# Evaluation of Mekong River Commission Operational Flood Forecasts, 2000-2012

Thomas C. Pagano (1)*

* Corresponding Author
Email: Thomas.C.Pagano@gmail.com
Phone: +61 04 3997 3069

1. Bureau of Meteorology
700 Collins Street,
Docklands VIC 3008
Australia

## 1. **Introduction**

The Mekong River is one of the few large rivers where its flow has not yet been drastically modified by human development. It is a complex and varied system, both naturally and institutionally, originating in the Tibetan Plateau, flowing through six countries, and discharging to the Mekong Delta in Viet Nam. The region and the River are less developed, and there are anticipated major geopolitical, economic, social, and environmental changes - such as the planned five-fold increase in reservoir storage in the next ten years (Johnston and Kummu, 2012) - to support the irrigation and hydropower needs of a rapidly growing population (Pech and Sunada, 2008). Deforestation and urbanization are likely, along with the construction of roads, embankments, and flood protection works.

Flood forecasts help the economic development of the region while mitigating flood damages and mortalities. The first flood forecasting program was established following a very large flood in 1966 (Plate and Insisiengmay, 2005), and a sequence of nearly unprecedented floods in 2000-2001 lead to the establishment of the Mekong River Commission's (MRC) Regional Flood Management and Mitigation Center (RFMMC) in Phnom Penh, Cambodia. The RFMMC and the flood forecasts it produces are part of a broader water management plan that includes both structural measures designed to keep floods away from people and non-structural measures designed to keep people away from floods.

The RFMMC generates 1 to 5 day-ahead forecasts, updated daily, during the wet season (June-October) and 1 to 7 day-ahead outlooks, updated weekly, during the dry season (November-May). It also creates qualitative flood forecasts, which describe the expectation of flooding (i.e. may not refer to a specific place but could be used for flash flood advice or for seasonal outlooks). The forecasts are bundled with recent observed data and distributed as the

Mekong Bulletin to 39 water-related government, non-government, and United Nations agencies in Viet Nam, Thailand, Lao People's Democratic Republic (PDR), and Cambodia; and made publicly available on the Internet (MRC, 2013). National television, radio broadcasting, telephone, facsimile, e-mail, websites, and newspaper networks are used to deliver flood information to the public. However, many people find it difficult to obtain real time alerts as they do not have access to email and websites (Keoduangsine and Goodwin, 2012).

Performance evaluation is a critical component of any forecasting system. Comparison of actual operational forecasts (and/or retrospectively generated hindcasts) to observations can highlight strengths and weaknesses of a system, helping to identify opportunities to improve forecasts. Performance evaluation can also show the value of forecasts to program managers and demonstrate the improvements realized from past investments in system upgrades. Users of the forecasts can consider information about the expected error of any given forecast to manage risks associated with taking action to protect against anticipated floods. Further, performance of operational systems can be compared to experimental and research systems to evaluate the potential adoption of new techniques and technologies. There have been increased calls for study of "hydrologic forecasting science" as a way for forecasts to improve our understanding of natural systems and vice versa (Welles et al., 2007).

This article is the first evaluation of the performance of the entire history of operational flood forecasts of the RFMMC. This study is intended not only as an external and independent investigation into forecast accuracy, but as a basis for considering and implementing further improvements to the RFMMC flood forecasting system. Additionally, the operational performance evaluation methods in use at RFMMC and outlined in this article may serve as templates for others in the region and overseas.

The article begins with a discussion of the study locations and the available data. It discusses the data inputs for models and tools used to generate the forecasts. It reviews past efforts at evaluating Mekong River forecasts and outlines the forecast evaluation method used here. The performance of the forecasts is then measured and the implications discussed.

## 2. Study Locations

The Mekong Basin (Figure 1) has several geographic features that make forecasting challenging. According to MRC (2005)

[FIGURE 1]

> *"Kratie is generally regarded as the point in the Mekong system where the hydrology and hydrodynamics of the river change significantly. Upstream from this point, the river generally flows within a clearly identifiable mainstream channel. In all but the most extreme flood years, this channel contains the full discharge with only local over-bank natural storage. Downstream from Kratie, seasonal floodplain storage dominates the annual regime and there is significant movement of water between channels over flooded areas, the seasonal refilling of the Great Lake and the flow reversal in the Tonle Sap. There is extreme hydrodynamic complexity in both time and space and it becomes impossible to measure channel discharge.*

> *Water levels, not flow rates and volumes, determine the movement of water across the landscape… As the water level in the mainstream falls in late September, water flows out of the lake down the Tonle Sap back into the Mekong mainstream. Nowhere else in the world is there a flow reversal this large."*

The Tonle Sap is the largest freshwater lake in Asia. The Bassac River is a distributary of the Tonle Sap and the Mekong River downstream of Phnom Penh, flowing alongside the mainstream channel.

Above Kratie, the basin is further divided at Vientiane-Nong Khai. Upstream of this point, especially in China, the catchment is relatively steep and fast responding although a snowmelt component contributes to flow in the dry season. The lower basin is dominated by wet-season runoff originating in Lao PDR. RFMMC currently produces forecasts of water level at 22 locations and discharge at 14 locations; there are no discharge forecasts below Kratie (Table 1).

[TABLE 1]

The forecast points are the locations of river gauges; additional information is necessary to translate the forecasts at gauges to water levels in the many local villages along the floodplain. Each forecast point has a defined Flood Level (e.g. 11.8 meters at Chiang Saen) at which point local and national authorities need to take urgent measures to prevent significant damage. Flood Levels are determined by the member states, with the definition of Flood Level dependent on national standards. Alarm Level is typically exceeded three days before Flood Level is reached or exceeded. Alarm Levels are determined by the RFMMC and member states based upon the defined Flood Level and an analysis of historic flood records (MRC, 2013).

