Weather and Climate Extremes 33 (2021) 100325

Contents lists available at ScienceDirect

Weather and Climate Extremes

journal homepage: www.elsevier.com/locate/wace

Verification of forecasts for extreme rainfall, tropical cyclones, flood and storm surge over Myanmar and the Philippines

David MacLeod [a], Evan Easton-Calabria [b,*], Erin Coughlan de Perez [c], Catalina Jaime [c]

[a] Atmospheric Oceanic and Physics Department, Clarendon Laboratory, University of Oxford, Sherrington Road, Oxford, OX1 3PU, UK
[b] Department of International Development, University of Oxford, 3 Mansfield Road, Oxford, OX1 3TB, UK
[c] Climate Science Team, Red Cross Red Crescent Climate Centre, Anna van Saksenlaan 50, 2593 HT, Den Haag, Netherlands

ARTICLE INFO

Keywords:
Forecast verification
Forecast-based financing
Anticipatory action
Extreme rainfall
Tropical cyclones
Floods
Storm surge
Myanmar
Philippines
Southeast asia

ABSTRACT

Within the humanitarian sector, there is a pressing need to scale up anticipatory action as climate change-related disasters increase. This article evaluates forecasts relating to extreme weather events – extreme rainfall, tropical cyclones, river flooding and storm surge – in Myanmar and the Philippines to assess the feasibility of using such forecasts to develop early warning systems and responses. To make best use of limited extant data, a variety of methods (reliability diagrams, hit rates, false alarm ratios, correlations) are employed. We review the skill of the European Centre for Medium-Range Weather Forecasts (ECMWF) tropical cyclone forecasts and find that whilst errors in cyclone position are relatively small, forecasting intensity is more difficult. When a tropical cyclone has formed, the probabilities provided in the ECMWF track forecast are highly reliable and only slightly overconfident. A tropical cyclone activity product is relatively reliable for forecasts less than a week ahead for North Indian Ocean cyclones affecting Myanmar, but becomes very overconfident beyond this. Looking at flood forecasting models, a comparison of the Global Flood Awareness System (GloFAS, produced by the ECMWF and the European Commission as part of the Copernicus Emergency Management Services) with the Global Flood Forecasting Information System (GLOFFIS, produced by Deltares) demonstrates that both GloFAS and GLOFFIS have difficulty simulating 1 in 2 year return period flows or higher, although GloFAS performance is better than GLOFFIS. GloFAS reforecasts show significantly overconfident probabilities over Myanmar where discharge observations are available, whilst the lack of a GLOFFIS reforecast prevents evaluation of this forecast system directly. Evaluation of the ten-day operational storm surge forecast (the Global Storm Surge Information System, GLOSSIS) produced by Deltares was attempted but lack of any data prevented assessment. These findings present valuable insights into how well forecasts perform, which is crucial information for establishing effective humanitarian action mechanisms.

1. Introduction

In recent years countries in Southeast Asia have experienced more extreme weather events that in cases have turned into disasters (IPCC 2019). This has led humanitarians and others involved in preventive and disaster risk reduction responses to seek to understand how well extreme weather events in the region can be forecasted. Regular hazards include extreme rainfall, cyclones, river floods, and storm surge. Despite their significance for human welfare and economic wellbeing, little work has hitherto analysed the effectiveness of forecasting hazards in these areas.

Two countries in the region particularly at risk of these hazards are Myanmar and the Philippines. Extreme rainfall events can lead to significant damages and socio-economic impacts, from both fluvial and pluvial flooding. Flooding impacts are particularly large in tropical and monsoonal regions such as in these countries, where sustained rainfall in a catchment can lead to significant flow anomalies downstream. The Philippines is also one of the most at-risk countries from tropical cyclones (TC) (NOAA 2010), whilst the few TCs that form per year in the North Indian Ocean can be devastating to Myanmar. Both countries are at risk from storm surge.

This article evaluates forecasts of these events over Myanmar and the Philippines with the aim of improving understanding of how

* Corresponding author.
E-mail addresses: david.macleod@bristol.ac.uk (D. MacLeod), Evan.easton-calabria@qeh.ox.ac.uk (E. Easton-Calabria), Coughlan@climatecentre.org (E.C. de Perez), jaime@climatecentre.org (C. Jaime).

https://doi.org/10.1016/j.wace.2021.100325
Received 20 April 2020; Received in revised form 25 March 2021; Accepted 7 April 2021
Available online 10 May 2021

anticipatory action can be undertaken. The Red Cross Red Crescent Climate Centre conducted this study to support next steps in conducting forecast-based financing (FbF). There has been growing interest in adding a forecast-based financing component in South Asian countries to the Southeast Asia Disaster Risk Insurance Facility (SEADRIF), a new insurance mechanism meant to strengthen countries' financial resilience in the face of climate and disaster shocks (SEADRIF 2019).[1] Practitioners need to understand how well forecasts perform in order to set up effective action mechanisms. However, in countries such as these, very few observations exist which can be compared to historical forecasts in order to assess skill, meaning that opportunities for anticipatory action such as FbF are unavailable (Coughlan de Perez et al., 2015).

This article seeks to assess the feasibility of using forecasts for extreme rainfall, cyclones, river floods and storm surge in both countries as input to an anticipatory financing mechanism. In the face of very little information, a variety of methods (reliability diagrams, hit rates, false alarm ratios, correlations) are employed to make best use of existing data. Several global forecast products have been suggested for use in Southeast Asia, including rainfall forecasts for flash floods as well as river flood forecasts and storm surge forecasts. While FbF systems seek to use national forecasts,[2] two aspects of global forecast products make them particularly suitable to be also evaluated for FbF: they are often freely available and they are not subject to subjective modification. While many other forecast evaluations have been carried out, they do not generally focus on extreme events, and are less available in developing countries. Here we evaluate the European Centre for Medium-Range Weather Forecasts (ECMWF) for extreme rainfall and tropical cyclone forecasts. River flooding forecasts from the Global Flood Awareness System (GloFAS) produced by the ECMWF and the European Commission as part of the Copernicus Emergency Management Services are compared with the Global Flood Forecasting Information System (GLOFFIS) produced by Deltares. However, as no forecasts are available for GLOFFIS, only the historical simulations are compared. The ten-day operational storm surge forecast produced by Deltares as the Global Storm Surge Information System (GLOSSIS) is also reviewed.

In the following section we discuss our methods for evaluating forecasts relating to each hazard. In the results section we then present findings for each hazard. Extreme rainfall forecasts; the skill of tropical cyclone forecasts; and forecasts of river discharge of the GloFAS are evaluated and compared with the GLOFFIS. The discussion analyses these findings and presents recommendations for both operational use of forecasts as well as future analysis. Section five concludes.

2. Material and methods

Here we outline the forecast and observational data used in the study, for extreme rainfall (section 2.1) tropical cyclones (section (2.2) and for flooding (section 2.3).

2.1. Verification of extreme rainfall events

2.1.1. Reforecast data and verification observations

Twice per day, ECMWF produces a 15-day ensemble forecast of 51 members (hereafter referred to as the ENS; see footnote for full description of the ENS).[3] To evaluate the skill of the ENS, we used a version of the hindcast produced for all ENS-extended start dates during 2018 (initializations for all Monday and Thursdays for the period 1998–2017), which provides 105 start dates over 20 years, for a total of 2 100 separate forecasts (see supplementary information for more information on the ECMWF forecasting system used). While the ENS-extended hindcast is available for lead times up to 46 days, only the data up to 15 days is considered here.

The Climate Hazards Infra-Red With Stations (CHIRPS, Funk et al., 2015) dataset is used to evaluate the skill of the ENS hindcast. CHIRPS is a gridded dataset of daily precipitation, which merges precipitation estimates from gauge-based ground observations with infra-red satellite retrievals. CHIRPS data is available at a 5 km resolution, which was downloaded before being interpolated to the ENS 18 km grid for comparison with the hindcast.

2.1.2. Verification

Here extreme rainfall is defined as a period which falls above the 99th percentile of the historical distribution. Percentile thresholds are calculated across the entire dataset separately for each gridpoint, for the observations and the reforecast separately. This provides an implicit correction of any model bias, as threshold exceedance probabilities are calculated relative to the model's climatological distribution.

Results are presented for multi-day periods of rainfall accumulation (one, three, five and seven day). Analysis of multi-day rainfall ENS forecasts is assessed at a range of lead times, from zero lead up to 15 days ahead. The lead time labelling corresponds to the start of the period of the forecast target.

We use reliability diagrams to evaluate the reliability of probabilities (Joliffe and Stephenson, 2012). The reliability diagram indicates the bias in model probabilities; the degree to which they are over or under confident. Hit rates – the fraction of events which do result in an event and for which action would be triggered – and false alarm ratios – the fraction of action triggers which do not result in an event (i.e. an action in vain) – are also assessed.

Evaluation of extreme events needs a large sample of forecast-observation pairs. Approximately seven million forecast-observation pairs for each region are obtained and the reliability diagrams, hit rates and false alarm ratios are calculated across all spatial points for each country. In order to assess potential spatial heterogeneity in skill, the Pearson's product-moment correlation coefficient of ensemble mean with observed anomalies, for each gridpoint, across all 105 start dates and 20 years. Anomalies are calculated for each start date by removing a 20-year mean climatology (smoothed over five start dates centered on the date). Statistical significance at the 99% level is calculated using a *t*-test (although with a sample size of 2100, any correlation marginally above zero passes this test, so statistical significance should not be taken as an indication of useful forecasts).

2.2. Tropical cyclones methods and materials

There is a large amount of existing verification of TCs. Here we collate results across various studies to synthesize an assessment on the skill of TC forecasts relevant for the two countries. Two ECMWF TC products are reviewed. While other TC forecasts are available (e.g. Yamaguchi et al., 2019), ECMWF forecast represents the state-of-the-art of TC forecasting and is consistently the best-performing global forecast model (e.g. Yamaguchi et al., 2012; Yamaguchi et al., 2015; Lee et al., 2018; Titley et al., 2020). At ECMWF a forecast of TC track and intensity is produced automatically once a TC has formed, from the next available high resolution (HRES) and ENS forecasts, initialized twice-daily. Hereafter this product is referred to as the track.

[1] Countries involved are Cambodia, Indonesia, Lao PDR, Myanmar, and the Philippines with the support of Japan and Singapore (SEADRIF 2019: 2).

[2] Although FbF would ideally use national forecasts we have not carried out evaluation of relevant national forecasts, and are not aware of others who have done the same. This is primarily due to two main barriers: Firstly, target countries do not produce the forecasts. Secondly, when they do, they are often produced without an accompanying reforecast dataset with which to estimate skill.

[3] https://www.ecmwf.int/en/forecasts/documentation-and-support/extended-range-forecasts/monthly-forecasting-ensemble-generation.

