall sum fraction (ensemble)	Minimum fraction of ensemble member fulfilling condition on rainfall sum	-
r_d_f	rainy day fraction (ensemble)	Minimum fraction of ensemble member fulfilling condition on wet day	-

Table B.5: Description of the thresholds used in the forecasting algorithm of the local ORS.

Threshold	South region	North region
c_r_d (days)	3	3
r_s_f	0.7	0.7
r_d_f	0.6	0.6
rain_trsh_fract10	0.4	0.4
rain_trsh_fract20	0.4	0.4
raintype	'Medium'	'Low'

Table B.6: Optimal thresholds' values for the local ORS predicting algorithm.

B.2.2 Hindcast Validation (2016-2021)

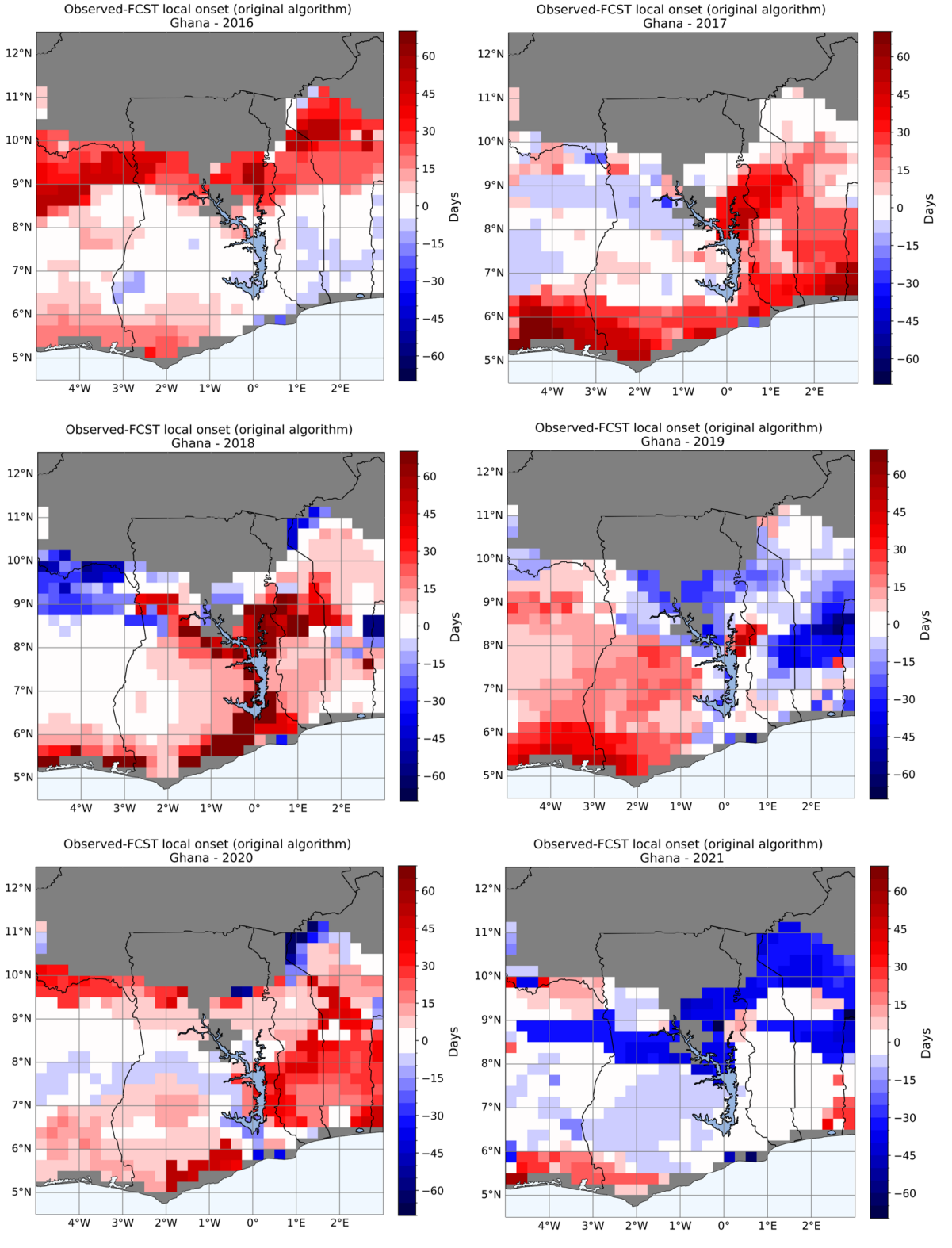

Figure B.4: Local ORS predictions' validation (South region). Forecast error (local ORS observation minus local ORS predictions). Reds shades indicate that the forecast predicted the ORS too early, blue ones the opposite. The top right shows that late forecast errors are widespread in the north of the region.