In the lower parts of the basin, maximum river level is not the only flooding concern. Prolonged periods of flow above a given discharge can cause the weakening and collapse of protection dikes. Also, rice paddies can be submerged in water for 8 to 10 days and survive, but longer than that and the crop begins to die (MRC, 2005). Total annual volume of flow is sometimes used as a proxy for the damages caused by long-duration floods. The RFMMC currently only produces 1 to 5 day-ahead forecasts but there is strong interest in medium-range and seasonal forecasts.

The flow has strong seasonality with a well-defined wet season during June to October (Figure 2). The upstream station, Luang Prabang, routinely has six or more peak flows during a single season, with the greatest peak typically occurring in June. Pakse, downstream, is less variable, with fewer peaks later in the season (August is a typical peak period but in 2007 floods occurred as late as October). Tan Chau at the Viet Nam/Cambodia border and near the Delta is nearly completely dominated by the seasonal cycle and there are instances of river heights exceeding Flood Level for more than a month. When Tan Chau river height is below 2 meters (usually December-July), the station is affected by ocean tides. These tides have an effect as far upstream as Phnom Penh at the nadir of the dry season.

[FIGURE 2]

Total travel time between Chiang Saen and Phnom Penh is about 10 days (Niko Bakker, personal communication, 7 August 2013). In the steep river reach between Chiang Saen and

Vientiane, floods can travel at approximately a speed of 400 km per day. Downstream of Vientiane, the speed is half of this or less, especially near the Delta. Below Phnom Penh, depending on the level of the Tonle Sap and tides, the river can stagnate and change direction.

Rain gauge density (but not spatial distribution) in Thailand and Viet Nam is sufficient, but the networks are inadequate in Cambodia and Laos (Pengel et al., 2008). There is little automation and telemetry of measurements, in part because human observers remain relatively inexpensive and provide reliable quality data. In 2006, the RFMMC had realtime access to 20 rainfall stations across 250,000 $km^2$ between Chiang Saen and Pakse. This is less than one tenth the density recommended by the World Meteorological Organization (Malone, 2006). Runoff coefficients (runoff/precipitation) vary between 0.34 and 0.52 for individual locations, with 0.41 for the whole basin (Hapuarachchi et al., 2008).

## 3. Forecast Methods

The RFMMC relies on observed river height data as well as precipitation estimates as inputs for models and to develop situational awareness. Ground-based stations are primarily selected based on their realtime availability. In recent years, the RFMMC has expanded its use of satellite-based precipitation estimates to supplement the sparse ground-based rain gauge network. The RFMMC uses two satellite-based products from the National Oceanic and Atmospheric Administration - Satellite Rainfall Estimation and the Tropical Rainfall Measuring Mission (MRC, 2010). The RFMMC has developed statistical methods for removing bias from the satellite-based products.

The RFMMC inherited several forecasting tools, including the Streamflow Synthesis and Reservoir Regulation (SSARR, Rockwood, 1968) installed in 1967 to simulate flows in the main river from Chiang Saen to Pakse (Johnston and Kummu, 2012). Following the recommendations of a comprehensive review (Malone, 2006) the forecasting system was updated in 2008 to use additional data sources, improve and extend use of rainfall forecasts and adopt improved hydrologic models.

The RFMMC currently uses human expertise and a combination of statistical, hydrologic and hydraulic models to generate flood forecasts. Empirical methods such as statistical regression are used downstream of Pakse, for example, estimating the recent rate of change of river height at the upstream river station and regressing this against the downstream station height change to make a future forecast. The statistical model output serves as a "sanity check" for the other model outputs, but is also useful when a lack of rainfall observations prohibit the running of other models.

In 2008, the RFMMC shifted to the Delft-FEWS platform using the URBS event-based hydrologic model with Muskingum hydraulic routing (Tospornsampan et al., 2009). URBS can be forced with spatially semi-distributed station and/or satellite based rainfall. Manually-tuned loss parameters control the rates of rainfall excess. The routing model is then forced with the rainfall excess and the observed recent streamflow. MM5 (Fifth Generation Mesoscale Model operated by the US Air Force, Cox et al., 1998) gives three, 24-hourly forecasts of rainfall for consecutive days and zero rainfall is assumed subsequently (Malone, 2006).

The RFMMC also uses the ISIS hydrodynamic model, a generic one-dimensional model for the simulation of unsteady flow in channel networks, by providing an implicit numerical solver for the Saint Venant equations. At selected intervals, it computes water levels and discharges on a non-staggered grid. The ISIS model is used for forecasts from Stung Treng to the ocean, receiving tributary inflows from the URBS model. ISIS is more computationally intensive than URBS and therefore the latter is run routinely whereas ISIS is run for retrospective analyses and as demand arises.

Over time, the operational forecasters have improved and gained experience with the system. The system was tested by major floods in 2008 and 2011, after which the forecasters re-tuned the URBS model parameters. Hydrologists use their situational awareness to quality control data, adjust model parameters/outputs and synthesize the results before generating the official forecasts.

## 4. Data

The primary distribution channel of the RFMMC's forecasts is the Mekong Bulletin. The Bulletin's tables and graphics are created using templates in Excel spreadsheets. For this study, processing scripts were used to extract the numerical values of the forecasts from the spreadsheets in order to place them in a consistent structure. The layout of the spreadsheets has changed over time and is designed to be human-readable (as opposed to having a strict and consistent format for machine-readability). Therefore care was taken to visualize the end results to detect outliers and possible processing errors.