In addition to this irregular cyclone track forecast, a forecast of tropical cyclone activity is provided from the ENS forecast. For this product, strike probability is defined as the probability that a storm will pass within 300 km of a particular location within a two day window, and the product is provided for two-day windows up to ten days ahead. Hereafter this forecast is referred to as the cyclone activity. This is notably different to the track probability as it also attempts to anticipate potential formation (genesis) of TCs, whilst the track probability forecast is only produced after a TC has already formed. Activity forecasts are generated for three tropical storm categories (cyclones >8 m/s, storms >17 m/s and hurricanes/typhoons >32 m/s).

2.3. *Flooding methods and materials*

To evaluate the potential for early action in advance of high impact flooding and storm surge events, three forecasting systems are evaluated. Two systems provide global river discharge forecasts: GloFAS and GLOFFIS. GloFAS and GLOFFIS are ensemble forecasting systems using meteorological forecasts to drive hydrological models and produce probabilistic forecasts of river flow out to 10 (GLOFFIS) and 30 (GloFAS) days ahead. GLOSSIS provides a deterministic storm surge forecast using meteorological forecasts to drive a tide and surge model. The third system is GLOSSIS, a storm surge forecast. The supplementary material provides an overview of each.

2.3.1. *Evaluating GloFAS reforecasts*

Verification of GloFAS forecast skill is made possible by the existence of a reforecast dataset generated from GloFAS simulations driven by historical forecasts from ENS. Here we use forecasts for the period 1980–2018 corresponding to version 2.0 of the system. Daily discharge data was available for several stations covering the GloFAS reforecast period by the Global Runoff Data Centre (GRDC) (see map below) and the GloFAS reforecasts are verified directly against the observations; evaluating reliability of probabilities, hit rates and false alarm ratios.

No relevant discharge observations are available at GRDC for the Philippines, so it is not possible to evaluate the GloFAS forecast directly. Instead an indirect approach is taken, using the GloFAS discharge ‘reanalysis’, created by driving GloFAS with the ERA5 reanalysis. Discharge reforecasts and reanalysis were extracted for all GloFAS reporting points in the Philippines and Myanmar. Comparison of reforecasts with reanalysis at each station suggests relative performance between stations as well as an upper limit on forecast skill; the skill which would be obtained if the reanalysis perfectly reproduced discharge observations. A comparison of results for stations across Myanmar and the Philippines then provides a general picture of potential GloFAS skill in the two countries.

2.3.2. *Comparing GLOFFIS and GloFAS using discharge reanalysis*

Unfortunately reforecast dataset exists for the GLOFFIS system, precluding a direct evaluation of the forecast skill. However a GLOFFIS “reanalysis” of historical discharge has been produced for several stations in Myanmar and this was made available by Deltares. This reanalysis is produced by running the discharge model with meteorological forcing to produce a best-guess estimate of historical discharge. We then compare the GLOFFIS reanalysis with the GloFAS reanalysis[4] to demonstrate which system is better at transforming meteorological forcing into discharge.

2.3.3. *GLOSSIS review*

For GLOSSIS no reforecast or reanalysis is available, and no data was made available for evaluation in this report. Therefore a review of the evaluation of GLOSSIS historical simulations is presented.

3. Results

The following sections present the evaluation results for the forecast skill in Myanmar and the Philippines by hazard: extreme rainfall (3.1.), tropical cyclones (3.2.), and flooding (3.3.).

3.1. *Extreme rainfall*

ECMWF forecasts over both Myanmar and the Philippines show some skill. Results indicate that forecasts of extreme one and three day rainfall perform better than longer period targets. Increased probabilities of extreme rainfall indicate enhanced risk of extreme events; this holds for forecasts up to at least one week ahead. However forecasts are significantly overconfident and raw probabilities from the forecast ensemble should not be relied upon directly.[5] For Myanmar probabilities are around five times larger than the eventual probability of an extreme outcome, whilst for the Philippines probabilities are at least three times as large.

3.1.1. *Myanmar*

The reliability diagram for 99th percentile three-day rainfall over Myanmar gridpoints starting two days ahead is shown in Fig. 2a (see supplementary material for information on ensemble mean correlation for three-day average rainfall) (Fig. 1). Three-day rainfall two days ahead is found to show the highest level of reliability; other targets and lead times are shown in supplementary Fig. A.3. The highest correlations (around 0.3–0.4) are found for one and three day averages, and correlation drops with lead time; for six days ahead and longer some parts of the country have correlations below statistical significance. Forecasts for seven-day averages at longest lead are severely overconfident, with an almost-flat reliability diagram.

In Fig. 2, as well, probabilities are heavily overconfident at all lead

[4] Note that different meteorological forcing datasets are used to produce the GLOFFIS and GloFAS discharge reanalysis. The GLOFFIS discharge reanalysis uses meteorological forcing from the Multi-Source Weighted-Ensemble Precipitation product (MSWEP, Beck, 2017), whilst the GloFAS discharge reanalysis uses the new ERA-5 reanalysis. No consistent discharge reanalysis data is available for a clear comparison between the two systems. However a global comparison of precipitation datasets over the United States showed that both MSWEP and ERA-5 were both ranked highly, with the MSWEP performing best overall (Beck, 2019). Although this ranking should not be extrapolated directly to other locations, particularly when observational station density may be vastly different, it does at least indicate that both discharge simulations take advantage of the best possible rainfall input data available. It is unlikely then that poorer performance in simulating discharge in one model compared to the other is due to using a significantly worse meteorological forcing. However without additional simulations it is impossible to quantify the contribution of meteorological forcing input to any error in simulated discharge.

[5] Other options include post-process forecasts to calibrate probabilities or to treat all probabilities greater than the climatological frequency as a forecast of “increased risk” without a quantitative value. This however limits use for FbF.

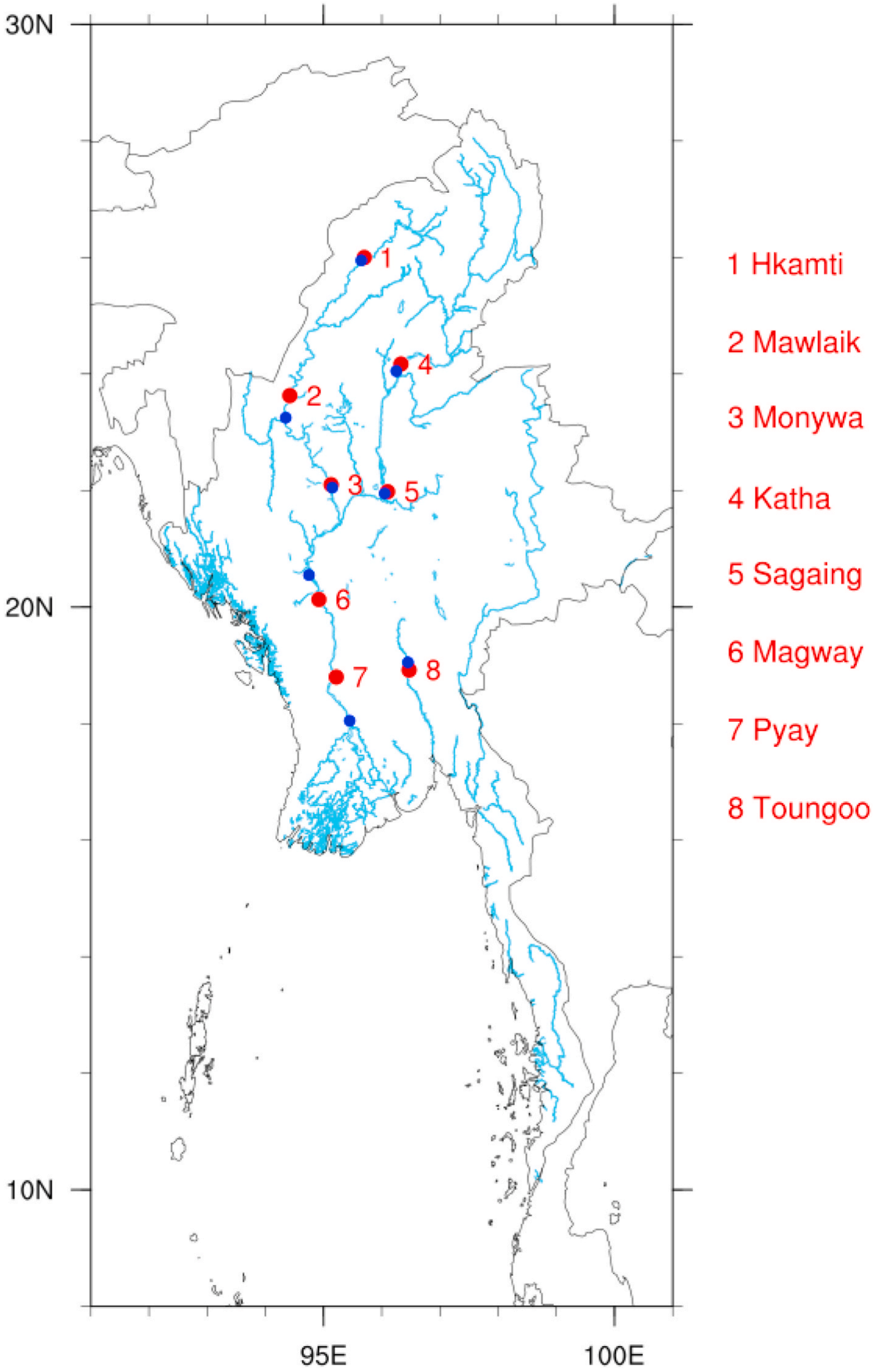

Fig. 1. Stations providing discharge observation records to GRDC (red). Blue dots indicate the nearest GloFAS reporting point to each station and light blue shows the major river network of Myanmar. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

times and for all targets. The slope of the reliability diagram is at best 1/5th of the expected slope of a perfect forecast (e.g. three-day average at zero-day lead). That is, the issued forecasts are at best five times too confident; when the issued forecast is 100%, the relative frequency is closer to 20%. The implication of this for forecast-based action is that if a 100% probability trigger for 99th percentile three-day average rainfall at zero-day lead is used, a false alarm is expected 80% of the time. For lower probability thresholds and forecast targets with a shallower reliability diagram, the chance of a false alarm will be higher.

False alarm ratios and hit rates for three-day rainfall starting two days ahead is shown in Fig. 2b. In all cases false alarm ratios are higher than 80%. The best achievable hit rate is around 30%, although this is achieved by acting on any non-zero probability and results in acting around 80 times over 10 years in a single location (eight times per year). Increasing the probability trigger threshold to 10% roughly halves the number of actions, whilst reducing the hit rate to around 15%.

