Report to Securing Water for Food (SWFF) and USAID on
Evaluation of Ignitia Daily Rainfall Forecasts for Subscribers in West Africa
conducted by the International Research Institute for Climate and Society (IRI)
The Earth Institute, Columbia University

Abstract

This report provides an independent verification of daily rainfall forecasts over West Africa from a private sector weather service, Ignitia, during the wet season for 2016 and 2017. The question asked by USAID was, "How accurate are Ignitia's weather forecasts for rainfall?" The forecast product under evaluation is daily rainfall likelihood, targeted to the locations of specific subscribers. The forecasts provide qualitative statements, which are underpinned by quantitative probability ranges, for daily rainfall for the day of forecast issuance (24-hour forecast) and the subsequent day (48-hour forecast). Given the probabilistic format of the forecasts, and the associated event specific (rain versus no-rain) outcome, an assessment of accuracy was not possible. Therefore, we assessed the forecast performance based on the probabilistic reliability and their ability to discriminate rainy days from non-rainy days over the period in question (2016-near present) using reference standard forecast and a simple alternative forecast computed from the widely used and adapted Global Ensemble Forecasting System (GFS), which also happens to be an input to the Ignitia forecasts.

We note that we were unable to obtain competitor's forecasts (the National Meteorological Services, such as the Ghana Meteorological Agency, International news and media outlets, such as weather.com and BBC, and other private sector companies that produce forecasts on comparable spatiotemporal scales) available over the study period, which would allow us to fully examine the veracity of Ignitia's claim of that other forecasts can only achieve 39% accuracy. We were, however, able to determine the reason for Ignitia's claim that their own forecasts were 84% accurate. The claim is misleading. It is based on only a small subset of their forecasts: those for a high probability (>80%) chance of rain, which make up only about 5% of forecasts issued. Overall it cannot be said that their forecasts are 84% accurate. The results of the analysis show that while Ignitia's forecasts for rainfall perform well for certain categories of rainfall, and better than the raw output from a weather prediction model, in consideration of the full forecast spectrum, their claims are not substantiated.

Ignitia's daily rainfall forecasts were found to be reliable, though slightly over-confident. In particular, their forecasts are considerably more reliable than raw model-based predictions. They were also found to have reasonable discrimination between the probabilities issued on rainy days versus those issued on dry days, though this aspect of forecast quality was similar to that found in the raw model output. There is also discernible impact to the reliability using the ENACTS data for Ghana. Particularly for the 24-hr forecasts in 2016, the reliability is nearly perfect for the Ignitia forecasts, but remains very poor and over-confident for GFS. The effect on discrimination is to lower it slightly for both Ignitia and the GFS baseline.

EXECUTIVE SUMMARY

Purpose and Background of the project being evaluated

Many countries in West Africa face complex development and environmental challenges related to various climate and non-climate drivers. Agricultural sustainability is a primary concern for the most vulnerable populations in this region. In this region, it is not standard practice for rain-fed agricultural sector actors to use climate and weather information to inform decisions. Lack of availability and access to climate and weather information is one of the primary drivers of this, while other factors include lack of trust in the information, perception of value and general perceptions of accuracy and timeliness. Securing Water for Food: A Grand Challenge for Development supports building farmers resilience to climate risks. To improve water resource management, the program supports linking farmers with innovative applied solutions from entrepreneurs and scientists. This report aims to evaluate specific claims made by one of the entrepreneur organizations, Ignitia, who provides added value to weather prediction models for farmer subscribers in specific locations.

The International Research Institute for Climate and Society (IRI) provided technical support under the United States Agency for International Development (USAID) and Securing Water for Food (SWFF): A Grand Challenge for Development Project. The purpose of the assessment is to conduct an independent validation study of the forecast quality of weather forecasts produced by Ignitia and disseminated to smallholder farmers in West Africa. Farmers subscribing to Ignitia forecasts were located in Ghana, Senegal, Mali and Burkina Faso.

Assessment Objectives and Methods

Main evaluation questions

The team of consultants conducting this assessment was tasked with broadly answering the question - 'How accurate are Ignitia's weather forecasts for rainfall?'. Here the forecasts predict rainfall occurrence. Given the probabilistic format of the forecasts, and the associated event specific (rain versus no-rain) outcome, an assessment of accuracy was not possible. Therefore, we assessed the forecast performance based on the probabilistic reliability and their ability to discriminate rainy days from non-rainy days over the period in question (2016-near present). The initial verification question was rephrased as: *How skillful are Ignitia's forecasts compared to other available forecast model(s), based on actual in situ observation data?*

Methods

IRI has made considerable contributions to the body of literature on forecast validation methods, including a range of academic papers and guidance reports produced for multi-national organizations, such as the World Meteorological Organization (WMO). Upon identifying the most appropriate methods for validation, IRI outlined a draft work plan, which includes a clear method for addressing the question of interest.

The most appropriate verification metrics, given the forecast format as well as the type of event being forecast, required additional specificity to address skill and to recast the issue of 'accuracy', which is the term used by Ignitia. In line with the industry standards, the analysis addresses the original question in two parts, which encompass the dual nature of probabilistic forecasts for specific categories or events:

Q1a. *Reliability*: Does rainfall occur more frequently when confidence is high that it will rain?

Q1b. *Discrimination*: When it rains, do the forecasts indicate higher confidence in rainfall occurring compared to when it is dry?

Reliable forecasts are ones in which the probabilities mean what they say. For example, when a rainy day is forecast as 60% likely, 60% of the time that specific forecast is issued, a rainy day should occur. Forecasts that can discriminate events provide more confident forecasts for a rainy day on days that it rains compared to days with no rain. These terms may seem equivalent, but rather they are complementary. One could issue forecasts that gave the historical odds of rainfall in a given month, or over a season. Those forecasts would be reliable, since they were designed to give a probability for rain that is consistent with its frequency of occurrence. However, since the same forecast would be issued all the time, there would be no discrimination in forecast confidence on a wet day versus a dry day. These two aspects of forecast quality are both relevant to accuracy, in a probabilistic setting, but give a more complete picture. The 'Hit Rate', especially if focused on only one of several categories, as used by Ignitia to represent acccuracy, could miss important elements of forecast quality. If one were to forecast rain to occur every day, the hit rate would be perfect for "rainy day forecasts" because every day it rained it was forecast to rain. However, the discrimination would be poor because all the forecasts would have been the same, and the reliability would likely be bad (unless it actually did rain every day), because the confidence would be high (>80%) even though [in West Africa] a rainy day was not observed 80% of the time. Discrimination accounts not just for the hit rate, but also penalizes for false alarms. Further elaboration of the terminology and verification metrics is given in Section 3.

The verification work plan was presented to USAID-SWFF and Ignitia. The ability for co-agreement on a work plan for this assessment is a key element of this project. Upon acceptance of the draft work plan by SWFF and Ignitia, IRI conducted robust analyses to address the questions above.

Findings

Ignitia's daily rainfall forecasts were found to be reliable, though slightly over-confident. In particular, their forecasts are considerably more reliable than raw model-based predictions. They were also found to have reasonable discrimination between the probabilities issued on rainy days versus those issued on dry days, though this aspect of forecast quality was similar to that found in the raw model output.

Ignitia's stated performance in their 2016 Annual Report White paper, of "84% accuracy" is verified as a point on the reliability curves (See Figure 3, and Plate 1). However, ignitia's statement of 84% as an overall claim of high accuracy is flawed. The actual interpretation from the verification is that when their forecast says "Rainfall is highly likely", which carries a confidence of >80%, 84% of the time that forecast is issued, it rains. So, for that category, which is probably interpreted by forecast subscribers as "a forecast for rain", it is 84% accurate. The problem with this interpretation is that such a forecast represents only a small percentage of forecasts issued. This single statistic of Ignitia's performance does not represent the complete range of their forecasts – the majority of which are in the less confident categories.

For the observations, two different products were used. The first was a satellite product: the Rainfall Estimate (RFEv2) from NOAA, and the ENACTS[1] dataset, which merges country-held, quality-controlled, station data with satellite information. Because ENACTS uses many more station records in the calibration of the satellite information, it is likely more representative of what farmers actually experience on the ground. This higher resolution data (4km) was upscaled to the same resolution of the RFEv2 (100km) to provide a fair comparison.
The performance of Ignitia's forecast and the NOAA Weather Model prediction both seem slightly lower when the ENACTS 100km daily precipitation dataset is used for the verification. However, the differences are very small and not significant. But given the higher resolution, and reduction in the limitations faced by satellite estimates, the ENACTS data may provide a useful input to the Ignitia forecast processing and a point for future collaboration with the national meteorological services.

Conclusions and Lessons Learned

Ignitia appears to provide a valuable product to small-holder farmers in West Africa. The fact that Ignitia reports good skill on a very low percentage of forecasts, however, means they have not given the full picture of their forecast performance. Only 7% of the forecasts are issued in the very likely wet and very likely dry categories. We note that Ignitia performs well in these categories, with discrimination and reliability better than the comparative baseline NOAA Weather Model output. However, 93% of the time, Ignitia's forecast are not significantly better than the comparison model output that we have used as a baseline forecast.

[1] https://iri.columbia.edu/enacts

We note that the skill comparisons in this evaluation study do not properly address other existing forecast products, as such products were not readily available over the time period of interest and thus could not be included in this analysis. The baseline assessments used in this report consider only raw model output from NOAA's Weather Model. Private sector services such as that from BBC and Weather.com, do involve post-processing. However, those services also appear to address mainly cities, rather than subscriber-specific locations, and also may not focus on bias corrections specific to West Africa as Ignitia has done.

Table of Contents

1) Background

USAID approached the International Research Institute for Climate and Society (IRI), Columbia University, to conduct an independent verification study on skill of the daily rainfall forecasts that Ignitia provides by subscription to smallholder farmers in West Africa. The purpose of the assessment is to conduct an independent verification study of the forecast quality of those weather forecasts. Farmers subscribing to Ignitia forecasts were located in Ghana, Senegal, Mali and Burkina Faso.

In the ISKA White Paper of 2017, Ignitia claims high skill of their forecasts relative to other existing forecast products. According to their verification, the ISKA forecast product is said to be accurate 84% of the time, while competitors' forecasts are only 39% accurate (USAID SWFF statement). USAID tasked IRI to evaluate Ignitia's overall forecast performance and specifically to address their claim of 84% accuracy.

The forecast product under evaluation is daily rainfall likelihood, targeted to the locations of specific subscribers. The forecasts provide qualitative statements, which are underpinned by quantitative probability ranges, for daily rainfall for the day of forecast issuance (24-hour forecast) and the subsequent day (48-hour forecast) are of interest. Given the probabilistic format of the forecasts, and the associated event specific (rain versus no-rain) outcome, an assessment of accuracy was not possible. Therefore, we assessed the forecast performance based on the probabilistic reliability and their ability to discriminate rainy days from non-rainy days over the period in question (2016-near present).

The countries included in this assessment include: Burkina Faso, Ghana, Mali, and Senegal. IRI, as an independent party evaluator of forecast quality, has used industry accepted methods and standards for this project (WMO's Standard Verification Scores[2]). As shown in Figure 1, the majority of 2016 forecasts occurred in Ghana and Mali. In 2017 there was a significant increase in subscribers in all countries addressed (Figure 1). For both years the quantity of unique subscribers increases rapidly in June through August, corresponding to the planting seasons in the respective national and sub-national areas.