Operationally, a new spreadsheet is saved for each day's forecasts, normally named "F" with a suffix of the issue day, month and year (e.g. F21Aug09.xls). File names may have slightly different suffixes (e.g. F21Aug09_Original.xls, F21Aug09_Isis.xls). The latter may contain raw model output and not official forecasts (i.e. forecaster-approved final values that are issued to the public). The suffix "Original" was allowed in the 0.65% of cases that a normal-named file (i.e.

with no suffix) did not exist for a given date. 3,531 spreadsheets were identified as potentially containing official forecasts.

There are many examples of multiple files with the same name existing in various locations in the RFMMC operational forecasting directory structure. The union of all forecasts was retained (i.e. non-blanks overriding blanks) and in the 0.41% cases where forecasts with the same location, issue date, and lead-time conflicted, the original files were manually inspected and subjective judgment used to select the numbers that best reflect the forecaster's intent (e.g. 4.17 is more likely than exactly 0.00). The forecasters have the option to issue a "first" (i.e. provisional) forecast at 10 am and a "follow-up" forecast a few hours later. This is only done around five times per season and the metadata insufficiently distinguish first and follow-up forecasts.

This study archived the forecasts in absolute heights above Mean Sea Level and relative to the gauge datum (”Zero Gauge Levels", Table 1). The Bulletins contain these Zero Gauge Levels but when one was missing, the Zero Gauge Level was inferred from earlier and later forecasts.

The observations were collected from several sources. The Bulletins often contain observed river height for the prior two days. This is the 7:00 am reading and the data are provisional. Unfortunately, during the dry season when the forecasts are issued every seven days and only extend to seven days ahead, there will be nearly no overlap between the Bulletins' forecasts and observations (see, for example, the lack of forecast-observation pairs during the dry season in Figure 2). The RFMMC also receives four other manual readings per day along with continuous automated hourly data where available. These data are reviewed and corrected for errors and archived as a daily average in the operational database. This second source of data was time shifted to match the interpretation of the RFMMC forecasts (i.e. instantaneous height at 7:00 am). Thirdly, the IKMP (Integrated Knowledge Management Programme) of the Technical Support Division of the MRC is the long-term custodian of the data and provides July-October data for 2008-2012 on the Internet (http://ffw.mrcmekong.org/historical_rec.htm).

The observations from these three sources (Bulletins, Operational Database, and IKMP) were visualized together to discover and remove obvious outliers. The data were merged in order of priority (lowest to highest): Bulletins, Operational Database, IKMP. There are 4598 days (12.6 years) of observations for 22 stations. 21% of these observations are missing, 58% came from the Operational Database, 16% from IKMP, and 4% from the Bulletins.

Finally, the forecasts and observations were visualized together to inspect for outliers. 73 of 353,547 forecasts (roughly 1 in 5000 or 5 per year) appeared as outliers and the original Bulletins were examined to determine the cause. In 32% of cases, the Bulletins contained forecasts for a date other than what was indicated by the filename and therefore were excluded. 12% of cases resulted from a keying error (e.g. 9.3 meant to be 6.3). 57% appear to be genuine model malfunctions. For example, during 13-17 November 2011 (during the dry season), the forecast contains unreasonably low discharges in the headwaters and errors in excess of 3 meters. When available, observed flow from China is used by the RFMMC as an input to the model and it is possible that 0 inflow was entered when it should have been listed as missing. The forecasts with keying errors and model malfunctions are available to the public and therefore are an actual part of the user experience. However, for the purposes of this study all forecast outliers were removed because they are extremely rare, are not systematic, and it is hoped that attentive users would know that the forecasts are unreasonable. When forecaster intent was clear, keying errors were corrected to the likely true value.

## 5. Previous Studies

Although this article is the first evaluation of many years of operational forecasts, the RFMMC has been evaluating its forecasts for practically as long as it has been issuing them. The purpose of the evaluations has mainly been to give users a realistic view of the accuracy that can be achieved, particularly by emphasizing the high uncertainty in the forecasts with longer lead-times (Pengel et al., 2007).

Plate et al. (2008) demonstrated general evaluation concepts using water level forecasts from the SSARR model during July – October 2005 (wet season) as examples. The study included standard performance measures such as the Nash-Sutcliffe (NS, Nash and Sutcliffe, 1970). The NS is the mean squared error of the forecasts, relative to the error if the long-term average water level were used in place of forecasts (1 is perfect, 0 is no-skill). The performance was exceptional (i.e. NS 0.99 for 1 day-ahead, 0.8 for 5 day-ahead forecasts at Pakse) but this is partly because of the strong seasonality of flows. Plate et al. presented a “Quality Index”, which is similar to NS but uses persistence instead of long-term average water level as a baseline and has a reverse orientation (i.e. 0 is perfect, 1 is no-skill). The formula for this index is the same as the Coefficient of Prediction (CP, described in the next section) except the orientation is reversed. This is a more difficult baseline to outperform and Quality scores at Pakse were 0.47 for 1 day ahead degrading to 0.74 for 5 days ahead (CP of 0.53 and 0.26, respectively) . They explored progressively more difficult baselines, such as persistence extrapolated by trend of the observations.