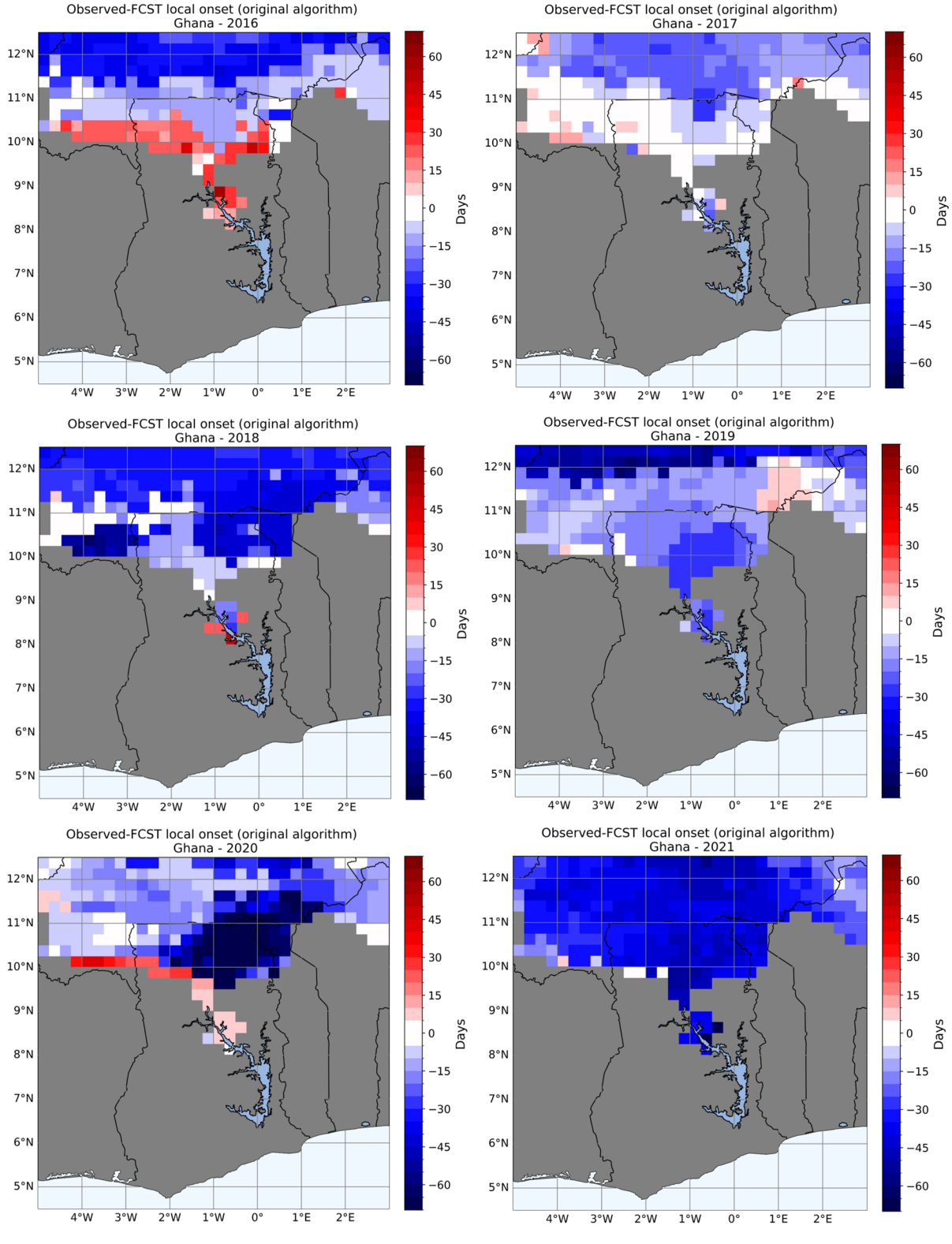

Figure B.5: Local ORS predictions' validation (North region). Forecast error (local ORS observation minus local ORS predictions). Reds shades indicate that the forecast predicted the ORS too early, blue ones the opposite. The top right shows that late forecast errors are widespread in the north of the region.