[2] http://www.bom.gov.au/wmo/lrfvs/Attachment_II-8.doc

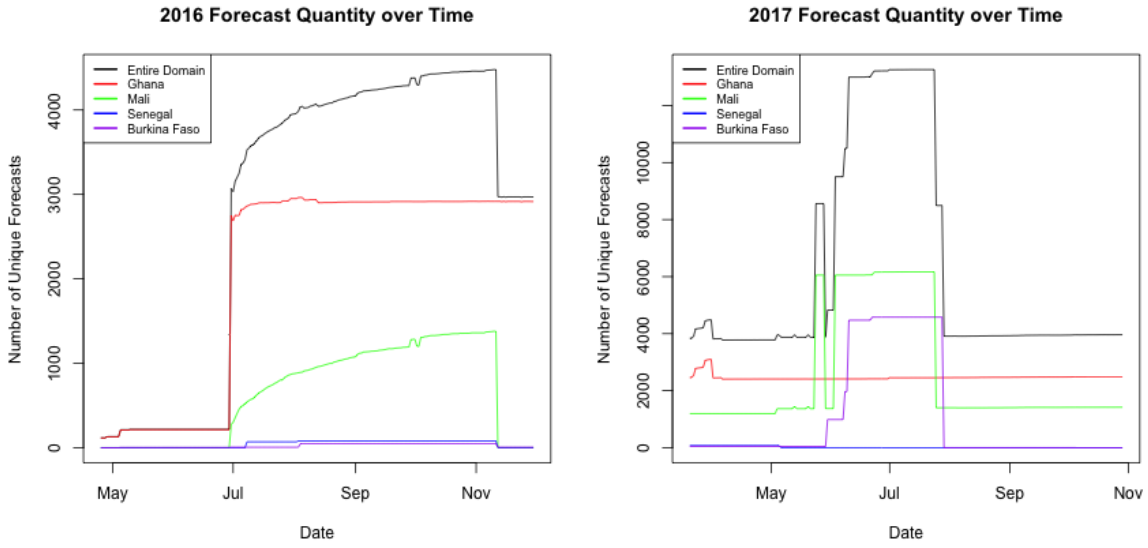


Figure 1: The number of unique forecasts over the wet season for 2016 and 2017. Note that the number of forecast subscribers increases dramatically during the rainy season in all countries under evaluation. See Plate A in section A1 for the locations of the stations on May 1 and July 1 of each year.

2) Objective and Main Evaluation Questions

The objective of this analysis is to provide an independent evaluation of forecast quality for daily rainfall predictions issued by Ignitia at lead times of 1- and 2-days out. Ignitia has performed their own evaluations based on the RFEv2 satellite products, and reported an 84% accuracy of their forecasts. They additionally state that this is more than double the accuracy available by other forecast providers (unnamed).

The main question behind this analysis is: *How skillful are Ignitia's forecasts compared to other available forecast model(s), based on actual in situ observation data?* Related to that, and in light of Ignitia's own evaluation: *Can their claim of 84% accuracy be validated?* We wish to note upfront that this question is ill posed given that the format of the forecasts is probabilistic, and the verifying observations are binary (i.e. rainy day or dry day).

3) Evaluation Methodologies

The two related questions stated in Section 2, are elaborated as follows:

Q1. How skillful are Ignitia's weather forecasts for rainfall?

- in comparison to observed rainfall estimates
- in comparison to raw output from NOAA's weather forecast model (GFS)

Q2. To what extent are Ignitia's statements based on their own verification analysis[3] true?

- Is the statement of 84% accuracy fair?
- Is the statement that Ignitia's forecasts provide more than double the skill of available forecasts fair?

As a by-product of addressing the first question, which is the primary intent of this analysis, the second question posed by USAID can also be answered. We do note, however, that we were unable to obtain a competitive forecast available over the study period, which would allow us to fully examine the veracity of Ignitia's claim regarding competitive forecasts. Instead, Ignitia's forecast performance was compared to a reference standard forecast and a simple alternative forecast computed using the widely used and adapted Global Ensemble Forecasting System (GFS), which also happens to be an input to the Ignitia forecasts. To justify the production of issued forecasts a provider should demonstrate skill over computer model predictions and other available forecasts (Le Blancq and Johnson 2000).

The most appropriate verification metrics, given the forecast format as well as the type of event being forecast, required additional specificity to address skill and to recast the issue of 'accuracy', which is the term used by Ignitia. In line with the industry standards, the analysis addresses the first question in two parts, which encompass the dual nature of probabilistic forecasts for specific categories or events:

Q1a: *Reliability*: Does rainfall occur more frequently when confidence that it will rain is high?

Q1b: *Discrimination*: When it rains do the forecasts indicate higher confidence in rainfall occurring compared to when it is dry

To explore quality of the forecasts it is necessary to review the outputs of Q1a and Q1b. It is not appropriate, and may be misleading, to use only the results from either question in isolation. Worse still is to address only a single aspect of these metrics for a particular forecast category. Given that few are familiar with the terminology used in the verification of probabilistic forecasts, some elaboration of these metrics is given in Section 3.1.

The metrics used to address the questions above encompass some of the verification aspects addressed in the ISKA White Paper (2017), but take a more comprehensive approach, as described in the IRI Description of Verification Scores[4].

In this section, we define deterministic and probabilistic forecasts in Section 3.1 and illustrate the interpretation of Ignitia's categories, based on their definitions, to form a probabilistic

[3] http://www.ignitia.se/our-story

[4] http://iri.columbia.edu/wp-content/uploads/2013/07/scoredescriptions.pdf

forecast. Section 3.2 discusses the two alternative forecasts used along with how they can be used to quantify the relative skill of Ignitia's forecast. An overview of the two observational datasets is given in Section 3.3. In 3.4 we provide definitions and interpretations of the methods used. Finally, Section 3.5 provides broad outline of the evaluation procedure.

3.1) Ignitia Forecast Data as Probabilistic Forecast

Ignitia translates their probabilistic precipitation forecast into four broad probability bins. We follow Annex 1 in their "Validation of forecast accuracy – project description", which collects the 25 categories outlined into four probabilities levels. We use probability ranges defined on page 7 of Ignitia's ISKA white paper, published April 30, 2017. For ease of communication in this report, we assigned the four levels to the letters A-D as follows:

IRI Code	Ignitia Description	Probability of Rain	Probability for Bin
A	Dry	<10% Chance of Rain	0%
B	Likely Dry	10-49% chance of rain	30%
C	Rain Likely	50-80% chance of rain	60%
D	High Chance of Rain	>80% chance of rain	90%

3.2) Alternative Forecasts

For a comparative measure of quality and potential value of a forecast, it is important to assess the forecast of interest relative to alternative forecasts that are available to potential users (McBride and Ebert 2000). Two different alternative forecasts were used in this verification.

First, a persistence forecast is used to determine a minimal acceptable baseline performance (Mittermaier 2008). Persistence forecasts follow the simple algorithm: use the weather today to predict the weather tomorrow (Ebert et al. 2003, Robertson et al. 2004). In the context of precipitation forecasting, a persistence forecast predicts that it will rain tomorrow if it rains today. Since the persistence forecast for precipitation makes a statement without uncertainty, it is deterministic, which translated into a probabilistic framework means that the forecast has either a 0% or 100% probability for rain.

The second alternative forecast is derived from Global Ensemble Forecasting System (GFS) from NOAA's National Centers for Environmental Prediction (NCEP). The GFS is a forecast model that drives the lateral boundaries of the higher-resolution regional WRF model used by IKSA predictions over West Africa (IKSA White Paper, 2017). The underlying weather model was run for each day from 1985 to present in NOAA's 2nd-generation Reforecast Project (Hamill et al. 2013). Using the daily 11-member ensembles, we are able to construct simple probabilistic

forecasts by determining the number of ensemble members that forecast 'significant rainfall'. To define 'significant rainfall', we use Ignitia's threshold of 2 mm. While the GFS is one of the most widely used global forecast models, it is to be considered raw model output and thus it does not have the status as a true alternative forecast. However, given the long history and large ensemble generated through the Reforecast Project (Hamill et al. 2013), most users of this product do perform some form of post-processing. For example The Weather Channel forecasts, available worldwide and produced by IBM, contain extensive post-processing. Similarly, BBC forecasts are to some extent post-processed as they use the UK MetOffice model, which employs post processing[5] (communication with BBC forecaster). Thus, while likely to be an improvement on the persistence forecast method, the GFS should be seen as a baseline forecast.

3.3) Observations

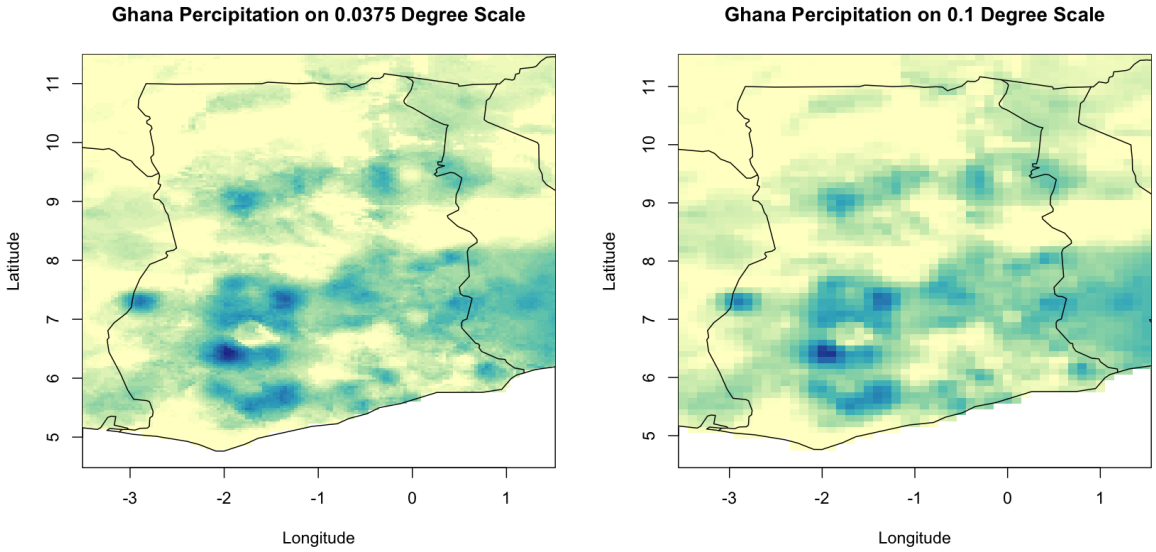
Two distinct observational datasets for rainfall are used in the verification: Rainfall Estimate of NOAA's Climate Prediction Center version 2.0 (RFEv2.0) and Enhancing National Climate Services (ENACTS).

RFEv2.0 was supplied by Ignitia, and confirmed for quality and consistency by IRI. Following the description provided by the ISKA White Paper (2017) this dataset uses microwave sensor (SSM/I and AMSU) and geostationary IR-channel data (NOAA, 2012). It is not a real-time product; the data is processed after the 24-hour collection period and calibrated against ground station measurements of rainfall in the region that report to the World Meteorological Organization's Global Telecommunication System (GTS) in real time. This means that nearby to a GTS ground station, the data is more representative of the real rainfall. However, away from the station the data is subject to the deficiencies known to satellite rainfall observations in the tropics, such as exaggerated spatial extent of cumulonimbus rainfall and difficulties in capturing warm cloud rainfall during the monsoon season. Additionally, RFEv2.0 is produced on a 0.1x0.1 degree grid which results in grid boxes of approximately 10km x 10km. Since a grid box value represents the average rain over the entire area of the grid box, very small-scale patterns of precipitation - such as that associated with convective fields - cannot be precisely located. Such localized detail in convective rainfall is nearly impossible to predict, unless it is associated with some geographic feature such as topography, large bodies of water or coastlines.