Kanning et al. (2008) expanded on these results using operational wet-season forecasts in

2006 and 2007. Their analysis included measures of forecasting system reliability, i.e. the percentage of days a forecast was not issued at all because of a lack of real-time data (typically 20% and most often missing on weekends and holidays, as well as during extreme floods when it was unsafe to continue manual readings). Furthermore, forecast performance at Kratie was shown versus lead-time, demonstrating 1 meter standard deviation of error at 5 days ahead. Average error (i.e. bias) and error standard deviation were shown for all forecast locations, illustrating the highest error in the upper catchment and very little error downstream of Phnom Penh. Interestingly, the raw SSARR model output was compared to the performance of the official forecasts that include adjustments based on hydrologist expertise; at Stung Treng the human-adjusted forecasts had better error standard deviation (about a 10% reduction in error at 3 days ahead lead-time but no reduction at 5 days ahead) and worse bias. Sources of error were discussed and quantified, such as rainfall forecast error and stream gauge rating curve uncertainty.

Following the major system upgrade in 2008, Smith (2009) was tasked with establishing a set of performance indicators and benchmarks for the RFMMC. These include a set of forecast accuracy measures such as mean error, mean absolute error, and error standard deviation; and categorical measures such as false alarm rate and probability of detection of conditions above Flood Level. It discussed benchmark values as well as targets for the improved system. It outlined measures of the quality of service, such as the timeliness of forecast release, number of website hits, customer satisfaction indices and number of staff changes during flood season, among others. These guidelines are largely modelled after those used by the US National Weather Service (Corby et al., 2002).

Informally, the RFMMC has monitored and communicated the performance of the forecasts on a daily, weekly and monthly basis through internal discussions and teleconferences with key users. For several years now the RFMMC has also published routine “Annual Flood Season Performance Evaluation” reports and “Seasonal Flood Situation” reports describing the character of the flood season and the activities of the RFMMC. Along with narrative of the meteorological systems and flood response, these reports often compare the accuracy of the official forecasts to several other systems (e.g. the raw model output when forced with ground based rainfall observations, or the model when forced with satellite rainfall estimates, etc). They include tables of the percentage of forecasts with an acceptable level of accuracy that vary by location and lead-time (Table 2); in 2011 roughly 60% of the raw model output forecasts were acceptable. In 2009, operational (expertise-enhanced) forecasts were, in total, 73% acceptable. Tospornsampan (2009) did similar side-by-side comparisons of old and new model performance, and also measured the (poor) performance of 10 day forecasts that assume zero precipitation after day 5.

[TABLE 2]

In external studies (e.g., Hapuarachchi et al., 2008) and the RFMMC’s reports, the most commonly cited challenge for modellers and forecasters is a lack of *in situ* data. (Pengel et al., 2007) stated that climate networks in Cambodia and Lao PDR, the major water-producing areas during flood season, were being upgraded from 59 to 86 realtime rainfall stations. Even under the expanded system, the coverage would be more than 4150 km$^2$ per raingage, which would be less than one fifth the minimum density recommended by the World Meteorological Organization. RFMMC uses several remotely sensed products but the satellite-based rainfall estimates commonly differ from the *in situ* measurements and each other by 20-60% on seasonal timescales (or over 200% in extreme cases).

In operational practice, the final products from the model are examined and analysed by the flood forecaster in charge, who may change the forecast based on his judgement by utilizing his knowledge of the system, relevant information (e.g. hydro-meteorological data, satellite images, weather charts, storm forecast etc.), and past experiences. These forecaster adjustments commonly occur upstream of Kratie and have been shown to yield substantial improvements to forecast skill over the raw model output (Kanning et al., 2008).

## 6. Method

Aspects of performance of the forecasts are measured in a variety of ways in this study. The deterministic forecasts are of a continuous variable at point locations (river height measured in the morning at specific gauges). The accuracy of the forecasts is calculated using the standard deviation of the error, with 0 being a perfect value;

$$\sigma(loc, lead) = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left\{[f_i(\text{loc}, \text{lead}) - \text{o}_{i+lead}(\text{loc})] - \left[\overline{f_i(\text{loc}, \text{lead}) - \text{o}_{i+lead}(\text{loc})}\right]\right\}^2}$$

where $f_i(\text{loc}, \text{lead})$ is the forecast issued on day $i$ for a given location and lead-time (lead = 1 to 5 days). The corresponding observation occurs at $\text{o}_{i+lead}(\text{loc})$. Forecasts and/or observations are missing on some days, and statistics were only calculated on days with valid forecast-observation pairs. This measure does not consider bias (average error).

While the error standard deviation is a highly relevant evaluation measure for an individual user at a single location, this measure is often highly influenced by the hydrological characteristics of the river and is less influenced by the quality of the forecasts. For example, the difference between maximum and minimum height for Luang Prabang during 2000-2012 is 18.2 meters whereas Tan Chau did not vary by more than 5.0 meters. Murphy (1993) lists the unconditional variance of the observations (“Uncertainty”) as one of ten aspects of forecast quality - highly variable observations are intrinsically more challenging to forecast (in absolute terms) than observations with low variability.

To facilitate easier comparison of performance across locations, it is useful to normalize the results. The Nash Sutcliffe (NS) is one minus the mean squared error of the forecasts divided by the variance of the observations;

$$\mathrm{NS}(\mathrm{loc},\mathrm{lead}) = 1 - \frac{\sum_{i=1}^{N}\left\{[f_i(\mathrm{loc},\mathrm{lead}) - \mathrm{o}_{i+lead}(\mathrm{loc})] - \overline{[f_i(\mathrm{loc},\mathrm{lead}) - \mathrm{o}_{i+lead}(\mathrm{loc})]}\right\}^2}{\sum_{i=1}^{N}\left[\mathrm{o}_{i+lead}(\mathrm{loc}) - \overline{\mathrm{o}_{i+lead}(\mathrm{loc})}\right]^2}$$

An NS of 1 is perfect, 0 indicates no skill over always guessing the long-term average, and values less than 0 imply negative skill.