3.1.2. *Philippines*

Reliability diagrams for Philippines grid points are shown in Fig. 2c for three day average rainfall starting two days ahead (results for other

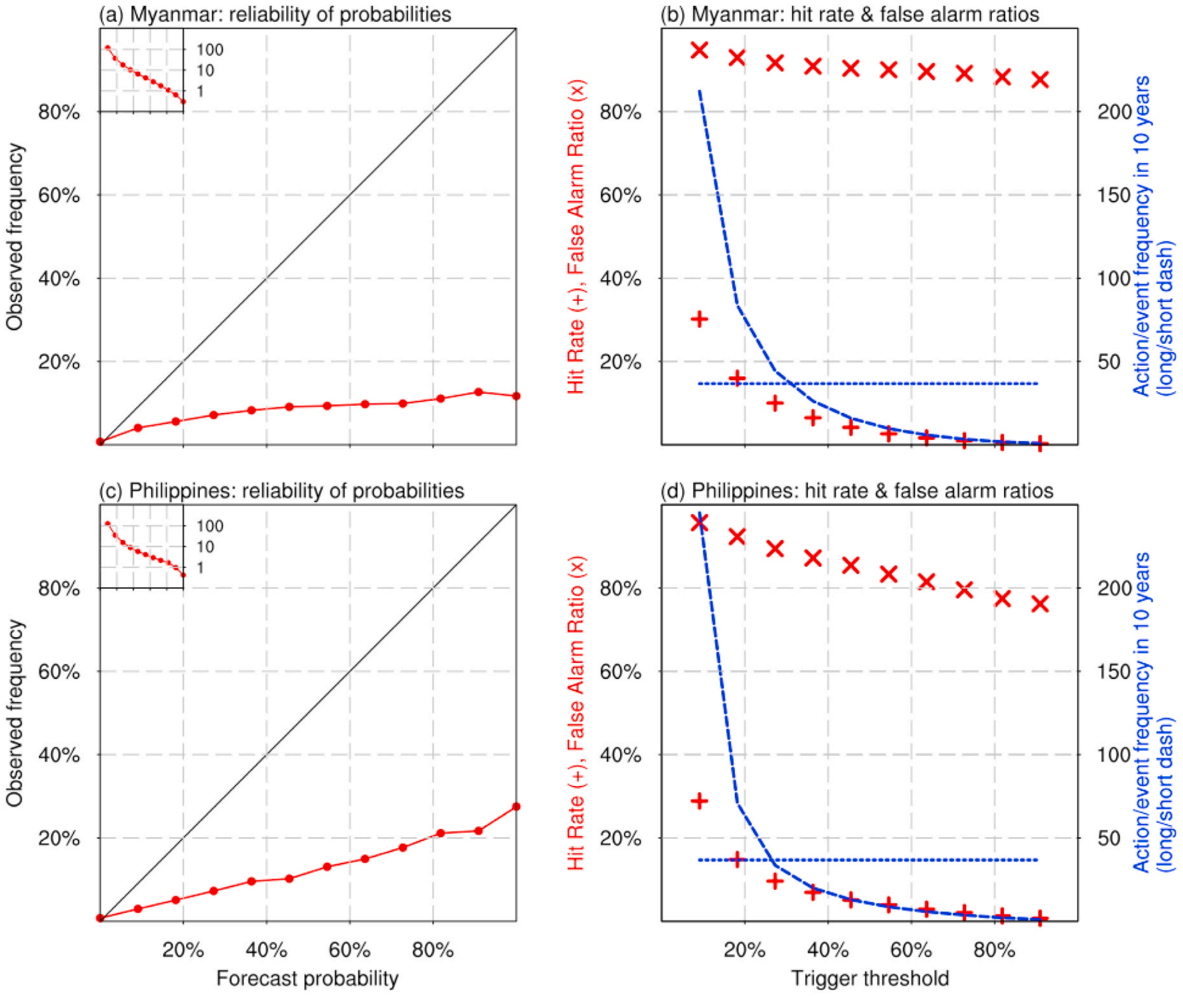

Fig. 2. Verification of a 99th percentile three-day precipitation over all Myanmar (a–b) and Philippines (c–d) gridpoints, for forecasts starting two days ahead. (a, c): reliability diagrams, indicating the relationship between forecast probabilities and outcomes. The smaller inset plot on the left plot shows the expected number of forecasts at each probability level at a point over a ten year period. Note that the X axis of the inset plot corresponds to the X axis of the larger plot and the markers are plotted at consistent probability bins (0% probability is not plotted in the smaller plot; the vast majority of forecasts show 0% probability of the event). (b, d): hit rate (red plus) and false alarm ratio (red diagonal cross), as a function of probability trigger (left axis), along with expected number of actions and events in a 20-year period associated with each probability trigger (long, short blue dash, right axis). Reliability diagrams for other periods and lead times are shown in appendix Fig. A.1, A.3 and hit rate/false alarm ratio plots are shown in appendix Fig. A.2, 4. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

periods and lead times are shown in supplementary information Fig. A.4; see supplementary material for information on ensemble mean correlation for three-day average rainfall). Like Myanmar, the probabilities are also highly overconfident, although slightly less so than Myanmar. However the reliability slope is still around 1/3rd of perfect reliability, indicating probabilities are at best around three times too confident. For instance, a trigger of a 100% probability of three-day rainfall at two-day lead would result in a worthy action around 30% of the time, with an action in vain 70% of the time.

False alarm ratios and hit rates across all probability thresholds are shown for Philippines grid points in Fig. 2d for three-day rainfall starting two days ahead (results for other periods and lead times are shown in supplementary Fig. A.5). The best achievable hit rate is around 35%, achieved by acting on any non-zero probability of one-day precipitation with a two-day lead. As discussed in Myanmar, this is associated with acting around eight times per year. Increasing the probability trigger threshold to 10% reduces the number of actions to around once per year, whilst reducing the hit rate to around 20%. The false alarm ratios associated with these thresholds is around 90% or higher.[6]

3.2. Results: skill of tropical cyclone prediction

The probabilities contained in the ECMWF tropical cyclone track forecasts are highly reliable once a cyclones has formed. Acting in

[6] For an insurance product, it is usually not cost-effective to have a payout more than once per year; most products are structured for payouts in only the most catastrophic events. The Forecast-based Action by DREF fund of the IFRC anticipates taking early action for events that happen with an approximately 5-year return period. However, more local and less costly actions and funding mechanisms might be able to pay out more frequently, based on less catastrophic events for actions of smaller scope.

regions where the probability of cyclone strike reaches 90% are associated with a probability of action in vain of 10%. A tropical cyclone activity product is also available from ECMWF, this incorporates potential cyclone activity before cyclone formation. The probabilities of this product are less reliable; for the North Indian Ocean basin forecasts become quite overconfident at lead times of one week (with forecast probabilities of 90% leading to hits 30% of the time), whilst for the Western North Pacific reliable anticipation of cyclones at longer leads than one week may be possible.

3.2.1. Track probabilities

In this section we present hit rates and false alarm ratios for action based on the high probability zone of track forecasts as calculated by Haiden et al. (2018); this evaluation is based on operational ECMWF forecasts made during the 12 month period ending May 31, 2018.

Fig. 3 shows several verification metrics and compares the HRES deterministic run with results from the ensemble ENS run (18 km resolution). The position error for three-day forecasts averages 150–200 km over the last few years, and at the five-day lead time the error is around 300–350 km (Haiden et al., 2018). Mean intensity errors for three-day ahead forecasts are also shown. These show an average positive bias in central pressure indicating that simulated TCs tend to be on average less intense than observed.

Reliability diagrams for the track probabilities are shown in Fig. 4 below.[7] The reliability diagram indicates that the issued probabilities in the track forecast have a positive bias: when the probability was between 90 and 100%, the actual chance of a strike was more like 85% (reading off the X, Y coordinates from the topmost point of the top panel of Fig. 4).

In addition the ROC curve is shown in Fig. 4 (bottom left panel). This indicates the expected hit rate as a function of false alarm rate. The false alarm rate is defined as the fraction of non-events which are false alarms. For TCs many gridpoints indicate 0% probabilities and the contribution of these points gives a very low overall false alarm rate (Haiden et al., 2018). To remove the influence of non-events, the modified ROC is plotted (Fig. 4, bottom right). This instead plots the hit rate as a function of false alarm ratio, defined as the fraction of actions which are false alarms. This modified ROC can be used to estimate potential hit rate and false alarm rate associated with triggering action on certain probability thresholds. This shows that acting based only in the region of 90% strike probability from a track forecast would result in a 20% hit rate and a 16% chance of action in vain. Extending that region to an 80% trigger would provide a 30% hit rate with a 20% chance of action in vain. It is important to note that in FbF, a choice of action should be commensurate with false alarm ratios and probability of detection results. Practitioners might choose to take a more costly action in the 90% zone (e.g. evacuate people), and only prepare supplies in the 80% zone. FbF can work with the forecast skill that is available, to select what actions are appropriate.

3.2.2. Cyclone activity forecasts (including genesis)

The verification provided here corresponds to the model version between 2010 and 2012. Fig. 5 shows an estimate of the reliability of the TC activity forecast which includes genesis (with thanks to Munehiko Yamaguchi). For the Western North Pacific probabilities are relatively reliable up to nine days, whilst for the North Indian Ocean probabilities are relatively overconfident at short leads (at 0–3 day lead when the issued probability is 95% the actual probability is closer to 70%; at 6–9 day lead an issued 95% probability is actually closer to 30%).

Verification of strike probabilities including genesis has not been carried out with the latest version of the model (Fernando Prates, ECMWF, personal communication). However the track probability forecast for the latest and the 2010–12 version of the ECMWF model can be compared (Yamaguchi et al., 2012 and Haiden 2018, Fig. 6) and these show no significant change in reliability. In the absence of verification of the latest operational model, it is reasonable to take Fig. 4 as an indication of the reliability of the current strike probability product.

3.3. Flooding results

3.3.1. Comparison of GloFAS and GLOFFIS historical simulations/reanalysis

Our evaluation found that the ability of GLOFFIS to simulate the timing of high flow events over Myanmar is inferior to GloFAS.[8] Simulated discharge is compared with observed discharge over the common period for each of the six stations (the period for each station where data is present across all three datasets). As such, this comparison refers only to historical simulation of events, and does not represent any prediction or forecasting. Table 1 provides details and shows the correlation of simulated with observed discharge for both GLOFFIS and GloFAS for each station. Results are shown for both Pearson's product-moment and Spearman's Rank correlation (hereafter Pearson's and Spearman's correlation).

For all stations and both metrics GloFAS performs better than GLOFFIS. An exception to this is Hkamti, where the Spearman's correlation is marginally higher for GLOFFIS than GloFAS. For all stations the Spearman's and Pearson's correlation values are high and relatively similar for GloFAS, which indicates that the simulated discharge has a relatively strong linear relationship with observed discharge. For GLOFFIS, the values of Pearson's correlation are significantly lower than Spearman's correlation across every station, indicating that whilst there is a strong monotonic relationship between simulated and observed discharge, the relationship is not linear. One implication of this is that any linear bias correction of GLOFFIS discharge may introduce errors, particularly for high flows and so bias correction should take into account this non-linearity.

No systematic biases was found in either model across all stations or common relative performance between the models. Both simulated distributions are relatively consistent with observations for Mawlaik and Monywa, whilst discharge is overestimated for Toungoo across all percentiles. For Hkamti the GloFAS distribution matches observations very well, whilst the GLOFFIS distribution underestimates discharge by around $^1/_3$ for high percentiles. However, GLOFFIS performs well for Katha and Pyay, whilst GloFAS overestimates discharge here. The GloFAS overestimation is particularly bad for Katha, where the 99th percentile in GloFAS is almost twice the 99th percentile in observations.