The second observational dataset, which was only used for Ghana in this assessment, is the ENACTS data (Enhancing National Climate Services)[6]. These datasets are produced within the national meteorological services of the countries, with methodology, software, training, and computing infrastructure for dissemination provided by IRI. In short, the motivation for the ENACTS process was established to address the under-utilization of regularly reporting rainfall stations in Ghana and other countries. Many countries have more data than is provided to the

[5] https://www.wmo.int/pages/prog/arep/wwrp/new/wwosc/documents/ACB-WWOSC2014-v2.pdf
[6] http://www.wmo.int/pages/prog/wcp/wcasp/RA-I/documents/reference_materials/climate_data_monitoring/CDMS/ENACTS-climate-data-we-links.pdf

GTS, and this data can be used for improved monitoring and historical analysis. The first step is to quality control the station data, and fill gaps, where necessary and/or possible. The second step is to merge those station data with satellite data (TAMSAT and TRMM), which is continuous in space and time, but has spatial and magnitude biases as indicated above. The ENACTS grid is 0.0375x0.0375 degree (or 4km x 4km), which allows for more precise quantification of rainfall at a point location. The ENACTS observed data is likely to be more representative of what the subscribers experience, not just because it is at higher resolution, but because much more station data is used to calibrate the satellite fields. Below is an example of the 4km version of ENACTS, next to an upscaled 10km version of the same data. For the purpose of this analysis, the upscaled 10km version was the only version used.



3.4) Forecast Verification Methodologies

How "good" a forecast is depends primarily on three components, its consistency, quality, and value (Murphy 1993). Forecasts that are said to be "consistent" are those which produce a true indication of what the forecaster believes will happen. Alternatively, "quality" usually describes a mathematical relationship between the forecast and what is actually observed. Forecasts that are thought to produce "value" are those that bring about benefits in areas such as economics and social actions. Trying to determine if a probabilistic forecast is "correct" is less intuitive than that of deterministic forecasts, given the nature of the forecasts, but there are some common practices to do this, including discrimination and reliability.

Reliability: Reliable forecasts are ones in which the probabilities mean what they say. For example, when a rainy day is forecast as 60% likely. Sixty percent of the time that specific forecast is issued, a rainy day should occur.

Discrimination: Forecasts that can discriminate events provide more confident forecasts for a rainy day on days that it rains than for days with no rain.

These two terms may seem equivalent, but rather they are complementary. One could issue forecasts that gave the historical odds of rainfall in a given month, or over a season. Those forecasts would be *reliable*, since they were designed to give a probability for rain that is consistent with its frequency of occurrence. However, since the same forecast would be issued all the time, there would be no *discrimination* in forecast confidence on a wet day versus a dry day. These two aspects of forecast quality are both relevant to accuracy, or more appropriately 'hit rate', but give a more complete picture. The 'Hit Rate', especially if focused on only one of several categories, as used by Ignitia, could miss important elements of forecast quality. If one were to forecast that every day were to have rain. The hit rate would be perfect for "rainy day forecasts" because every day it rained it was forecast to rain. However, the discrimination would be poor because all the forecasts would have been the same, and the reliability would likely be bad (unless is actually did rain every day), because the confidence would be high (>80%) even though [in West Africa] a rainy day was not observed 80% of the time. Discrimination accounts not just for the hit rate, but also penalizes for false alarms.

Discrimination looks to answer the question "Do the forecasts differ given different outcomes?". In other words, what is the ability to distinguish between forecast probabilities that lead to a specific event of interest from forecast probabilities that do not lead to that event. Statistically speaking, this implies that the less overlap that is present for the forecasts' distributions, the easier it is to distinguish events, and therefore the higher the discrimination. This is one of the most basic attributes of probabilistic forecasts, and cannot be improved by statistical recalibrations.

To measure discrimination, the relative operating characteristics (ROC; sometimes also referred to as the receiver operating characteristics) graph is generated, and the area beneath this curve is calculated. This requires binary calculations with each forecast category having separate results (given there are more than 2 forecasts). In the case of this analysis, 30%, 60%, and 90% chance categories were analyzed. The outcomes will vary from 0% to 100%, where a score of 50% is equivalent to guessing. Below 50% show that forecasts can be discriminated against but in the wrong tendency. Scores that are above 50% show improved usefulness up to 100% where the forecasts give perfect discrimination between events.

To construct a ROC curve, we proceed iteratively (for more details see Mason and Graham, 1999). Begin by choosing a probability threshold p0. Then, all forecasts that predict the outcome at or about the probability p0 are said to "warn" and the forecasts with probability below p0 are said to "not warn". Then, assuming that our outcome of interest is precipitation, we can make a 2x2 contingency table of the form:

	Warning	No Warning
Rain	a	b
No Rain	c	d

From here the hit rate and false alarm rates are calculated. The hit rate (HR) looks at the number of hits (number of events selected) within a dataset and compares this with the number of events within this dataset (see equation below). The false alarm rate (FAR) compared the number of false alarms (number of non-events selected incorrectly) with the total number of non-events within the dataset (see equation below).

$$\text{Hit Rate} = (\#\text{ hits})/(\#\text{ events}) = a/(a+b)$$
$$\text{False Alarm Rate} = (\#\text{ false alarms})/(\#\text{ non-events}) = c/(c+d)$$

The HR and FAR are then plotted against each other forming one point of the ROC curve where the point (x,y) = (FAR,HR). This whole procedure of setting a p0 warning level, making a contingency table, and calculating the HR and FAR is repeated at various p0 levels. In the case of Ignitia's forecast (See Table in Sec 3.1), we are only able to compute three points: warning for forecast of category D, warning for category C and D, and warning for categories B, C, and D. An example ROC curve is shown in Figure 2 below.

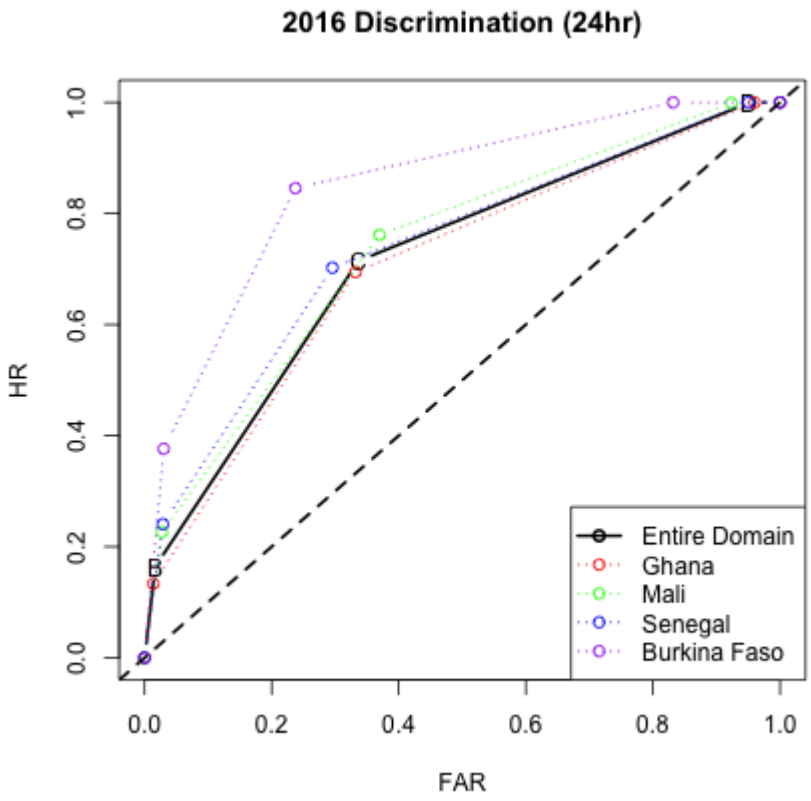


Figure 2: ROC curves from the study. The False Alarm Rate (FAR) is given on the horizontal axis, and the Hite Rate (HR) is given on the vertical axis. The black dashed line indicates that the forecast system cannot discriminate between hits and false alarms, and represents the line of no skill. Good discrimination is indicated by curves that reach into the upper left corner, indicating a high hit rate and few false alarms. The solid black line is the ROC curve for the entire domain. The points (B,C,D) represent the warning probability set at (30%, 60%, and 90%) respectively (see table in Sec 3.1). For the individual countries, circles are used instead of the probability labels (B,C,D)

When the HR and FAR are equal, the forecast offers no useful information since it cannot discriminate, or beat a random forecast. When the HR is higher than the FAR, the forecasts can discriminate events. On the ROC curves shown in the results section, the best outcome possible is when HR is much larger than FAR, causing the graph to be steep on the left, and have a shallow angle on the right. Typically, forecasts will have better discrimination for the more confident categories (i.e. higher probabilities).

Since the ROC curves can be difficult to interpret or compare by eye, we summarize the discrimination of a forecast by calculating the area under the curve as a metric of separation from the unity line (Hogan and Mason 2012).

Reliability is used to assess, "Are the probabilities given to forecast an event appropriate to the associated frequency with which the event occurs?" It the other main tool for determining the quality of a forecast (Murphy 1973). The reliability scores range from 0.0 (perfectly reliable forecasts) to 1.0 (perfectly bad forecasts). When graphing this, the best outcome will be at 45-degrees, following a 1:1 ratio between the forecast probability (%) and the observed relative frequency (%) of this event. Lines that that are more horizontal are said to be "over-confident", while lines that are steeper than 45-degrees are said to be "under-confident".

Both over-confident and under-confident forecasts are not desirable. An overconfident probabilistic rainfall forecast is one that attaches higher probabilities to rainfall occurring compared to what is observed. For example, if a forecast system has issued 100 forecasts over time that indicated 90% probability of a rainy day, but it only rained on 60 of those 100 days, then those forecasts would have been overconfident. One would need to examine the forecasts, at all levels of confidence, in this manner to draw conclusions about the reliability quality of the forecast system. Both Ignitia and GFS 2016 24-hour forecasts for Rain Likely are examples of overconfident forecasts. This can be seen in Figure 3. For example, GFS indicates for Rain Likely (which translates into approximately a 60% chance of rain) and only 30% of the time (as noted by the position on the Y-axis) does it actually rain following that forecast. It should be noted that Ignitia's 2016 24-hour forecasts for Rain Likely are much better than GFS, however only 40% of the time these forecasts are correct.