For slowly varying rivers and/or rivers with a strong seasonal cycle, the long-term average is an uninformative baseline. Instead, researchers commonly use a Coefficient of Persistence (CP) that is similar to NS but the baseline uses the value of the observation at the start of the forecast issuance (Kitanidis and Bras, 1980)

$$\mathrm{CP}(\mathrm{loc},\mathrm{lead}) = 1 - \frac{\sum_{i=1}^{N}\left\{[f_i(\mathrm{loc},\mathrm{lead}) - \mathrm{o}_{i+lead}(\mathrm{loc})] - \overline{[f_i(\mathrm{loc},\mathrm{lead}) - \mathrm{o}_{i+lead}(\mathrm{loc})]}\right\}^2}{\sum_{i=1}^{N}[\mathrm{o}_{i+lead}(\mathrm{loc}) - \mathrm{o}_i(\mathrm{loc})]^2}$$

This study also uses a baseline of persistence extrapolated using the trend of the two observations prior to forecast issuance:

$$\acute{f}_i(\mathrm{loc},\mathrm{lead}) = \mathrm{o}_i(\mathrm{loc}) + \mathrm{lead} * [\mathrm{o}_i(\mathrm{loc}) - \mathrm{o}_{i-1}(\mathrm{loc})]$$

RFMMC commonly calculates a Percentage Satisfactory index, measuring the percentage of forecasts where the error is less than a prescribed threshold B(loc,lead).

$$\mathrm{PS}(\mathrm{loc},\mathrm{lead}) = \frac{1}{N}\sum_{i=1}^{N} \begin{matrix} |f_i(\mathrm{loc},\mathrm{lead}) - \mathrm{o}_{i+lead}(\mathrm{loc})| < \ \mathrm{B}(\mathrm{loc},\mathrm{lead}) \rightarrow 1 \\ |f_i(\mathrm{loc},\mathrm{lead}) - \mathrm{o}_{i+lead}(\mathrm{loc})| \geq \ \mathrm{B}(\mathrm{loc},\mathrm{lead}) \rightarrow 0 \end{matrix}$$

PS of 1 is perfect and 0 is completely unsatisfactory. The thresholds depend on the user's concept of "satisfactory". They could be based on maintaining a consistent level of service (e.g. are this year's forecasts at least as good as last year's?) or based on the decision-making context (e.g. is the accuracy sufficient for planning purposes?).

Finally, perhaps the most visible and important forecasts of the RFMMC are those that predict a passing into Flood Level conditions. The continuous forecasts of water level can be converted to categorical forecasts of "Yes flood" and "No flood", based on the Flood Levels published in the Bulletins. A contingency table can then be constructed measuring the fraction of observed and/or forecast events that were correctly predicted. The false alarm rate is the fraction of times that the forecast indicated an event (e.g. flood) but no event occurred (0 is perfect). The probability of detection is the fraction of times that the forecast indicated an event, relative to all the times the event occurred (1 is perfect). The Equitable Threat Score combines hits, misses, and false alarms in a manner that considers the rarity of the event (Gandin and Murphy, 1992):

$$\mathrm{ETS} = \frac{\mathrm{H} - \mathrm{He}}{\mathrm{H} + \mathrm{FA} + \mathrm{M} - \mathrm{H}_e}$$

Where H is hits (forecasts said flood, observed was flood), M is misses (forecasts said no flood, flood occurred) and FA is false alarms (forecast said flood, no flood occurred). $\mathrm{H_e}$ is the expected hits by chance and is given by

$$\mathrm{H}_e = \frac{(\mathrm{H} + \mathrm{FA})\ (\mathrm{H} + \mathrm{M})}{\mathrm{N}}$$

Where N is the total events and non-events. For rare events, the worst value of ETS is near 0 whereas a perfect score is 1.

Throughout this study, only forecasts issued during the wet season (June to October) were evaluated. During the dry season the rivers remain predictably near baseflow and can be affected by ocean tides.

## 7. Results

Upstream of Kompong Cham, with the exception of Luang Prabang (which is the lowest accuracy location), 1 day-ahead forecasts have an error standard deviation of approximately 0.17 meters, increasing to 0.83 meters at 5 days ahead. Below Pakse, the 1 and 5 day-ahead forecasts have higher accuracy with an error standard deviation of 0.06 and 0.26 meters respectively (Figure 3). Most locations upstream of Phnom Penh have a wet-season observed standard deviation near 2.5 meters although Kratie has a value as high as 3.6 and Chiang Saen (the most upstream point) is as low as 1.4 meters. The river height at Kratie is naturally more variable than neighboring locations because of Kratie's W-shaped channel cross section and nearly vertical 15-meter tall banks. Below Phnom Penh, the observed standard deviation is typically close to 1.5 meters. Some of the observed variability is due to the seasonal cycle. The standard deviation of August observations (near the peak of the wet season) is also shown at the top of Figure 3.

[FIGURE 3]

When compared to the baseline of the long-term average, the forecasts appear exceptionally skilful; all locations except Chiang Saen have 1 day ahead NS scores greater than

0.99 (1.0 is perfect). Upstream of Kratie, 5 day ahead NS are typically 0.90, and the NS are still above 0.98 for the points downstream. Undoubtedly, a substantial amount of this apparent skill comes from the strong seasonal cycle and the slow variations of such a large river system. When compared to persistence, the skill is more modest, with CP scores between 0.4-0.8 for 1 day-ahead and 0.1-0.7 for 5 day-ahead forecasts (bottom of Figure 3). These results are similar to but somewhat better than what is reported by research models (e.g., Shahzad et al., 2009, reported NS ~ 0.9 and Persistence Index of 0.2-0.5). For lead-time 1 day, persistence extrapolated by a linear trend of the two observations prior to forecast issuance outperforms the operational forecasts for 12 out of 22 locations, however, for 2 days and greater, persistence with trend is consistently worse than simple persistence only.