Fig. 6 evaluates the ability of the discharge reanalyses to identify high flow days. All days in the observed record above a specific threshold are identified, and the percentage of these days also indicated as above that threshold in the model reanalysis is calculated, and repeated for all percentiles 1 to 99. Both systems have difficulty in accurately simulating the highest flow events when run with meteorological reanalysis. For all thresholds of up to the 60th percentile, the results are similar across both systems and stations, with hit rates of over 80%. Above the 60th percentile thresholds the hit rate declines steadily to a minimum for the highest thresholds. For all stations the hit rate of 99th percentile events is less than 50% for both models. In some instances it is significantly lower.

The analysis also shows that the hit rate for above-threshold events decreases steadily for more extreme events. It is important to put Fig. 5 in terms of return periods, which are often used in flood forecasting. For instance, at Hkamti the 2, 5 and 20-year return periods of the GloFAS

[7] These have been evaluated for the last few years of operational ECMWF systems and are created by stratifying all gridpoints of all available forecasts (with a TC present at initialization) into forecast probability bins of width 10% starting at 0% (with forecast of exactly 0% treated as a separate bin).

[8] It should be emphasised that these results only strictly hold for regions where the comparison has been made.

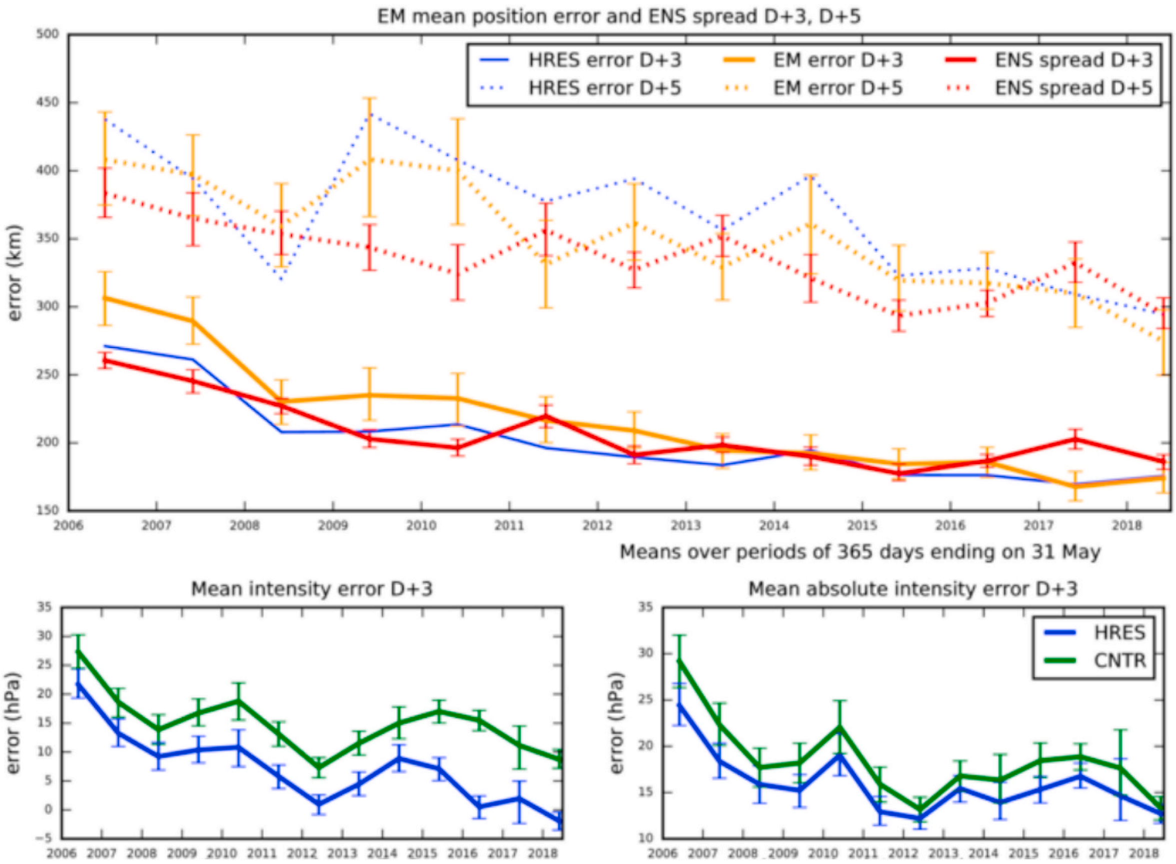

Fig. 3. Verification of ECMWF TC predictions from the operational high-resolution (HRES) and ensemble forecast (ENS). Results shown for all tropical cyclones occurring globally in 12-month periods ending on 31 May. It is also important to note that results can vary by model and basin, as discussed in Titley et al. (2020). Verification is against the observed position reported via the GTS. Top row: mean position error of ensemble mean forecast with respect to the observed cyclone (orange curve) and ensemble spread (mean of distances of ensemble cyclones from the ensemble mean; red curve); for comparison, the HRES position error (from the top panel) is plotted as well (blue curve). Bottom row: left, the mean error between forecast and reported central pressure (positive error indicates the forecast pressure is less deep than observed), right, the mean of the absolute errors of the intensity. Figure from Haiden et al., (2018). (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

reanalysis are 13,900m3s-1, 15,800m3s-1 and 19,400m3s-1. The 99th percentile, with a hit rate of 35%, corresponds to a discharge of around 12,000m3s-1. This suggests then that a hit rate around 35% and a false alarm ratio of around 65% might be expected for a 2-year return period. The decrease in hit rate with the highest flow values shown in Fig. 6 suggests results for 5- and 20-year return period events will be worse.

3.3.2. Evaluation of GloFAS reforecasts for high flow events

We next turn to evaluation of flood forecasts directly against discharge observations. Reliability diagrams for Hkamti are shown for 2, 5 and 20-year return periods in Fig. 7, with plots of possible hit rate and false alarm ratio are shown in Fig. 8.

Reliability diagrams show that forecasts are sharpest at short lead time, with all probabilities either zero or 100%, however the reliability at this short lead is not necessarily highest. For instance, when forecasts for a day-ahead forecast for a 2 year return period event indicate 100%, the event in question occurs only one in five times. This may be related to the poor performance of GloFAS at simulating the highest flows as previously shown, or may be related to biases arising from the land data assimilation with initialization (Zsoter et al. 2019, 2020). Whatever the source of the problem, the resultant operational forecasts at the shortest lead times are highly overconfident.

Whilst forecasts become more sharp at short lead times (i.e. all probabilities are either zero or 100%), the reliability at this short lead is not necessarily highest; with less than 20% of day-ahead probabilities of 100% chance of two-year return period discharge corresponding to a two-year exceedance event. This is likely related to the poor performance of GloFAS at simulating the highest flows, shown in the previous section. Coupled with underdispersion at the shortest lead times results in overconfident inaccurate forecasts.

The evaluation of hit rates and false alarm ratios found that for some stations a false alarm ratio of below 40% is potentially possible for two-year return period events with up to ten day lead time, whilst the false alarm rates associated with five-year return period events are much higher: hit rates are below 20% at all lead times and false alarm rates are above 50% for all probability triggers.

Results for Hkamti indicate the most promising performance for two-year return periods for 10-day lead forecasts. The hit rate and false alarm ratios are then shown for two-year return period forecasts at 10-day lead for all eight Myanmar stations with streamflow observations, in Fig. 9.

The skill attributes of GloFAS are location-dependent. For some stations, probability thresholds may be chosen which provide relatively low false alarm ratios: for example action on greater than 30% probability at Hkamti or Katha would result in false alarm ratio lower than 40%, although only 20% of events are successfully anticipated. Performance for Mawlaik, Pya and Toungoo is worst of all stations, with false alarm rates higher than 60% for all thresholds. For all other stations probability thresholds may be chosen which result in a false alarm ratio of 40% or lower.

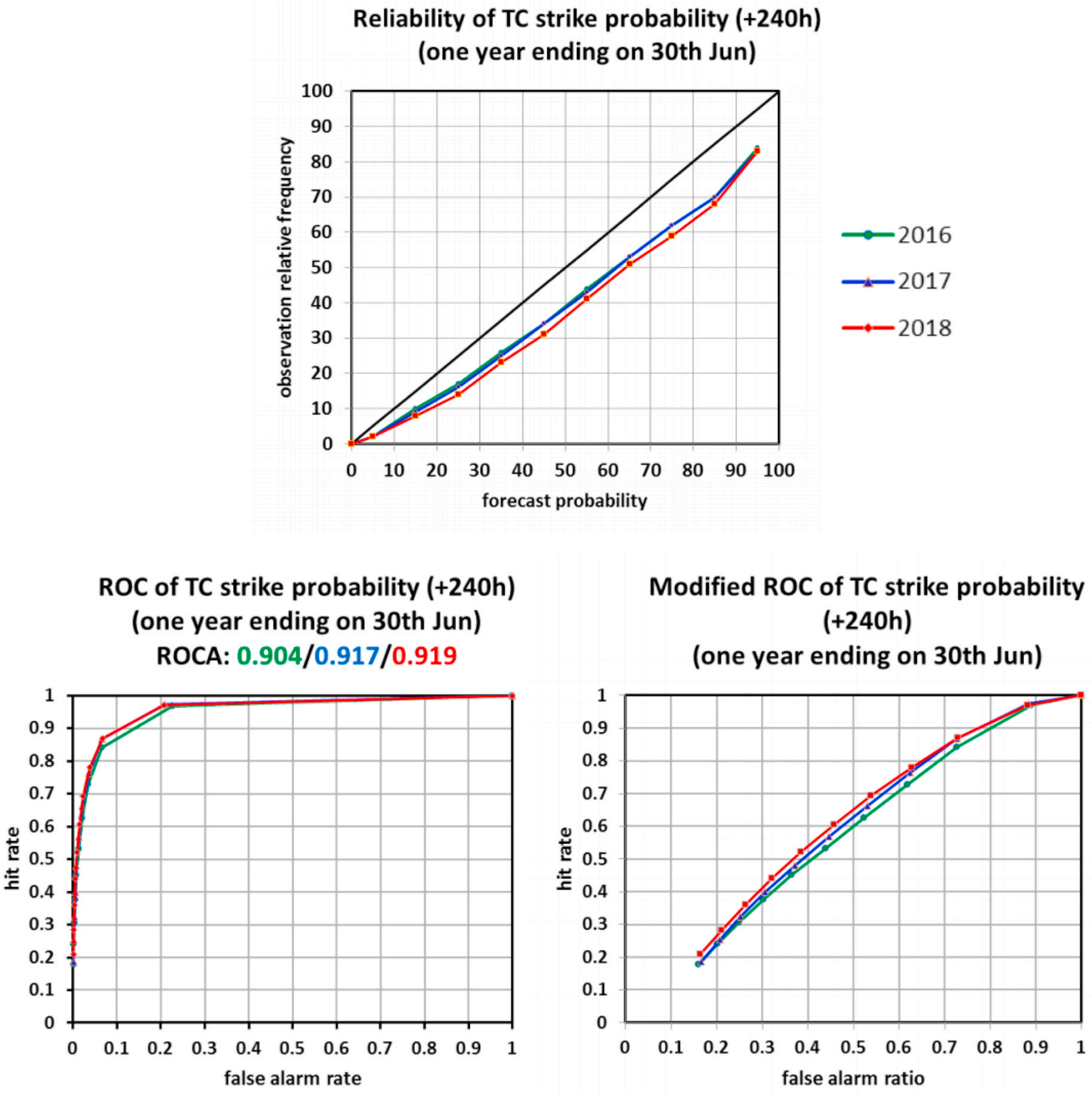

Fig. 4. Verification of ECMWF TC strike probabilities from the ensemble forecast (ENS). Strike probability is defined by the chance a TC will pass within 120 km of a point within the next 120 h. Skill for 10 day forecasts is shown for all global gridpoints but only for the subset of forecasts where a TC already exists at the initial forecast time. Results are shown for the ECMWF ENS forecast, based on all forecasts during the 12 month period ending on June 30th, 2016, 2017 and 2018 (green, blue, red). The top panel shows reliability diagrams for strike probabilities, whilst lower panels show the standard ROC diagram (left) and a modified ROC diagram, where false alarm ratio is used instead of false alarm rate. Consistent probability bins are used in all three figures. Figure from Haiden et al., (2018). (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