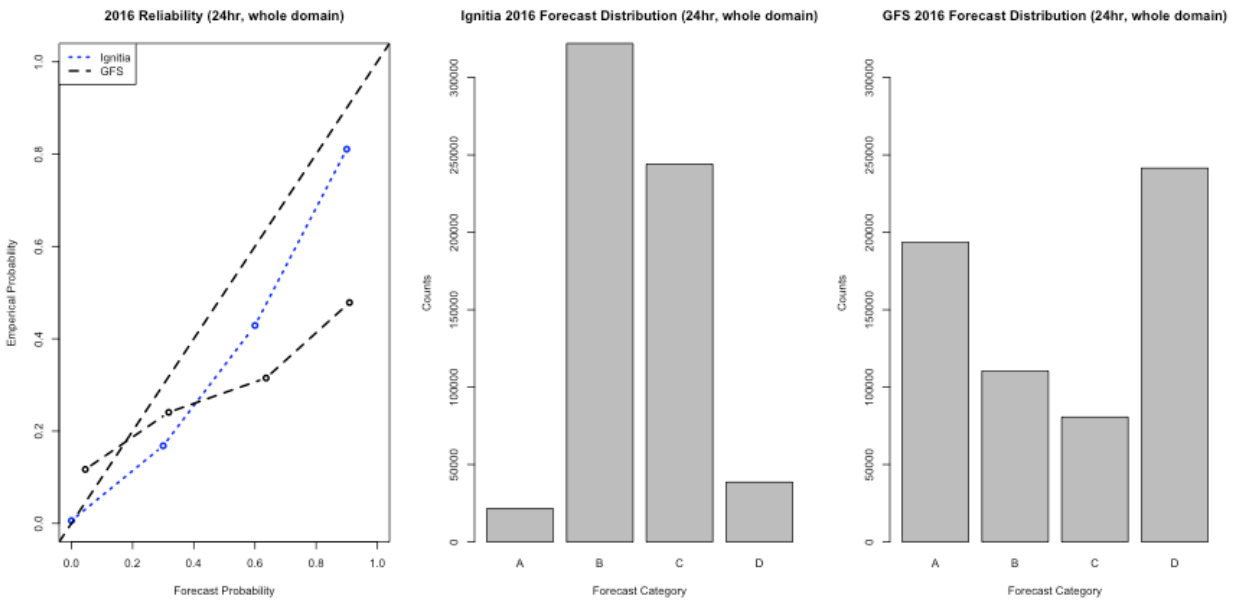


Figure 3: Reliability curves and forecast distributions from the study. The Forecast Probability for a 'Rainy Day' is given on the horizontal axis, and the associated Observed Frequency of Occurrence is given on the vertical axis. The black dashed line at 45-degrees indicates perfect reliability.

$$reliability = \frac{1}{n}\sum_{k=1}^{d} n_k(\bar{p}_k - \bar{y}_k)^2$$

where there are d discrete probability values, n_k is the number of forecasts for the K^{th} probability value (p_k), and y_k is the observed relative frequency for that value.

3.5) Evaluation Procedure Outline

For the evaluation, the open source language R was used. R is a powerful statistical analysis tool that allows for comprehensive and replicable analyses of large weather datasets. It is important to note that a day of rainfall is defined using Ignitia's definition of 'any day in which 2 mm of rainfall is observed in an observational product'.

An overview of the analysis workflow for a single year is as follows:

1. The Ignitia forecast data was trimmed such that there was only one forecast at each lat/lon coordinate. Removing duplicate forecasts prevents biased forecast assessment as hits/misses for a single location will only be counted once.
2. For each day a forecast was issued (recall that the range of forecast dates differs between 2016 and 2017), we compiled for:
    - The Ignitia forecast (24 hour forecasts-for that calendar day, 48 hour forecasts-for the next calendar day)
    - The persistence alternative forecast

    - The GFS-based alternative forecast
    - The outcome for each forecast given RFE2
    - The outcome for each forecast given ENACTS (Ghana only)
3. For each of the three forecasts (Ignitia, persistence, GFS-based) and the RFE2 outcome, we evaluated the discrimination using the ROC curve methods and the reliability using a reliability diagram for:
    - The entire forecast domain
    - Each of the 4 countries in the Ignitia 2016 white paper individually (Senegal, Ghana, Mali, Burkina Faso).
4. For each of the three forecasts (Ignitia, persistence, GFS-based) and the ENACTS outcome, we evaluated the discrimination using the ROC curve methods and the reliability using a reliability diagram for:
    - Ghana 2016

The following table contains a summary of reliability and discrimination analyses conducted with RFE validation dataset. Each row corresponds to a set of analyses run:

Country	**Years**	**Forecast Range**	**Alternative Forecast**
Entire Domain	2016/2017	24hr/48hr	Persistence/GFS
Senegal	2016/2017	24hr/48hr	Persistence/GFS
Ghana	2016/2017	24hr/48hr	Persistence/GFS
Mali	2016/2017	24hr/48hr	Persistence/GFS
Burkina Faso	2016/2017	24hr/48hr	Persistence/GFS

Additionally, the following table contains a summary of reliability and discrimination analyses conducted with the ENACTS validation dataset.

Country	**Years**	**Forecast Range**	**Alternative Forecast**
Ghana	2016	24hr/48hr	Persistence/GFS

4) Limitations to the evaluation

Limitation 1: There is a lack of consistency between verification and forecast datasets. The verification products used are 00:00 - 23:59 and the 24-hour forecast is for 6:00 - 5:59 the next day. This timing discrepancy should be minor for West Africa, but could potentially shift the

"rain day" if the rain occurred during the time gap between observation and forecast target day.

Limitation 2: We could not get access to an archive of a reasonable alternative forecast, i.e. one that would actually be available to the subscribers. Such forecasts exist, but we were not able to obtain the past forecasts in the short timeframe of the project. Our GFS-based forecasts are not equivalent to a user-oriented forecast product. Examples of forecast products that are both available to users and can be defined as 'user-oriented' include those that are produced by the National Meteorological Services, such as the Ghana Meteorological Agency, International news and media outlets, such as weather.com and BBC, and other private sector companies that produce forecasts on comparable spatiotemporal scales. These forecasts are expected to be improvements on the GFS output data.

5) Findings and conclusions

5.1) Summary of Findings:

Q1. How skillful are Ignitia's weather forecasts for rainfall?
- *in comparison to observed rainfall estimates*
- *in comparison to raw output from NOAA's weather forecast model (GFS)*

Finding 1: The reliability of the Ignitia rainfall forecasts are superior to those of the raw GFS model baseline. Both sets of forecasts are over-confident, but this problem is much worse for the GFS. The observed frequency-of-occurrence of rainfall for each forecast category were slightly below the target probabilities. However, since the categories represent a range of probabilities, the reliability curves contain inherent uncertainty, and may be more or less reliable than shown. The differences between the 24-hr and 48-hr forecasts seem small in 2016; in 2017 the 48-hr forecast is noticeably less reliable (Plates 1 & 2).

Finding 2: The distribution of forecast confidence is dramatically different between the Ignitia forecasts and those from GFS. The Ignitia forecasts rarely issue a confident (high probability) forecast for a rain day, or from a no-rain day. The GFS forecasts, on the other hand indicate these high-probability forecasts much more frequently. This finding is not surprising. The reduction of confident forecasts is common to calibrated forecasts. In addition, it seems appropriate given the improved probabilistic reliability in Ignitia compared to GFS. (Plates 1 & 2)

Finding 3: Over the region as a whole, the discrimination quality of the Ignitia forecast is not significantly different from that of GFS. There are some differences across the countries, but these do not seem to be systematic across years. (Tables 1 & 2)

Finding 4: There is a discernible impact to the reliability using the ENACTS data for Ghana. Particularly for the 24-hr forecasts in 2016, the reliability is nearly perfect for the Ignitia

forecasts, but remains very poor and over-confident for GFS. The effect on discrimination is to lower it slightly for both Ignitia and the GFS baseline. (Plate 4 and Table 3)

Q2. To what extent are Ignitia's statements based on their own verification analysis[7] true?

- Is the statement of 84% accuracy fair?

- Is the statement that Ignitia's forecasts provide more than double the skill of available forecasts fair?

Finding 5: The statement of Ignitia regarding 84% accuracy is not correct. Their claim is based primarily on the 'Hit Rate' from the most confident category for a rainy day. (Plate 1 or 2). In other words, when they forecast rain is very likely (>80% probability), about 85% of the time the rain day occurs. The problem is that those forecasts are a small percentage of the forecasts issued. Further, as noted in the literature, the results of reliability for lesser-used categories must be regarded with some caution, and the results of reliability of categories with the largest amount of forecasts issued are regarded to be more robust results. When considered in isolation, a level of reliability or accuracy, for a limited category, does not indicate the skill of the forecast system (Le Blancq and Johnson 2000).

Finding 6: The statement that other [available] forecasts are only 39% accurate is also ill-posed for the same reason as stated in Finding 5. Additionally, we were not really able to support even that part of the claim since we were unable to obtain actual competitive forecasts for the study period. However, we do acknowledge that for the high-probability category, the GFS reliability showed only about a 40% occurrence of rain days.

5.2) Reliability: Does rainfall occur more frequently when confidence that it will rain is high?

As described in section 3.1, a robust, intuitive, and accessible method for determining the reliability of a forecast is a reliability diagram. We constructed reliability diagrams for 2016 and 2017 forecasts over the entire Ignitia domain for both the 24- and 48-hour forecasts. These diagrams are shown in Plate 1.

It is clear that the Ignitia forecasts add value over the raw model output from GFS, as indicated by the improved reliability. This is true of most forecast post-processing that can account for systematic biases. As one might expect with rainfall forecasts, the majority of the forecasts fall in the middle two categories corresponding to 30% and 60% rain for both 2016 and 2017 in all four countries. The separate reliability analyses, performed for each of the four countries of interest, yielded results that were more or less consistent with the aggregated results over the entire domain. Other forecast providers (e.g. BBC, Weather.com) do issue forecasts through

[7] http://www.ignitia.se/our-story

smart phones in the countries of this study. However, as far as we can tell, the forecasts are for towns and cities, and thus not as location-specific as Ignitia's service.

5.3) Discrimination: When it rains do the forecasts indicate higher confidence in rainfall occurring compared to when it is dry?

The relative operating characteristics (ROC) curve is used to visually determine the discrimination of a forecast. To quickly and quantitatively compare the discrimination of two forecasts, we calculated the ROC area under the curve (ROC AUC). As stated in finding 3, the discrimination of Ignitia's model when using the RFEv2 validation set was generally similar to the GFS-based method. This indicates that, on average, Ignitia adds minimal discrimination skill to the raw GFS output. As described in Sec 3.4, discrimination seeks to assess whether different outcomes (e.g. rain day versus dry day) are accompanied by notably different forecast probabilities. The calibration of the Ignitia forecasts have greatly reduced the number of forecasts issued in the outer, more confident categories, which means that the forecasts differ little between the different outcomes. The GFS, on the other hand readily yields very confident – in fact over-confident – forecasts. This trade-off between discrimination and reliability may be a realistic calibration. Given that the discrimination is comparable, but the reliability is better in Ignitia, these metrics taken together indicate that Ignitia has added value to GFS.

In addition to the GFS alternative, the ROC diagrams contain a point with the hit rate and false alarm rate for a persistence forecast. Both the Ignitia and GFS-based significantly outperform the persistence model. We note that the persistence forecast does exhibit positive skill as it appears above the unity line.

We summarize the ROC area under curve results in Tables 1 and 2. These AUCs are calculated from the ROC curves used for this analysis, which are presented in Plates 3 and 4. We also break down the ROC calculation by country to determine if there was any systematic bias in the forecast for specific countries. We do not find any evidence of such bias and present these figures in Annex 1 in Plates A2-A5 alongside the corresponding reliability diagrams.

5.4) Effect of Verification dataset

We found that depending on which verification dataset is used to define observed daily rainfall, reliability varies. For the Ghana component of this study for 2016 we repeat the full analysis on the 2016 Ghana forecasts using the 10x10km ENACTS, the resolution of Ignitia's forecast, and present the results in Plate 4.

For reliability, we find that Ignitia's forecast was robust to changes in validation forecast. The reliability improved under ENACTS validation relative to RFEv2 validation.

For discrimination, Ignitia's forecast and the GFS exhibit slightly lower discrimination with the ENACTS data when compared to the verification with the RFE. These results suggest a deeper look into the sensitivity of the forecast performance to the verification used.