Despite the large range of error standard deviations from one location to another, the CP indicates that the skill of forecasts is relatively even across the basin. There is a larger difference in 1- and 5-day ahead CP for the upstream locations than there is for the downstream locations between Kratie and Neak Luong, which may be the attributed to the greater uncertainties in initial conditions, recent and future precipitation and other meteorological influences at the smaller scale watersheds found upstream. Indeed, the lowest performing forecasts (5-days ahead at Chiang Saen) rely almost exclusively on the signal contained in observed upstream flows due to the lack of access to rainfall observations in China. Downstream, where hydraulic routing effects have a greater influence than local precipitation, there is nearly no loss of skill with leadtime. The exception is the two furthest downstream forecast points, where low flow forecasts have relatively high error when the river height is affected by the ocean (e.g. observe the poor performance of Tan Chau forecasts in June-July, relative to those in September-October in Figure 2).

As mentioned in previous sections, the RFMMC commonly reports the Percentage Satisfactory forecasts as a measure of performance. Three benchmarks are available, the first of which has been used operationally for many years ("Legacy", included in old seasonal and annual RFMMC reports), the second and third were proposed by an Australian consultant ("Malone") and a US consultant ("Operational", Table 2), the last two extend to 10 days ahead and are reported in Smith (2009). Smith's benchmarks are more stringent than the others and were intended as stretch goals after the 2008 forecast system upgrade. Smith's benchmarks have been adopted as the operational standard since 2011. All of the above benchmarks were typically based on the mean absolute error of operational forecasts and/or raw model output over a single year, rounded, and smoothed by a human expert. The long-term historical performance is shown in figure 4.

[FIGURE 4]

The challenge in measuring the Percentage Satisfactory with baselines derived from mean absolute error statistics, is that the results will depend on the distribution of errors. The Mekong's operational forecasts' errors are leptokurtic in that the absolute errors are positively skewed, more so for short lead-time forecasts. Therefore, long lead-time forecasts and forecasts at certain locations will consistently appear less satisfactory than others without any special circumstances. In contrast, basing the benchmarks on median absolute error ensures that performance at all locations and lead-times will, over the long run with a stable system, be satisfactory half of the time.

However, the existing measure is an established performance indicator at RFMMC and users are familiar with it. Adjusting the benchmarks so that forecasts are typically 50% satisfactory (instead of the current 65-80%) may leave users and program managers with the false impression of a dramatic loss of skill. Instead, this study defined new benchmarks (Table 2, right) based on the 70th percentile of historical errors at each location and lead-time for the wet-season forecasts. Values greater than 0.1 meter were rounded to the nearest 0.05 meter, and values less than 0.1 meter were rounded to the nearest 0.01 meter, to ease presentation of the results.

Compared to the existing operational benchmarks, these new benchmarks are stricter for short lead-times at nearly all locations and more lenient for long lead-times between Chiang Khan and Kratie. Compared to the Legacy benchmarks, the new benchmarks stricter at short leadtimes but relatively unchanged at long leadtimes. As can be seen in Figure 4, this study's proposed benchmarks give performance levels that are (by definition) more consistent across locations and lead-times.

The Percentage Satisfactory forecasts for all locations and lead-times are displayed versus time in Figure 5. The year-to-year variability of performance under existing benchmarks is nearly identical to that of this study's benchmark. Although there is a gradual (albeit likely insignificant) upward trend in skill between 2006 and 2012, there is no obvious cause for the higher skill in 2002-2004. Individual stations and/or leadtimes do not have significant trends for either Percentage Satisfactory or average absolute error (not shown).

[FIGURE 5]

A contingency table of Yes/No forecasts for conditions above Flood Level is shown in

Table 3. Only shown are forecasts where the preceding observation was below the Flood Level; such forecasts are the most important for users because after the flood has started there are fewer options to take protective action. Do note that further information is necessary to translate Flood Level at a specific gauge into local flood impacts directly upstream and downstream of the gauge, given that the height of the embankment varies.

[TABLE 3]

Threshold crossing events (i.e. going from non-flood to flood) are very rare; at 11 of 22 stations there has never been a forecast at any lead-time that indicated that the Flood Level would be crossed. This may be because Flood Levels are based on local vulnerability and many places are highly protected. Therefore, the collection of forecasts were pooled for all locations.

The vast majority (>99.7%) of forecasts correctly predict the persistence of below-Flood Level conditions. Forecasts with 1 day lead-time have a moderate Probability of Detecting floods (48%) and a very low False Alarm Rate (13%). Forecasts with 5 day lead-time have a lower Probability of Detection (31%) and a high False Alarm Rate (74%). The 1 day ahead forecasts have a higher ETS than 5 day forecasts. Between days 1 and 5 (i.e. days 2-4, not shown), the skill declines nearly linearly with leadtime. Although the sample sizes are very small, forecasts below Phnom Penh are somewhat better at predicting threshold crossing events than are points upstream, presumably due to the dominance of hydraulics over hydrology in the lowest reaches of the mainstream channel.

## 8. Discussion and Conclusions

This study analyzed thirteen years of data from the operational flood forecasts for 22 locations along the Mekong River. The forecasts had very low error particularly in the region downstream of Phnom Penh. When measured by standard skill scores, the forecasts perform exceptionally well, although a substantial part of this apparent skill is due to the strong seasonal cycle and the narrow natural variability at certain locations.