3.3.3. *Comparison of GloFAS performance over Myanmar with the Philippines*

Spearman's and Pearson's correlations are shown in Fig. 10 for all Philippines and Myanmar GloFAS stations. The performance of GloFAS against its own reanalysis is on average better for stations in Myanmar than the Philippines. However there is significant within-country variation in skill, and the skill of the best performing stations in both countries is comparable. This spatial variability is shown in Fig. 11, where the average Spearman's rank correlation overall lead times is plotted for each station. Overall GloFAS performs better against its own reanalysis for stations on the Irrawaddy, for stations around Mandalay, with correlations of 0.6–0.8. In addition, stations on the Salween appear to perform relatively well (correlations of 0.6–0.8), indicating potential first targets for further analysis. For the Philippines the highest correlation (0.6–0.8) is found for relatively smaller Panay river on Panay island (with a drainage basin of 2 181 square kilometers). Correlations for the larger rivers of the Philippines (the Cagayan, the Rio Grande de Mindanao and the Agusan, with drainage of 27,753, 23,169 and 11,937 km^2) correlations found are around 0.4–0.6. This analysis does not then find evidence to suggest that GloFAS will perform better over the Philippines than it does over Myanmar. The verification of GloFAS against real discharge observations presented in section 3.3.2 could therefore be considered as an upper estimate of forecast skill over the Philippines.

3.3.4. *Evaluation of GLOSSIS*

Little is known about the skill of the GLOSSIS storm surge forecast and they must be treated with caution. The GLOSSIS storm surge model is able to accurately simulate historical TC storm surges (Bloemendaal et al., 2019) and reproduce historical storm surge heights with high accuracy – however this accuracy relies on near-perfect meteorological forcing. While the analysis of Bloemendaal et al., (2019) suggests that a

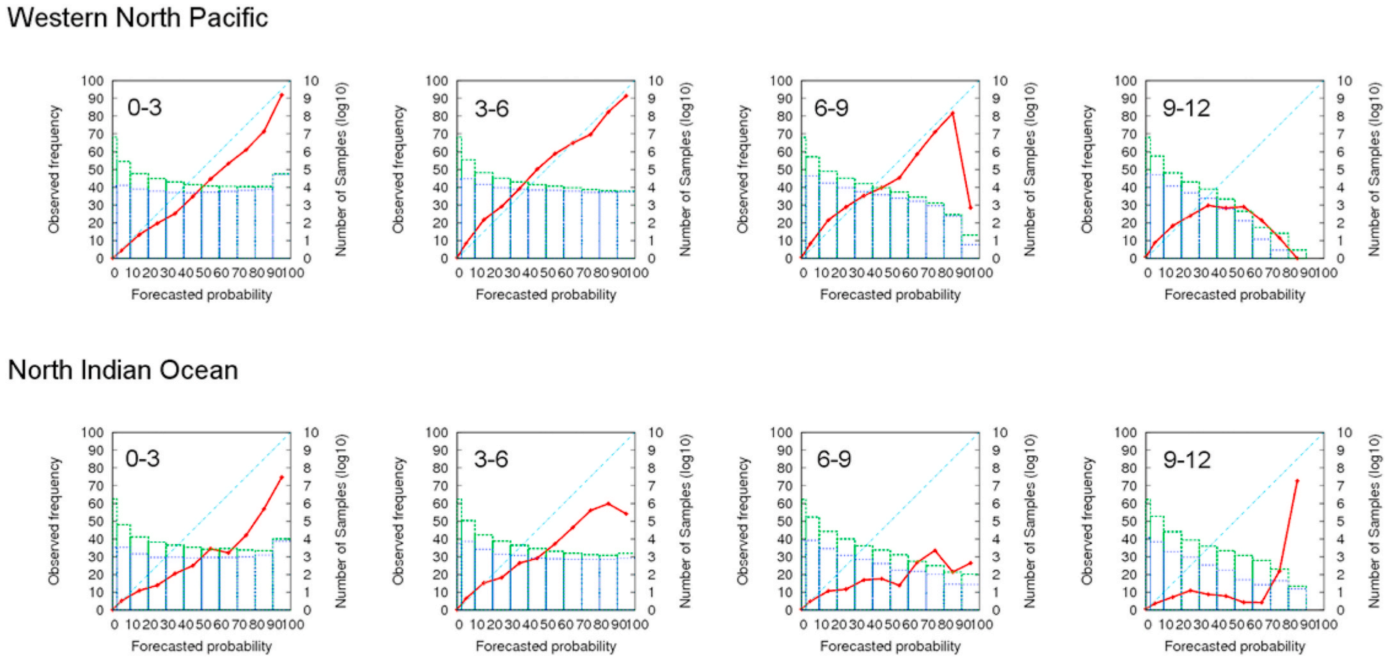

Fig. 5. Reliability of ECMWF tropical cyclone forecasts across all forecasts, including genesis; based on the operational model 2010–2012. The top row shows forecasts for grid points in the Western North Pacific basin, whilst the bottom row shows results for the North Indian Ocean basin. Forecasts for increasing lead times are shown left to right; the inset number range indicates the target period in days (i.e. 0-3 days ahead of the forecast initialization). Results provided by Munehiko Yamaguchi, following analysis in Yamaguchi et al., (2015).

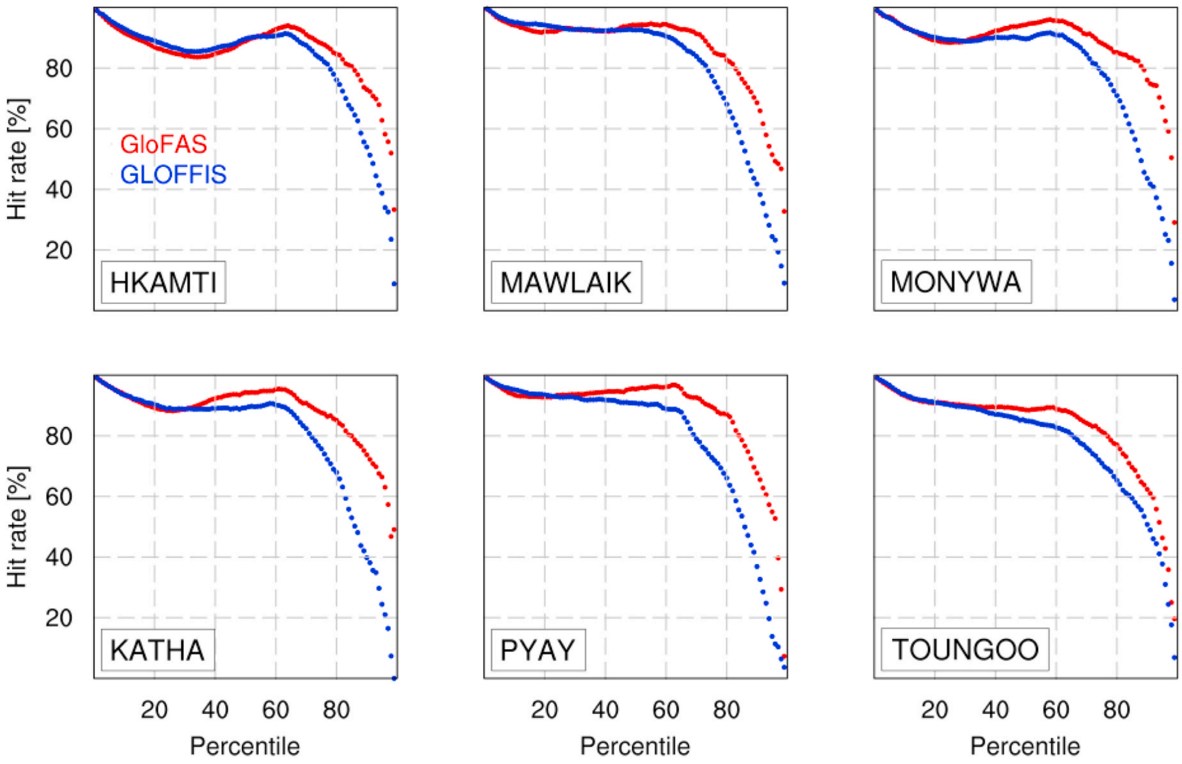

Fig. 6. Showing the hit rate of above-threshold days. That is, the percentage of days where observations show discharge above a percentile threshold which simulations also show as above threshold. Results are shown for all thresholds 1–99%, for GloFAS (red) and GLOFFIS (blue) historical simulations. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

reasonable simulation of storm surge is possible up to 9 h ahead, section 3.2 has shown that for a forecast window of up to ten days there is significant uncertainty forecasts of TC position. This in turn is a strong factor in storm surge forecast uncertainty. The large spread in TC position forecasts and the low skill in intensity forecasts suggests that the uncertainty in a storm surge forecast is potentially huge.

A probabilistic ensemble forecast would offer the chance to quantify this uncertainty in real-time, however the existing GLOSSIS system is

Table 1
Correlation of historical simulated discharge over Myanmar stations, from GLOFFIS and GloFAS, compared to observations from GRDC. Pearson's product-moment and Spearman's rank correlation are shown; for each metric and station where one model shows a statistically higher correlation than the other it is highlighted in bold.