Plate 1: Entire Domain Reliability Diagrams 24 hour forecast (RFE validation)

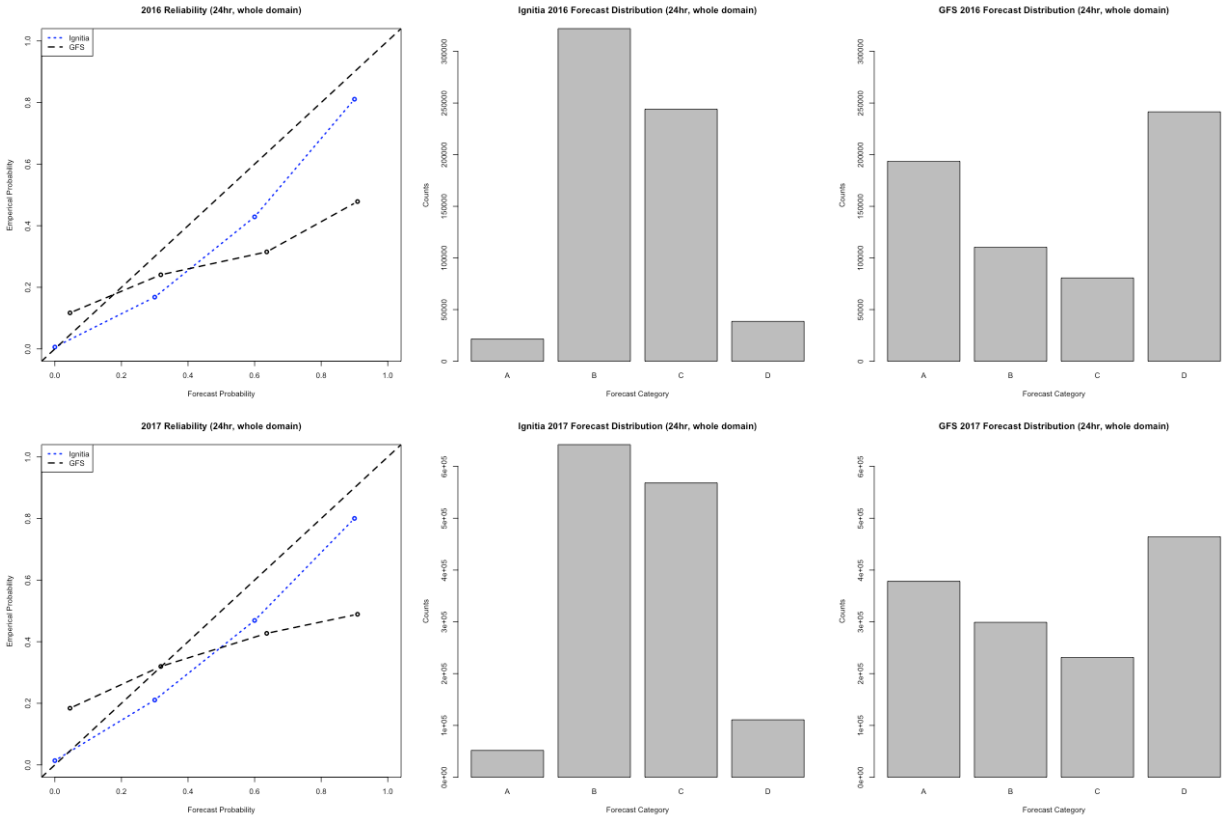


Plate 2: Entire Domain Reliability Diagrams 48 hour forecast (RFE validation)

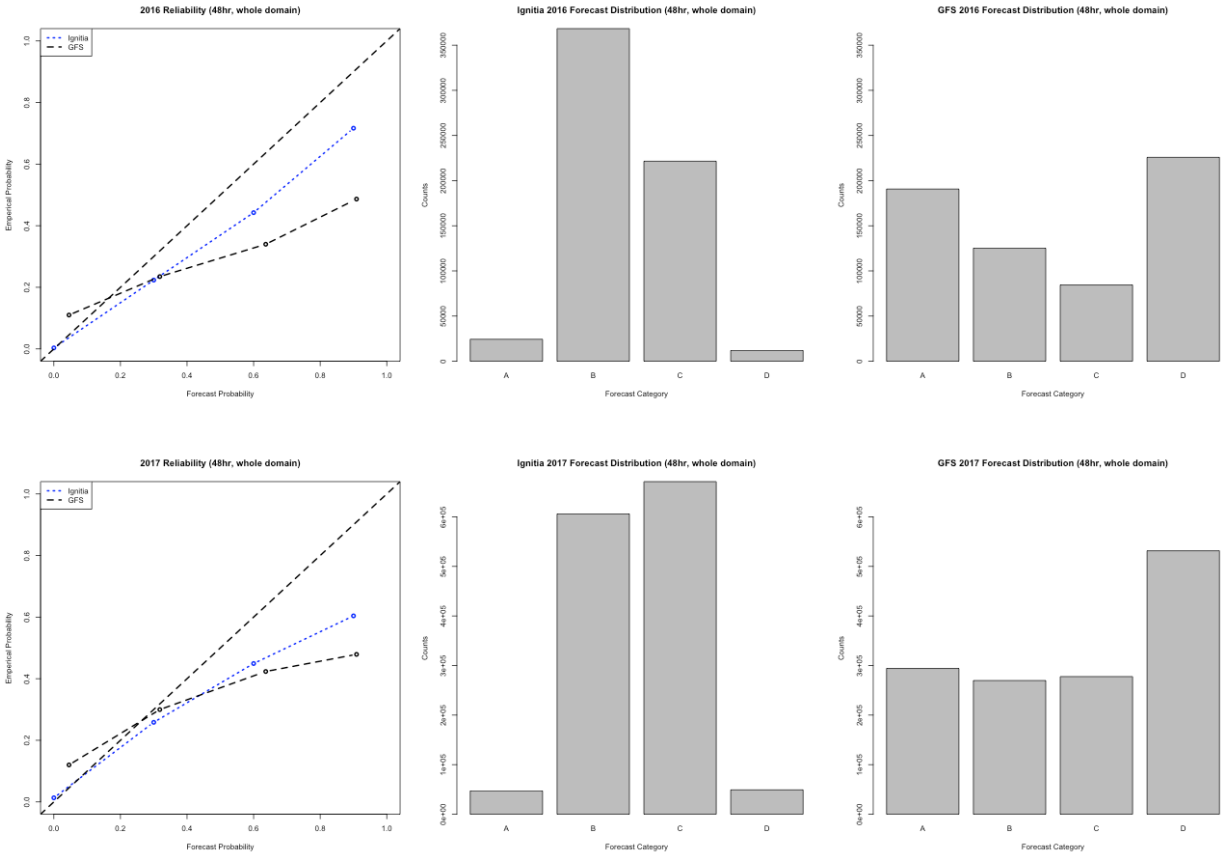


Table 1: ROC Area under curve using the RFE validation dataset for 2016.

ROC AUC	Ignitia 2016 (24hr)	GFS 2016 (24hr)	Ignitia 2016 (48hr)	GFS 2016 (48hr)
Total	0.717	0.717	0.654	0.716
Ghana	0.705	0.700	0.634	0.688
Mali	0.736	0.777	0.683	0.776
Senegal	0.736	0.837	0.661	0.84
Burkina Faso	0.849	0.839	0.747	0.809

Table 2: ROC Area under curve using the RFE validation dataset for 2017

ROC AUC	Ignitia 2017 (24hr)	GFS 2017 (24hr)	Ignitia 2017 (48hr)	GFS 2017 (48hr)
Total	0.710	0.656	0.632	0.664
Ghana	0.697	0.652	0.639	0.657
Mali	0.720	0.701	0.628	0.693
Senegal	0.742	0.491	0.495	0.495
Burkina Faso	0.680	0.570	0.564	0.550

Table 3: ROC Area under curve using the ENACTS 100km upscaled validation dataset for Ghana in 2016. The 24 hour discrimination of the two forecasts is similar, but the GFS significantly outperforms Ignitia in the 48 hour forecast

ROC AUC	Ignitia 2016 (24hr)	GFS 2016 (24hr)	Ignitia 2016 (48hr)	GFS 2016 (48hr)
Ghana	0.685	0.690	0.613	0.666

Plate 3: Entire Domain Discrimination Figures (RFE validation)

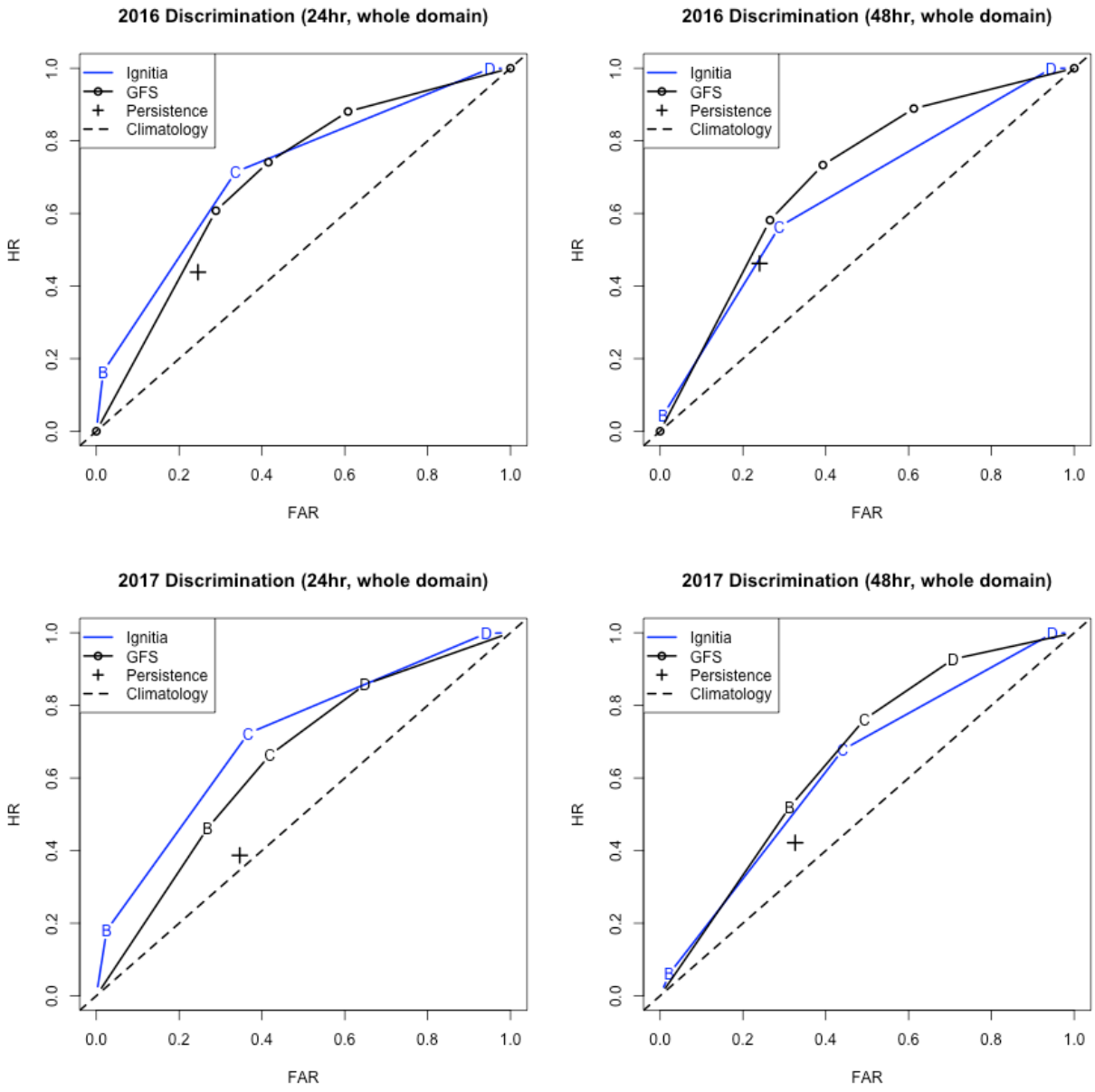
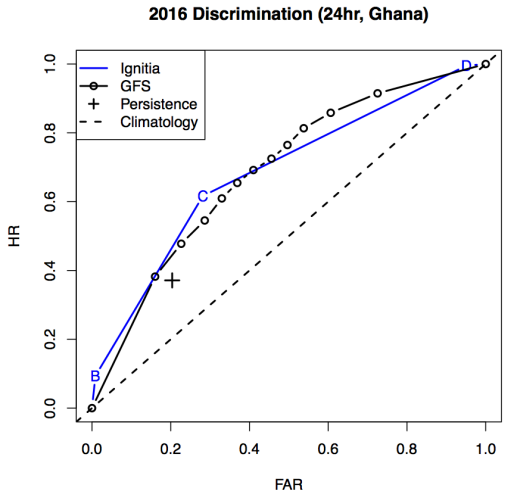
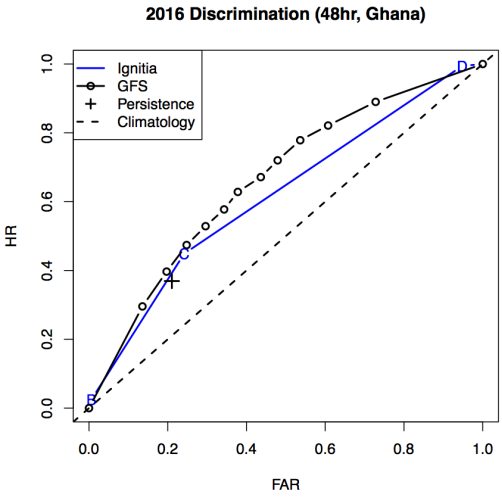
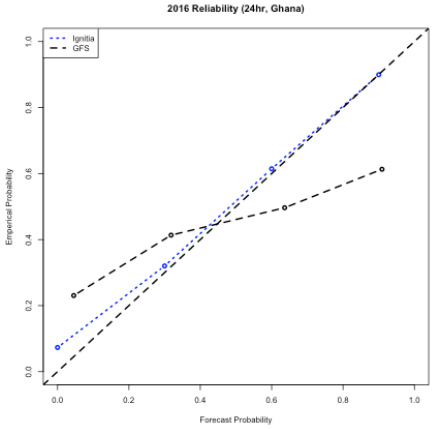
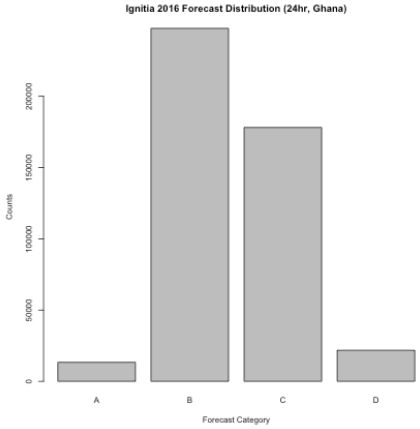