When compared to the baseline of a persistence forecast, the operational skill is more modest but still positive even at the longest lead-times suggesting that RFMMC could be reasonably confident in extending its lead-times beyond 5 days. At several locations, persistence with trend outperformed the 1 day-ahead operational forecasts. Given that RFMMC makes extensive use of recent observed flows when generating forecasts, this result may be partly an artefact of the real-time use of provisional data that has since been revised. In other words, persistence with trend using provisional observations (what is available in real-time) might not outperform the operational forecasts.

RFMMC currently creates an overall index of Percentage Satisfactory forecasts using an established set of (deemed) acceptable error levels. This study showed that the current benchmarks make certain locations and lead-times consistently appear to have less acceptable forecasts than others. If the error levels are based on user requirements, the existing benchmarks should be retained, otherwise minor modifications were proposed to the benchmarks to make the results more stable and consistent.

During historical forecast processing, occasional but rare outliers were detected, often resulting from keying errors or model malfunctions. RFMMC should strive to minimize keying errors by programmatically populating forecasts into product templates from a digital database (something that should be easier under new modelling software). Likewise, RFMMC should use automated routines and manual checks to prevent forcing the models with obviously bad data. The forecasts should be visualized in the context of the recent observations and historical climatology to ensure that unreasonable forecasts are not issued. For example, the recent observation can be extended into an envelope of possibilities in the future based on simple autocorrelation of historical river levels at a given location (e.g. the river depth has rarely changed more than 1 meter per day); the operational forecast can go outside this envelope if anomalous conditions are predicted (e.g. significant rainfall has occurred and/or a flood wave has been observed upstream).

These analyses would not be possible without the existence of archived forecasts. Operational agencies are strongly encouraged to systematically preserve historical operational forecasts, as well as observations, in a consistent machine-readable format to facilitate easy processing. If possible, such forecast databases should include official products as well as original model inputs and outputs. Adoption of a culture of continual forecast evaluation helps agencies in demonstrating the value of their forecasts to users and assessing the potential benefits of innovations in their forecasting systems.

## Acknowledgements

Thanks are extended to Seqwater’s Terry Malone and Deltares’s Alex Minett for their discussions of Mekong forecasting concerns during a site visit to the RFMMC in Phnom Penh during November-December 2012. Thanks are extended to RFMMC operational forecasters and managers for providing the archive of historical forecasts and observations, published reports, and review of this manuscript, particularly Dr. Lam Hung Son, Mr. Nicolaas Bakker, Mr. Hort Khieu, and Dr. Pichaid Varoonchotikul. Tanya Smith provided valuable editing assistance.

## Tables

Table 1. Characteristics of forecast points along the Mekong River. ID is the identifier in the RFMMC forecasting system and number is the identifier of the station in the MRC’s Master Catalogue. Zero level is the datum of the river gauge. Anglicised names may vary by source (e.g. Pakse versus Pakxe or Paksé). Contributing area for locations below Phnom Penh vary seasonally due to the reversal of flows.

| ID | Number | Lat. | Long. | Distance upstream (km) | Travel time to Phnom Penh (days) | Upstream area | Alarm Level | Flood Level | Zero Level | Name |
|---|---|---|---|---|---|---|---|---|---|---|
| CSA | 010501 | 20.274 | 100.089 | 2364 | 10 | 185 | 11.5 | 11.8 | 357.11 | Chiang Saen |
| LUA | 011201 | 19.893 | 102.134 | 2010 | 9 | 262 | 17.5 | 18 | 267.20 | Luang Prabang |
| CKH | 011903 | 17.900 | 101.670 | 1716 | 8.5 | 289 | 17.32 | 17.4 | 194.12 | Chiang Khan |
| VIE | 011901 | 17.931 | 102.616 | 1584 | 8 | 295 | 11.5 | 12.5 | 158.04 | Vientiane |
| NON | 012001 | 17.881 | 102.732 | 1548 | 8 | 295 | 11.4 | 12.2 | 153.65 | Nong Khai |
| PAK | 012703 | 18.376 | 103.644 | 1395 | 7 | 332 | 13.5 | 14.5 | 142.13 | Paksane |
| NAK | 013101 | 17.425 | 104.774 | 1218 | 5.5 | 365 | 12.6 | 12.7 | 130.96 | Nakhon Phanom |
| THA | 013102 | 17.396 | 104.796 | 1216 | 5.5 | 365 | 13 | 13.5 | 129.63 | Thakhek |
| SAV | 013402 | 16.583 | 104.733 | 1125 | 5 | 382 | 12 | 13 | 125.02 | Savannakhet |
| MUK | 013401 | 16.544 | 104.732 | 1123 | 5 | 382 | 12.5 | 12.6 | 124.22 | Mukdahan |
| KHO | 013801 | 15.318 | 105.500 | 909 | 3.3 | 408 | 16 | 16.2 | 89.03 | Khong Chiam |
| PKS | 013901 | 15.100 | 105.813 | 869 | 3 | 541 | 11 | 12 | 86.49 | Pakse |
| STR | 014501 | 13.533 | 105.950 | 684 | 2 | 631 | 10.7 | 12 | 36.79 | Stung Treng |
| KRA | 014901 | 12.481 | 106.018 | 561 | 1 | 647 | 22 | 23 | -1.08 | Kratie |
| KOM | 019802 | 11.995 | 105.469 | 439 | 0.5 | 653 | 15.2 | 16.2 | -0.93 | Kompong Cham |
| PRE | 020102 | 11.811 | 104.807 | 364 | | | 9.5 | 10 | 0.08 | Prek Kdam (Tonle Sap) |
| PPP | 020101 | 11.610 | 104.920 | 332 | 0 | 663 | 9.5 | 11 | 0.00 | Phnom Penh Port |
| PPB | 033401 | 11.563 | 104.935 | 332 | | | 10.5 | 12 | -1.02 | Phnom Penh (Bassac) |
| KOH | 033402 | 11.268 | 105.028 | 273 | | | 7.4 | 7.9 | 0.00 | Koh Khel (Bassac) |
| NEA | 019806 | 11.250 | 105.283 | 268 | | | 7.5 | 8 | -0.33 | Neak Luong |
| TCH | 019803 | 10.801 | 105.248 | 209 | | | 3.5 | 4.5 | 0.00 | Tan Chau |
| CDO | 039801 | 10.705 | 105.134 | 203 | | | 3 | 4 | 0.00 | Chau Doc (Bassac) |