Station	Period	Pearson's correlation		Spearman's correlation	
		GLOFFIS	GloFAS	GLOFFIS	GloFAS
Hkamti	1980–2014	0.79	**0.91**	**0.86**	0.85
Katha	1996–2010	0.82	**0.94**	0.91	**0.95**
Mawlaik	1996–2010	0.79	**0.95**	0.88	**0.93**
Monywa	1996–2010	0.79	**0.95**	0.87	**0.93**
Pyay	1996–2010	0.82	**0.96**	0.90	**0.95**
Toungoo	1980–2014	0.74	**0.86**	0.84	**0.89**

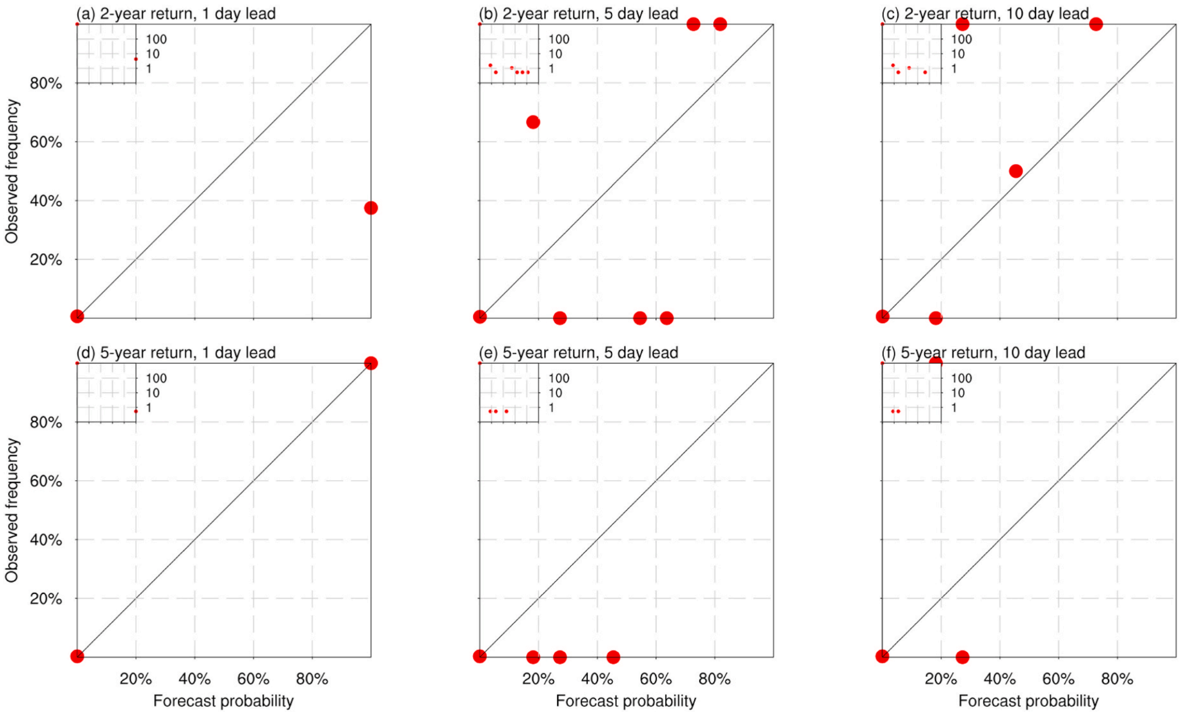

Fig. 7. Reliability diagram for GloFAS predictions of discharge at Hkamti exceeding a two-year return level (top row) and five-year return period (bottom row). Reliability of probabilities is shown for forecasts one, five and ten days ahead (left, centre, right). Forecasts are binned according to 11 equally spaced probability intervals from 0 to 100%, where no dot is plotted, no forecasts are present in the dataset within that interval. The inset figure shows the expected frequency of forecasts in each probability bin in 10 years: note the logarithmic scale.

unable to capture this uncertainty given that it is run only with a single deterministic forecast instance. The deterministic GLOSSIS forecast is therefore highly overconfident about the actual potential range of future outcomes and provides no estimate of uncertainty. The lack of any reforecast data or operational forecast data for verification means that the relationship between the ten day GLOSSIS forecasts product and actual storm surge is entirely unknown.

4. Discussion

In the following sections we discuss the findings and implications for each of the evaluations according to hazard (extreme rainfall 4.1, tropical cyclones 4.2, flooding 4.3.).

4.1. Extreme rainfall

Model probabilities for extreme rainfall are highly overconfident, as illustrated by Fig. 2 above. However, this result was for the 99th percentile forecasts over only a small location and was an assessment of model data, not an existing forecast product. The 'off-the-shelf' ECMWF products generally target much less rare events such as weekly rainfall total falling in the 'upper tercile' rainfall (i.e. above the 67th percentile) and the probabilities for these show a much smaller degree of overconfidence.[9] Indeed, analysis (not shown) found a lower degree of overconfidence for less extreme events, with higher percentile definitions of extreme events showing more overconfidence. Model-derived probabilities should not be taken at face value, and statistical calibration is necessary to generate reliable probability forecasts. It is also unlikely that the post-processed forecast probabilities will provide bold indications of extreme event probabilities. For users interested in forecasts of extreme rainfall events, using post-process forecasts to calibrate probabilities is one option, but one which may be beyond the capacity of an average user. In such cases it is recommended for users to engage with their National Met Service to provide them with reliable (calibrated) probabilities. The other option is to treat all probabilities greater than the climatological frequency as a forecast of "increased risk" without a quantitative value. This however limits use for FbF and so may only be appropriate in certain instances.

There are several possible reasons why this is an overly pessimistic view of skill, including spatial biases of individual high-impact events at the analysed scale of the data, and errors in observational data. The priority for further work should be to confirm this by verifying the reforecast against station data. However the current analysis suggests that the best possible false alarm ratio possible for a 99th percentile event is 80%, with less than 10% of events hit.

[9] See the ECMWF reliability diagram for weekly rainfall: https://www.ecmwf.int/en/forecasts/charts/catalogue/mofc_multi_verification_probability_family_reliability?facets=undefined&time=2020022000&parameter=Precipitation&week=Day%205-11&area=Tropics&threshold=in%20upper%20tercile [accessed March 20, 2020].

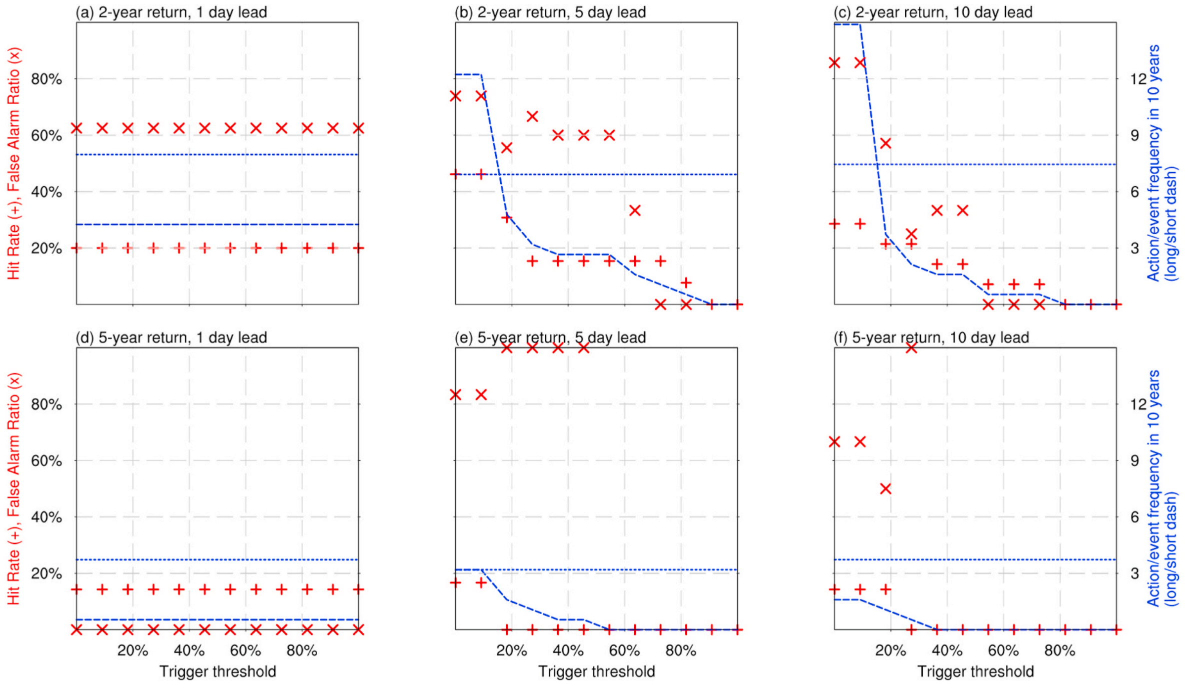

Fig. 8. Hit rate (plus) and false alarm ratio (crosses) associated with triggering on GloFAS probabilities of discharge at Hkamti exceeding a two-year return level (top row) and five-year return period (bottom row). Results are shown for forecasts one, five and ten days ahead (left, centre, right). The long dashed lines indicates the expected number of triggers in 10 years, and the short dash indicates the expected number of events over 10 years (note that this is projected based on the size and statistics of the specific verification sample for each plot and so is not necessarily consistent at different lead times).

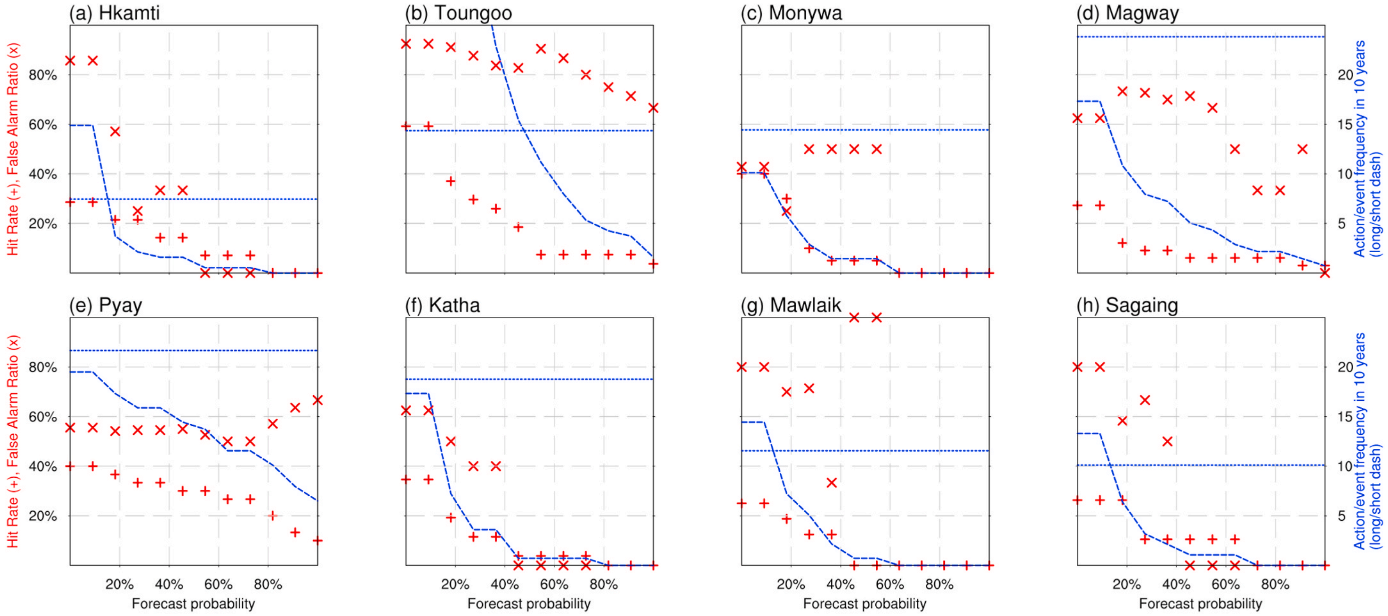

Fig. 9. Hit rate (plus) and false alarm ratio (crosses) associated with triggering on 10-day lead GloFAS probabilities of discharge exceeding the two-year return period, for all Myanmar stations with discharge observations. Top row: Hkamti, Mawlaik and Monywa. Middle: Katha, Sagaing and Magway. Bottom row: Pyay and Toungoo. The long dashed lines indicates the expected number of triggers in 10 years, and the short dash indicates the expected number of events over 10 year. Note that this is projected based on the size and statistics of the specific verification sample for each plot and so is not necessarily consistent at different lead times, in addition each day is treated as an independent event, so a sustained exceedance of a threshold is counted as multiple events in this analysis.