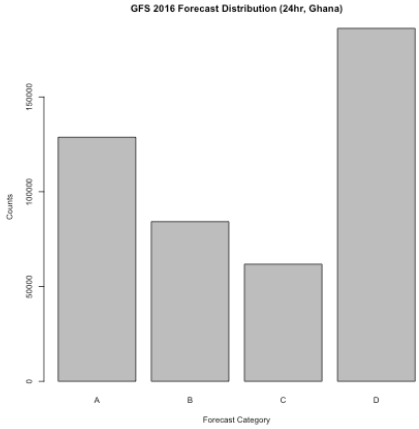
Plate 4: Upscaled Enacts Figures (Ghana 2016)

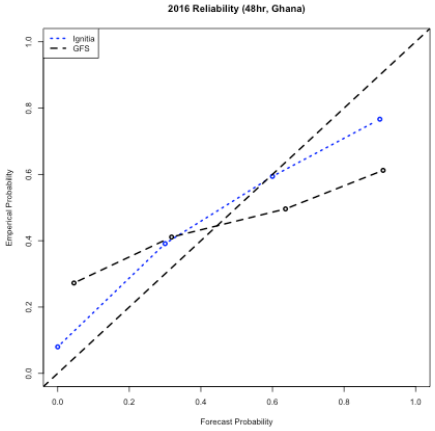


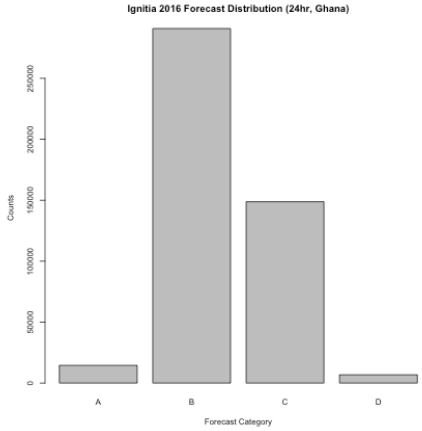


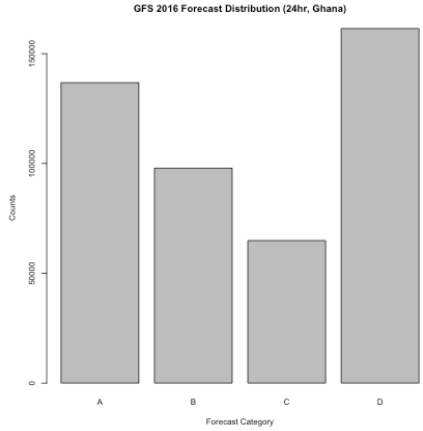






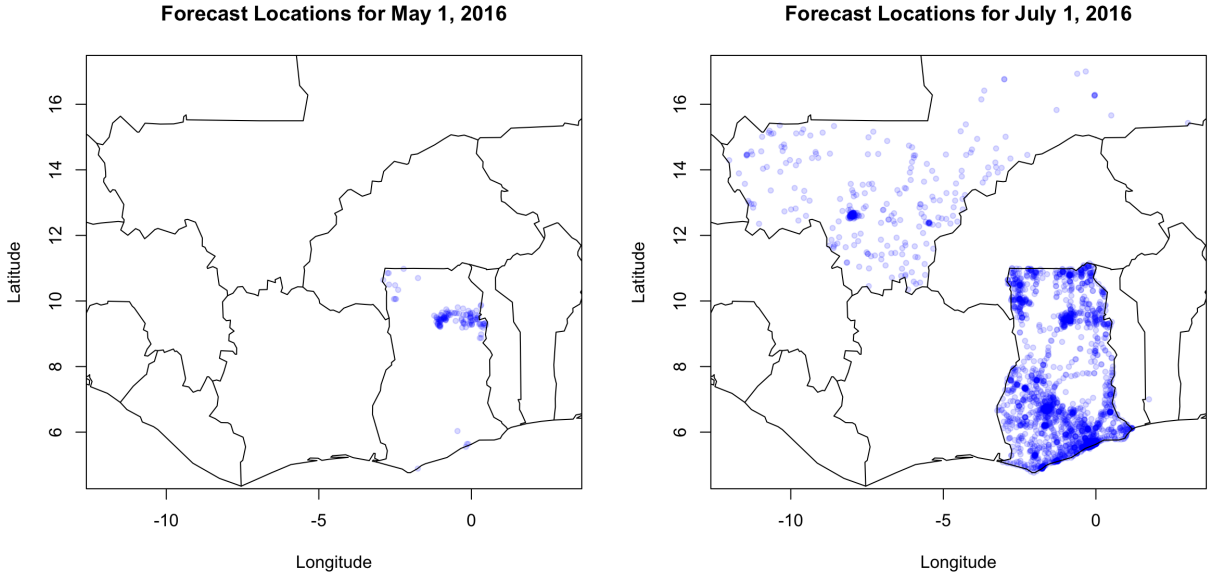




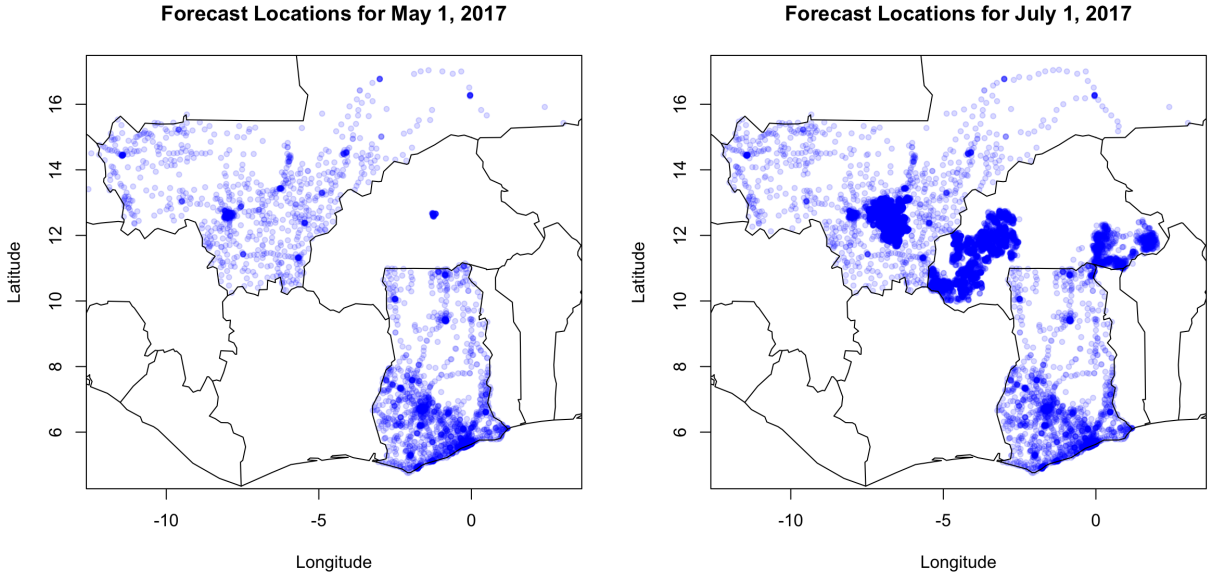


A1) Additional Figures

Plate A1: Spatial Distribution of Ignitia forecasts



Unique forecast locations at the beginning of the wet season (May) and the approximate peak subscriber time period (July) for 2016



Unique forecast locations at the beginning of the wet season (May) and the approximate peak subscriber time period (July) for 2017

Plate A2: Discrimination/Reliability by country for 2016/24hr:

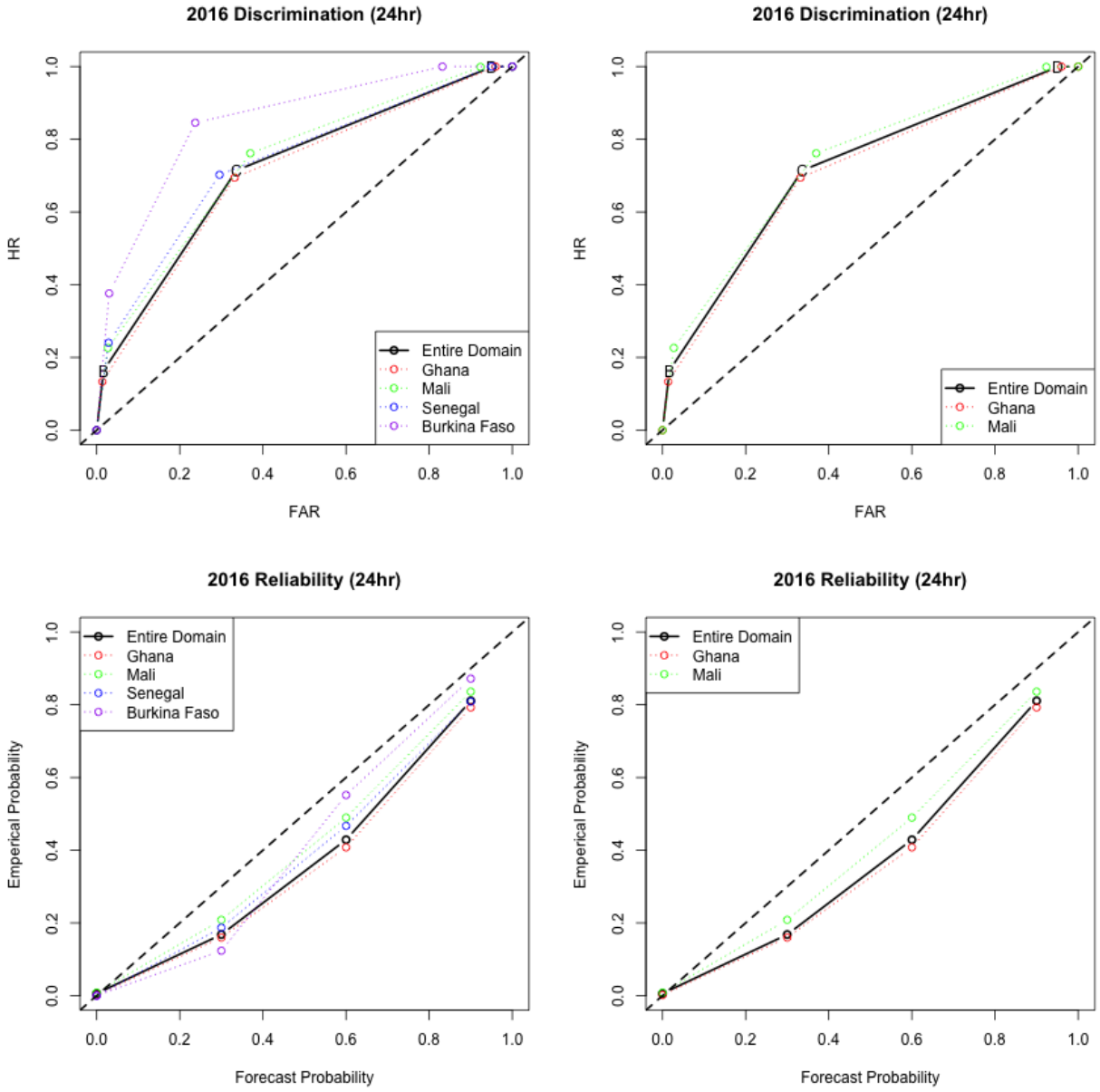


Plate A3: Discrimination/Reliability by country for 2016/48hr:

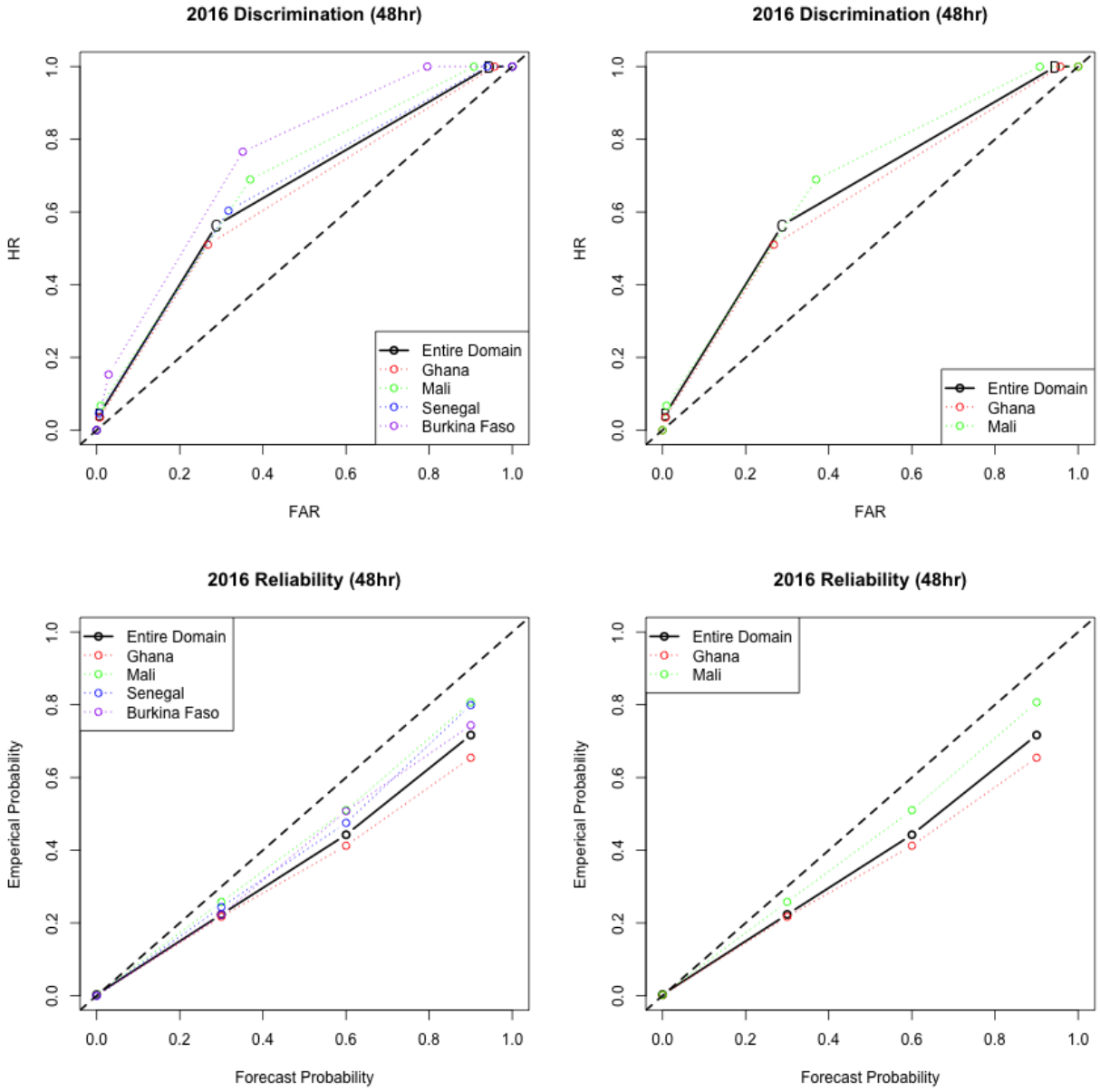


Plate A4: Discrimination/Reliability by country for 2017/24hr:

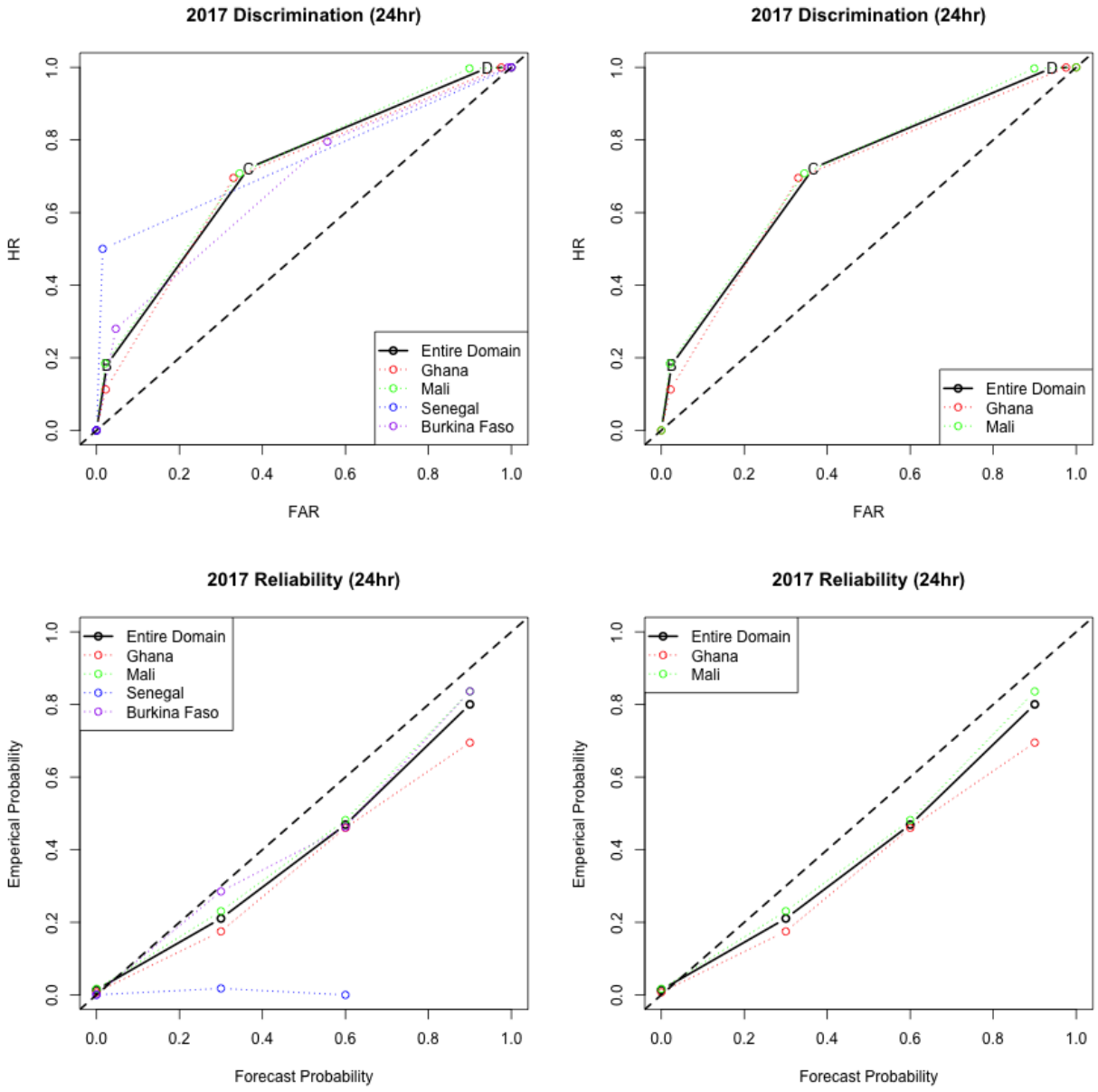
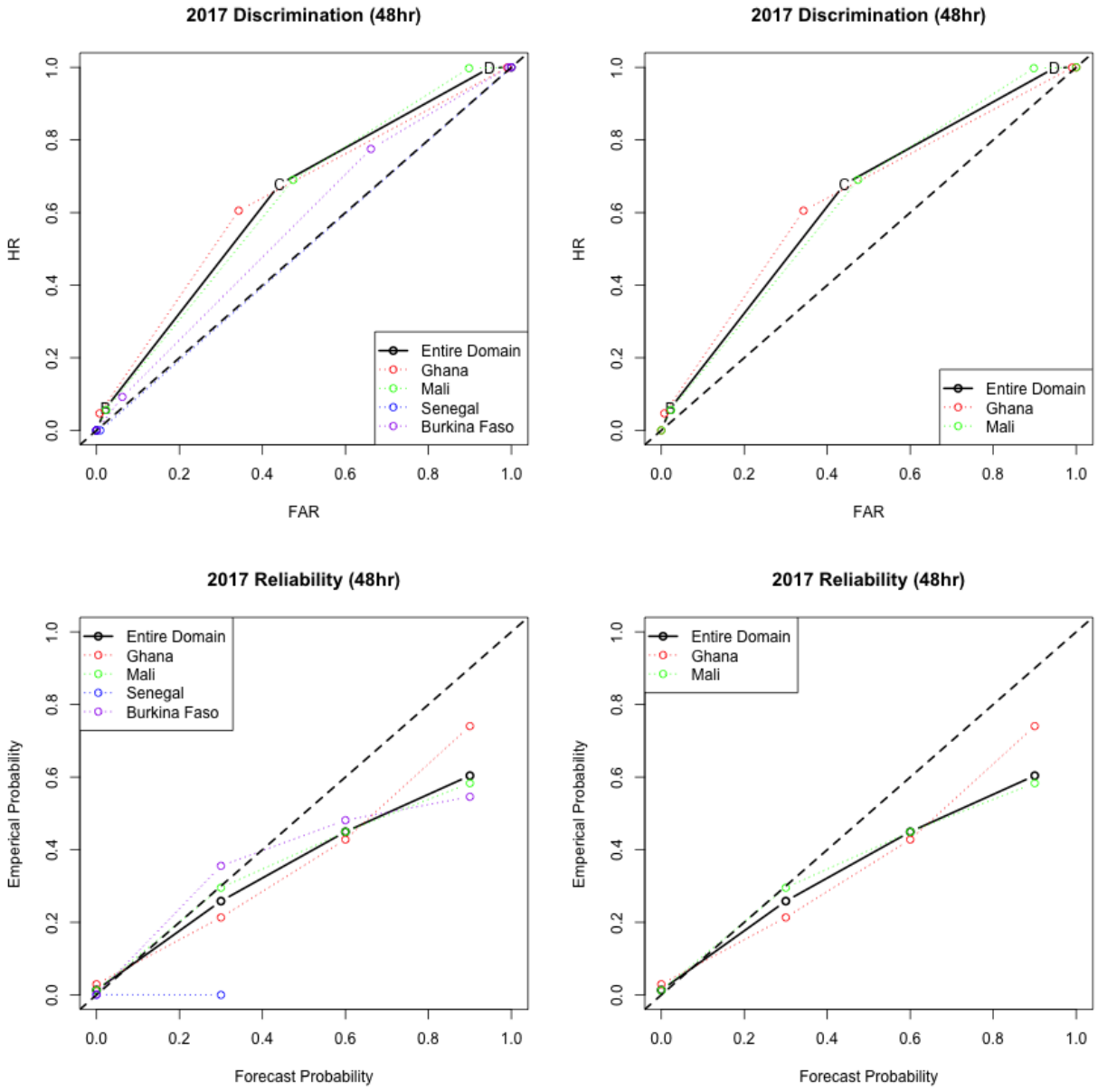