To form a more robust view on skill it is recommended to obtain station data from the national meteorological centres of the two countries and verify the ensemble forecast at points against this data. In addition, in order to assess the level of uncertainty in CHIRPS data over the region it would be worthwhile to make a multiproduct assessment over the region (e.g. Beck et al., 2017). Whilst the performance of CHIRPS has been shown to be superior to other products in some regions of the world (e.g. Toté et al., 2015; Dinku et al., 2018), the purpose of the dataset is originally drought monitoring and trend analysis and it should be evaluated directly for the representation of extreme rainfall events over Myanmar and the Philippines against other products. Future research should compare the representation of extreme rainfall in

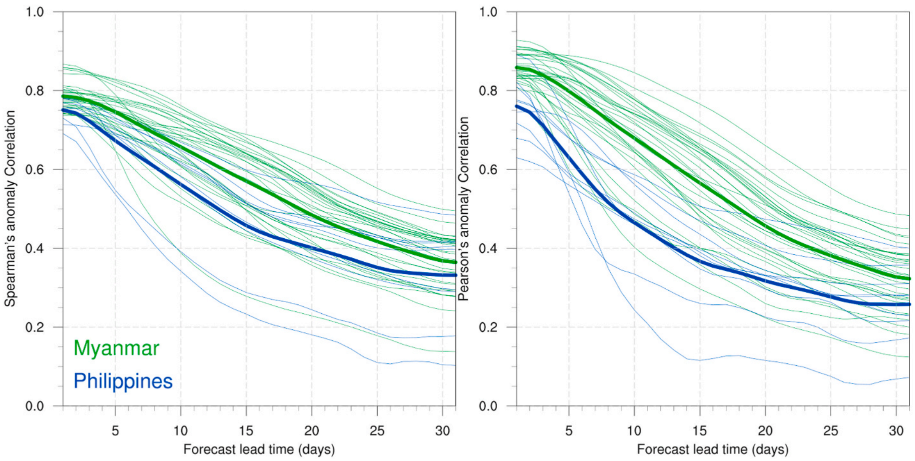

Fig. 10. Correlation of reforecast GloFAS discharge anomalies against GloFAS reanalysis, as a function of lead time. Results are shown for all GloFAS standard reporting points over Myanmar (green) and Philippines (blue). Thin lines indicate results for individual points, whilst the thick line shows the correlation averaged over all stations. Left panel and right panels show the Spearman's rank and Pearson's product-moment correlations respectively. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

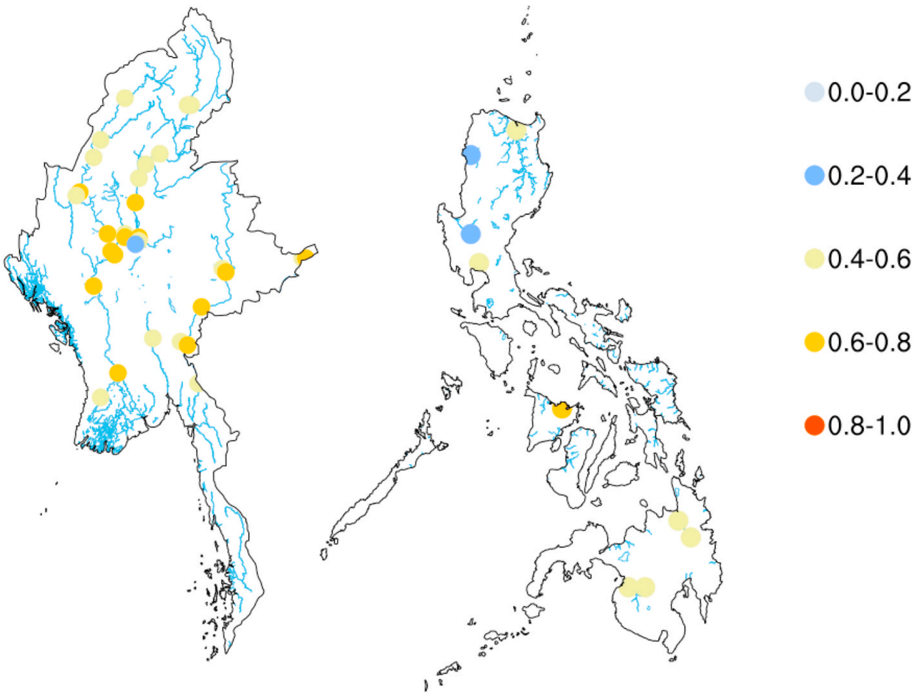

Fig. 11. Average correlation of reforecast GloFAS discharge anomalies against GloFAS reanalysis for all Myanmar and Philippines points. The values plotted are Spearman's rank correlation, which is calculated for each lead day and averaged across all, to get an indication of mean performance across the 30-day window (NB this value essentially corresponds to the average Y-axis height of each line plotted in the left panel of Fig. 10).

CHIRPS over the target countries to other datasets such as the Multi-Source Weighted-Ensemble Precipitation (MSWEP), as it is ranked highest in rainfall representation over certain regions of the globe (Beck et al., 2017).

Extant knowledge and data gaps regarding extreme rainfall could be addressed through repeating the verification of extreme rainfall forecasts in the ENS after first averaging both forecast and observations over larger regions, as averaging would reduce the impact of spatial errors in the precise location of rainfall events. This is particularly important for humanitarian preparedness actions, because as the spatial precision of forecasts drops, the utility for humanitarian actors becomes more limited.

It would also be valuable to clarify the definition of extreme rainfall in relation to flooding impact. This could occur through an impact assessment of flash flooding and links with precursor rainfall in order to define relevant rainfall thresholds. In addition in-country experts (i.e. meteorologists and hydrologists) could be consulted to document in-use definitions and evaluate the evidence that supports these thresholds. Local rainfall observations from gauge data could be used to build confidence in characterisation and verification of extreme rainfall; such local observations have been shown to be relevant for long range climate forecasting (Eakin 1999; Chisadza et al., 2015). This integration does however rely on having real-time observations available, which is extremely challenging for hydrological variables in particular.

4.2. *Tropical cyclone discussion*

The probabilities contained in the ECMWF tropical cyclone track forecasts are highly reliable once a cyclones has formed. Acting in regions where the probability of cyclone strike reaches 90% are associated with a probability of action in vain of 10%.

A tropical cyclone activity product is also available from ECMWF, this incorporates potential cyclone activity before cyclone formation. The probabilities of this product are less reliable; for the North Indian

Ocean basin forecasts become quite overconfident at lead times of one week, whilst for the Western North Pacific reliable anticipation of cyclones at longer leads than one week may be possible.

As Fig. 5 in the results section demonstrates, the cyclone activity forecast is particularly overconfident for the North Indian Ocean. While for the WNP basin the probabilities of the cyclone activity forecast are reliable up to nine days ahead, the product of North Indian ocean TC including genesis becomes unreliable at lead times greater than one week. This is shown by Yamaguchi (2012), but very recent models (Titley et al., 2020) also see a drop in skill for the North Indian ocean, and move toward overconfidence in forecasts at longer lead time (5+ days). The consistency of our findings with recent assessments of the skill of tropical cyclone track forecasts in 2017–2018 using ECMWF ENS forecasts (as here) as well as MOGREPS-G and NCEP GEFS (Titley et al., 2020) suggest that the TC forecasts for which we present evaluation (based on 2010–2012 forecasts) reasonably represent current levels of skill. Supporting our own findings, Titley et al. (2020) find that probabilities are overall reliable, showing potential utility, and tend toward overconfidence which is more pronounced at longer lead times. Skill over the Western North Pacific (North Indian Ocean) basin is generally higher (lower) than others.

Given these findings, if the cyclone activity forecast is to be used it is recommended that verification is carried out for the latest version of the model, and calibration of probabilities should be conducted. The only extant evaluation of the activity forecast is based on an older model version and so it is a priority that verification is carried out with the latest version of the model. Until that case, it is prudent to exercise caution in interpreting the cyclone activity product probabilities in the North Indian Ocean, particularly given track probability forecasts in this basin from state-of-the-art models remain relatively low-skilled compared to other regions (Titley et al., 2020). It should be noted however that other aspects of cyclone representation in forecasts have improved significantly over time, as shown in Fig. 3: the error in position and absolute intensity of tropical cyclones is around half the magnitude that it was around ten years ago. Future forecasts will no doubt improve further.

A multi-model ensemble also has the potential to provide higher skill than any single model (Yamaguichi et al., 2012), and may ultimately provide the most reliable and useful forecast, although no such operational product is currently publicly available. This is also suggested by Titley et al.'s (2020) findings, where it is demonstrated that a multi-model ensemble outperforms all individual models. However, when examining individual models, ECMWF forecasts display the best reliability skill and value compared to the other two, which demonstrates that it is a reasonable forecast to use in the absence of an operational multi-model product.

Future analysis of tropical cyclone forecasts should address the differential skill for the reliability of Western North Pacific and North Indian Ocean cyclone forecasts and stratify forecasts into post-formation and pre-formation. A key question for improved understanding of tropical cyclones forecasts in different basins remains: is the differential skill found for West North Pacific and North Indian Ocean present in pure genesis as well as in track forecasts? Answering this has implications for understanding and interpreting TC forecasts in the region.

4.3. Flooding

Both GLOFFIS and GloFAS have issues with simulating the highest flows most relevant for humanitarians preparing for high impact events – those at a 1 in 2 year return period and higher, even when driven by a perfect meteorological forecast. This finding stands apart from existing published work on GloFAS, which generally shows the skill evaluated over the whole hydrograph (Harrigan et al., 2020), or using 'high flow' thresholds which are not fully representative of the operational GloFAS forecast products. This choice is driven by the need for robust statistics; for instance, Alfieri et al. (2013) evaluate the skill of the reforecast for 90th percentile events and state, 'Such a percentile is a good tradeoff between being representative of high flow values and including a sufficient number of events to draw robust statistics.'