Plate A5: Discrimination/Reliability by country for 2017/48hr:



References

Ebert, E. E., Damrath, U., Wergen, W., & Baldwin, M. E. (2003). The WGNE assessment of short-term quantitative precipitation forecasts. Bulletin of the American Meteorological Society, 84(4), 481-492.

Hamill, T. M., G. T. Bates, J. S. Whitaker, D. R. Murray, M. Fiorino, T. J. Galarneau, Jr., Y. Zhu, and W. Lapenta, 2013: NOAA's second-generation global medium-range ensemble reforecast data set. Bull Amer. Meteor. Soc., 94, 1553-1565, http://dx.doi.org/10.1175/BAMS-D-12-00014.1

Hogan, R. J., and I. B. Mason, 2012: Deterministic forecasts of binary events. In Jolliffe, I. T., and D. B. Stephenson (Eds), Forecast Verification: A Practitioner's Guide in Atmospheric Science, Wiley, Chichester, 31–59.

Le Blancq, F., & Johnson, P. (2000). Introducing and verifying rainfall probability forecasts in a small meteorological office. Meteorological Applications, 7(4), 361-367.

Mason, S.J. and N.E. Graham, 1999: Conditional Probabilities, Relative Operating Characteristics, and Relative Operating Levels. Wea. Forecasting, 14, 713–725, https://doi.org/10.1175/1520-0434(1999)014<0713:CPROCA>2.0.CO;2

McBride, J. L., & Ebert, E. E. (2000). Verification of quantitative precipitation forecasts from operational numerical weather prediction models over Australia. Weather and Forecasting, 15(1), 103-121.

Mittermaier, M. P. (2008). The potential impact of using persistence as a reference forecast on perceived forecast skill. Weather and Forecasting, 23(5), 1022-1031.

Murphy, A. H., 1973: A new vector partition of the probability score. J. Appl. Meteor., 12, 595–600.

Murphy, A. H., 1993: What is a good forecast? An essay on the nature of goodness in weather forecasting. Wea. Forecasting, 8, 281–293.

Robertson, A. W., Kirshner, S., & Smyth, P. (2004). Downscaling of daily rainfall occurrence over northeast Brazil using a hidden Markov model. Journal of climate, 17(22), 4407-4424.

ANNEXES

i. The Evaluation SOW;
Attached.

ii. Any statements of difference regarding significant unresolved differences of opinion by funders, implementers, and/or members of the evaluation team;
N/A

iii. All data collection and analysis tools used in conducting the evaluation, such as questionnaires, checklists, and discussion guides;
All data collection and analysis tools are listed in the main report. Code is available upon request.

iv. Signed disclosure of conflict of interest forms for all evaluation team members, either attesting to a lack of conflicts of interest or describing existing conflicts of.

v. Summary information about evaluation team members, including qualifications, experience, and role on the team. (CVs available upon request)

- **Dr. Lisa Goddard** (IRI Senior Research Scientist in climate) is responsible for the oversight on the project. This includes report writing and presentation of results.
- **Dr. Simon Mason** (IRI Senior Research Scientist in climate) is responsible for the oversight on verification methodology selection and interpretation.
- **Nathan Lenssen** (MA in statistics and PhD student in Department of Earth and Environmental Sciences at Columbia) adapted the code and performed the calculations of the forecast evaluation.
- **Dr. Tufa Dinku** (IRI Research Scientist in environmental monitoring) is responsible for assessing the observational data, separate from the satellite data provided by Ignitia, and assuring quality.
- **Andrew Kruczkiewicz** (IRI Staff Associate in environmental monitoring) is the IRI point of contact. Andrew also explored the potential availability of alternate forecast data available for the period of interest.

Comments on IRI final report on iska forecast accuracy

The study was intended to answer three questions:
Q1a. Reliability: Does rainfall occur more frequently when confidence is high that it will rain?
Q1b. Discrimination: When it rains, do the forecasts indicate higher confidence in rainfall occurring compared to when it is dry?
Q1c. How skillful are Ignitia's forecasts compared to other available forecast model(s), based on actual in situ observation data?
An answer (to the degree it is possible) should be answered in the abstract, as many non-scientific readers only read the abstract and not the rest of the study. Following the same logic, the following section should be removed from the abstract, as it was not Ignitia's external communications that was under scrutiny, "*The study did find that Ignitia's claim of 84% accuracy was based on ….. their claims are not substantiated."* . It can be part of the analysis, but it shouldn't be a main result listed in abstract as it was not defined as a task for this study.

- While we believe that the ROC score is a relevant and useful indicator on certain aspects of forecast performance and complies with industry standards, we are not convinced that it can be used to draw any meaningful conclusions *when comparing* iska with the GFS benchmark dataset. The reason is that an apple-to-apple comparison cannot be made as the benchmark dataset has a strong bias (overpredicts rainfall). Given the concept of discrimination, i.e., that it looks at what the observation is (e.g., "did it rain?") and then checks what the forecast was saying, it is very likely that a rainfall observation will be matched by a forecast saying rain, contributing positively to the discrimination score. It is important to note that we are not questioning using the ROC as such, in fact, we believe it is a relevant metric that we will also use moving forward. The comment only concerns to which degree it is meaningful if a biased model is compared to a much less biased model, and what conclusions can be drawn and how the results should be interpreted under such circumstances.

- While accuracy as such is determined from the binary rain/no-rain (why only "dry" and "high chance" was used to determine it), we are happy to see that the numbers are confirmed by IRI. However, we do agree that it would be more meaningful to use other metrics in order to quantify our performance to represent the full range of forecasts, moving forward. We will look into whether it would be possible to use a combination of reliability and discrimination, such as the Brier Skill Score or the CRPSS, to produce a relevant and meaningful single score that can be used for apple-to-apple comparisons. We acknowledge the fact that there is no single metric that captures all aspects of forecast performance, which has added to the confusion regarding the use and meaning of "accuracy".

- We see that our ensemble-based algorithm can be improved by generating more of the extreme categories, as the frequency of occurrence of these are much lower than they need to be, while being well calibrated (reliable). This is one of the conclusions from IRI that was insightful and we will take action on moving into 2018.

- We also note that our 24h forecasts have improved from 2016 to 2017 relative to the GFS benchmark data. This tells us that our improved model system for initialization and the nowcasting components have added relatively more value this year, while we continue working on extending the lead time of this advantage to also be reflected in the 48h forecasts. Experiments addressing this are currently being performed and preliminary results hold promise for the next rainy season. At the same time, there is a physics barrier for the predictability of convective rainfall structures when using classic deterministic models. To improve the discrimination, we are therefore developing machine learning processes alongside the R&D we perform on convective origination.

- When comparing against ENACTS, we are converging our result to the "perfect reliability" curve for day 1. This is very promising, as the ENACTS dataset comprises many more ground observations, meaning that this gridded dataset is likely to be more accurate and representative of reality. This suggests that our model is more well-calibrated than previously thought when comparing against NOAA RFE 2.0. We believe that it could be very useful to use this dataset for future comparisons.

Minor corrections

- Page 3, last paragraph, and page 4, top paragraph: the resolution should read 10 km, not 100 km.
- Page 9, last paragraph: IKSA should be replaced with ISKA (two instances).

Finally, we would like to extend out thanks to IRI for the clear analysis. The report provided by IRI regarding iska forecast accuracy presents a clear analysis of iska performance across West Africa. The report presents us with a generally positive evaluation of our forecast performance, while also identifying some current weaknesses where there is potential for improvements. The methodology appears relevant for Ignitia moving forward in its tracking of forecast performance, including a range of more representative indicators than had been previously used.

Ignitia AB	HQ	Ignitia Ghana Ltd	Ignitia Weather Services Ltd
Organisation No Sweden: SE 556-802 07 46 info@ignitia.se	Bergsundsgatan 25 SE 117 37 Stockholm, Sweden	Kotoka International Airport PO Box KA 18183 Accra, Ghana +233 (0) 302 542 116	3rd Floor, Africa Re Building Plot 1679, Karimu Kotun Street Victoria Island, Lagos, Nigeria

Ignitia AB

Organisation No Sweden:
SE 556-802 07 46
info@ignitia.se

HQ

Bergsundsgatan 25
SE 117 37
Stockholm, Sweden

Ignitia Ghana Ltd

Kotoka International Airport
PO Box KA 18183
Accra, Ghana
+233 (0) 302 542 116

Ignitia Weather Services Ltd

3rd Floor,
Africa Re Building
Plot 1679, Karimu Kotun Street
Victoria Island, Lagos, Nigeria