However 90th percentile daily discharge is the flow level which is exceeded on average once every ten days, which is not an extreme event and of little interest to humanitarians. Operationally GloFAS provides forecasts for 1 in 2, 5 or 20 year return period events; the corresponding percentile threshold for a 1 in 2 year event is higher than the 99th percentile, and is even higher for 1 in 5, or 1 in 20. While the choice of 90th percentile is still scientifically valid, and Alfieri et al. (2013) rightfully point out the constraint of small sample sizes for the evaluation of rare high impact events, our findings demonstrate a significant decline in ability to correctly simulate high flow beyond the 90th percentile. While this may not be the case in all regions, this suggests that analysis of skill using a 90th percentile threshold is unrepresentative of these high impact events. Therefore, the existing evaluation of GloFAS showing good performance is not necessarily relevant for the operational forecasts of high impact events, where skill assessment is not generally provided. Despite the small sample size for these rare events this has been attempted here, and indicates significant overconfidence in GloFAS probabilities. If such forecasts are to be used by humanitarians, these findings must be taken into account. They also suggest areas for humanitarian action, such as utilising global models for awareness but perhaps only triggering action in cases of observed floodwater upstream.

A comparison of GLOFFIS and GloFAS historical simulations over Myanmar with gauge discharge observations indicates that GloFAS is better at simulating high flows, as this analysis suggests that forecast errors for high flow will be larger for GLOFFIS than for GloFAS. However no GLOFFIS forecasts are available for direct comparison of forecast skill. If there is any intention of using GLOFFIS forecasts, it is a priority that any available GLOFFIS historical forecasts are obtained and evaluated. Ideally a full GLOFFIS reforecast should be created and evaluated.

A lack of observational discharge data over the Philippines prevents a robust direct evaluation of GloFAS skill over the country. However analysis of GloFAS against its own reanalysis indicates slightly worse performance for the Philippines than Myanmar, although significant variation exists within each country. Stations on the Irrawaddy around Mandalay perform best in this verification, along with stations on the Salween. For the Philippines the station on the Panay river performs best, indicating priority targets for future verification. It should be ascertained if daily discharge observations exist over the Philippines (e.g. in collaboration with the Philippines Met Services), and if they do, they should be obtained and used to evaluate the GloFAS reforecasts directly. In addition the skill of any existing regional calibrated models already in use in the region should be evaluated.

Little to nothing is known about the skill of the storm surge forecasts from GLOSSIS, although the evidence suggests it is relatively accurate at simulating surge, *post hoc*. However the high uncertainty in tropical cyclone position and intensity suggests that the available deterministic GLOSSIS forecast is highly overconfident, which limits its potential use for triggering early warning. While it is not recommended to rely upon this forecast before further evaluation, there are several avenues for building an evidence base. Promisingly, Deltares keep an archive of the past year worth of deterministic GLOSSIS storm surge and have indicated this data could be provided under the data sharing platform in development.

While Deltares have not carried out assessment of these archived operational forecasts, it is recommended that an evaluation of these should be undertaken as the next step toward building trust in GLOSSIS. For example, a GLOSSIS reforecast of storm surge using ensemble reforecasts of 10m wind speed and atmospheric pressure (from the ECMWF ENS, for example) could be run and assessed. This would build understanding storm surge forecast skill and the importance of tropical cyclone position and intensity errors for storm surge. If this analysis demonstrates skill for surge prediction then it would support the development of GLOSSIS to run in probabilistic mode in real-time. Other work in this area includes Kowaleski et al. (2020), who evaluate how uncertainties in TC–induced storm tide predictions vary, and find significant variation of inundation across the ensemble, indicating low forecast confidence. However they also note the potential of dynamical TC–surge ensembles to highlight regions of high forecast confidence where early action (such as FbF) could reliably be taken.

5. Conclusions

As the world faces an uncertain future of natural hazards, early warning systems can be an effective way to prepare and protect society before extreme events. This is a critical adaptation strategy that can reduce the losses of lives and livelihoods from disasters. Within the humanitarian sector, there is a pressing need to scale up anticipatory action. Understanding how well forecasts perform is crucial for establishing effective action mechanisms such as forecast-based financing. Through evaluating the skill of forecasts for extreme rainfall, cyclones, river floods and storm surge in Myanmar and the Philippines, this study provides (1) an overview of which forecasts are most imminently useful for early action in these regions, and (2) a template for other regions to do forecast verification using limited available data. Several knowledge and data gaps directly related to the viability of forecast-based action have also been identified, which, if addressed, could improve the skill of forecasts and the potential for successful forecast-based response.

Author statement

The authors would like to gratefully acknowledge the support of World Bank grant 1262745/SEADRIF Replica: Feasibility Analysis and the German Federal Foreign Office and the German Red Cross. The authors would like to acknowledge the contributions of Nicola Ranger Cathy Ansell and in this research. Thanks to Ervin Zsoter for providing GloFAS data and to ECMWF for making ENS data available. Thanks to Joost Beckers and Deltares for provision and assistance with GLOFFIS data. We would also like to thank GRDC for providing discharge data and the Climate Hazards Group of UCSB for the CHIRPS dataset.

The authors would also like to acknowledge the support of funded projects that could use these results in the design of their Forecast-based Financing systems, including Forecast-based Financing in the Philippines funded by the German Red Cross and the project “Towards impact-based forecasting: upgrading InaSAFE and GeoSAFE to enableforecast-based action” funded by the World Bank.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.wace.2021.100325.

References

Beck, H.E., Vergopolan, N., Pan, M., Levizzani, V., van Dijk, A.I., Weedon, G.P., Brocca, L., Pappenberger, F., Huffman, G.J., Wood, E.F., 2017. Global-scale evaluation of 22 precipitation datasets using gauge observations and hydrological modeling. Hydrol. Earth Syst. Sci. 21 (12), 6201–6217.

Alfieri, L., Burek, P., Dutra, E., Krzeminski, B., Muraro, D., Thielen, J., Pappenberger, F., 2013. GloFAS–global ensemble streamflow forecasting and flood early warning. Hydrology and Earth System Sciences 17 (3), 1161–1175.

Beck, H.E., Pan, M., Roy, T., Weedon, G.P., Pappenberger, F., van Dijk, A.I., Huffman, G. J., Adler, R.F., Wood, E.F., 2019. Daily evaluation of 26 precipitation datasets using Stage-IV gauge-radar data for the CONUS. Hydrol. Earth Syst. Sci. 23 (1), 207–224.

Bloemendaal, N., Muis, S., Haarsma, R.J., Verlaan, M., Apecechea, M.I., de Moel, H., Ward, P.J., Aerts, J.C., 2019. Global modeling of tropical cyclone storm surges using high-resolution forecasts. Clim. Dynam. 52 (7–8), 5031–5044.

Chisadza, B., Tumbare, M.J., Nyabeze, W.R., Nhapi, I., 2015. Linkages between local knowledge drought forecasting indicators and scientific drought forecasting parameters in the Limpopo River Basin in Southern Africa. International Journal of Disaster Risk Reduction 12, 226–233.

Coughlan de Perez, E., Nerlander, L., Monasso, F., van Aalst, M., Mantilla, G., Muli, E., et al., 2015. Managing health risks in a changing climate: red cross operations in east Africa and Southeast Asia. Clim. Dev. 7 (3), 197–207.

Dinku, T., Funk, C., Peterson, P., Maidment, R., Tadesse, T., Gadain, H., Ceccato, P., 2018. Validation of the CHIRPS satellite rainfall estimates over eastern Africa. Q. J. R. Meteorol. Soc. 144, 292–312.

Eakin, H., 1999. Seasonal climate forecasting and the relevance of local knowledge. Phys. Geogr. 20 (6), 447–460.

Funk, C., Peterson, P., Landsfeld, M., Pedreros, D., Verdin, J., Shukla, S., Husak, G., Rowland, J., Harrison, L., Hoell, A., Michaelsen, J., 2015. The climate hazards infrared precipitation with stations—a new environmental record for monitoring extremes. Scientific data 2, 150066.

Haiden, T., Janousek, M., Bidlot, J., Buizza, R., Ferranti, L., Prates, F., Vitart, F., 2018. Evaluation of ECMWF Forecasts, Including the 2018 Upgrade. European Centre for Medium Range Weather Forecasts.

Harrigan, S., Zsoter, E., Alfieri, L., Prudhomme, C., Salamon, P., Wetterhall, F., et al., 2020. GloFAS-ERA5 operational global river discharge reanalysis 1979-present. Hydrol. Soil Sci. Hydrol.

Ipcc (Intergovernmental Panel on Climate Change), 2019. Climate Change and Land, Special Report. IPCC, Geneva.

Jolliffe, I.T., Stephenson, D.B. (Eds.), 2012. Forecast Verification: a Practitioner’s Guide in Atmospheric Science. John Wiley & Sons.

Kowaleski, A.M., Morss, R.E., Ahijevych, D., Fossell, K.R., 2020. Using a WRF-ADCIRC ensemble and track clustering to investigate storm surge hazards and inundation scenarios associated with Hurricane Irma. Weather Forecast. 35 (4), 1289–1315.

Lee, C.Y., Camargo, S.J., Vitart, F., Sobel, A.H., Tippett, M.K., 2018. Subseasonal tropical cyclone genesis prediction and MJO in the S2S dataset. Weather Forecast. 33 (4), 967–988.

Noaa, 2010. Which countries have had the most tropical cyclone hits? https://www.aoml.noaa.gov/hrd/tcfaq/E25.html.

Seadrif, 2019. Southeast Asia Disaster Risk Insurance Facility. SEADRIF/ASEAN.

Toté, C., Patricio, D., Boogaard, H., van der Wijngaart, R., Tarnavsky, E., Funk, C., 2015. Evaluation of satellite rainfall estimates for drought and flood monitoring in Mozambique. Rem. Sens. 7 (2), 1758–1776.

Titley, H.A., Bowyer, R.L., Cloke, H.L., 2020. A global evaluation of multi-model ensemble tropical cyclone track probability forecasts. Q. J. R. Meteorol. Soc. 146 (726), 531–545.

Yamaguchi, M., Nakazawa, T., Hoshino, S., 2012. On the relative benefits of a multi-centre grand ensemble for tropical cyclone track prediction in the western North Pacific. Q. J. R. Meteorol. Soc. 138 (669), 2019–2029.

Yamaguchi, M., Vitart, F., Lang, S.T., Magnusson, L., Elsberry, R.L., Elliott, G., Kyouda, M., Nakazawa, T., 2015. Global distribution of the skill of tropical cyclone activity forecasts on short-to medium-range time scales. Weather Forecast. 30 (6), 1695–1709.

Zsoter, E., Cloke, H., Stephens, E., de Rosnay, P., Muñoz-Sabater, J., Prudhomme, C., Pappenberger, F., 2019. How well do operational Numerical Weather Prediction configurations represent hydrology? J. Hydrometeorol. 20 (8), 1533–1552.

Zsoter, E., Prudhomme, C., Stephens, E., Pappenberger, F., Cloke, H., 2020. Using ensemble reforecasts to generate flood thresholds for improved global flood forecasting. Journal of Flood Risk Management, e12658.