538

Table 2: Performance benchmarks currently used operationally (left, from Smith, 2009) and proposed by this study (right). The table is ordered from upstream to downstream. The right-most numbers are the period of record standard deviation of wet season observations. Units are in centimeters.

Satisfactory forecast accuracy benchmarks

| | Operational | | | | | Pagano | | | | | Wet seas. | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| ID | 1 Day | 2 Day | 3 Day | 4 Day | 5 Day | 1 Day | 2 Day | 3 Day | 4 Day | 5 Day | observed std.dev | Name |
| CSA | 25 | 50 | 50 | 75 | 75 | 15 | 30 | 45 | 60 | 70 | 140 | Chiang Saen |
| LUA | 25 | 50 | 50 | 75 | 75 | 20 | 35 | 60 | 80 | 110 | 280 | Luang Prabang |
| CKH | 25 | 50 | 50 | 50 | 50 | 15 | 25 | 40 | 55 | 75 | 230 | Chiang Khan |
| VIE | 10 | 25 | 25 | 50 | 50 | 15 | 20 | 35 | 50 | 70 | 240 | Vientiane |
| NON | 10 | 25 | 25 | 50 | 50 | 10 | 20 | 35 | 50 | 65 | 240 | Nong Khai |
| PAK | 10 | 25 | 25 | 50 | 50 | 15 | 25 | 40 | 55 | 70 | 250 | Paksane |
| NAK | 10 | 25 | 25 | 50 | 50 | 15 | 25 | 40 | 55 | 70 | 255 | Nakhon Phanom |
| THA | 10 | 25 | 25 | 50 | 50 | 15 | 25 | 40 | 55 | 70 | 250 | Thakhek |
| SAV | 10 | 25 | 25 | 50 | 50 | 15 | 25 | 40 | 55 | 70 | 255 | Savannakhet |
| MUK | 10 | 25 | 25 | 50 | 50 | 10 | 20 | 40 | 55 | 70 | 255 | Mukdahan |
| KHO | 10 | 25 | 25 | 50 | 50 | 15 | 25 | 40 | 55 | 70 | 310 | Khong Chiam |
| PKS | 10 | 25 | 25 | 50 | 50 | 15 | 20 | 35 | 50 | 70 | 265 | Pakse |
| STR | 10 | 25 | 25 | 50 | 50 | 10 | 20 | 30 | 40 | 50 | 200 | Stung Treng |
| KRA | 10 | 25 | 25 | 50 | 50 | 15 | 20 | 35 | 50 | 70 | 360 | Kratie |
| KOM | 10 | 25 | 25 | 50 | 50 | 9 | 10 | 20 | 30 | 40 | 315 | Kompong Cham |
| PRE | 10 | 10 | 10 | 25 | 25 | 4 | 6 | 9 | 15 | 15 | 240 | (Tonle Sap) Prek Kdam |
| PPP | 10 | 10 | 10 | 25 | 25 | 5 | 7 | 10 | 15 | 20 | 235 | Phnom Penh Port |
| PPB | 10 | 10 | 10 | 10 | 25 | 5 | 7 | 10 | 15 | 20 | 235 | (Bassac) Phnom Penh |
| KOH | 10 | 10 | 10 | 10 | 25 | 3 | 4 | 6 | 10 | 15 | 160 | (Bassac) Koh Khel |
| NEA | 10 | 10 | 10 | 25 | 25 | 4 | 6 | 9 | 15 | 15 | 180 | Neak Luong |
| TCH | 10 | 10 | 10 | 10 | 25 | 3 | 5 | 8 | 10 | 15 | 130 | Tan Chau |
| CDO | 10 | 10 | 10 | 10 | 25 | 3 | 6 | 9 | 15 | 15 | 120 | (Bassac) Chau Doc |

Table 3. Contingency table of the forecast versus observed occurrence of river levels above Flood Level (defined in Table 1). All locations and years are pooled together due to the rarity of floods. The top table is for one day ahead forecasts and the bottom is for five day ahead forecasts. Forecasts are only included if observed river level was below Flood Level at the time of forecast issuance. Also shown are the False Alarm Rate (FAR), Probability of Detecting Floods (POD), and Equitable Threat Score (ETS).

| 1 Day-ahead | Event: | | FAR 13.3% |
|---|---|---|---|
| forecast: | Flood | No flood | POD 48.1% |
| Flood | 26 | 4 | ETS 44.8% |
| No flood | 28 | 34,087 | |

| 5 Day-ahead | Event: | | FAR 73.5% |
|---|---|---|---|
| forecast: | Flood | No flood | POD 31.0% |
| Flood | 31 | 86 | ETS 16.5% |
| No flood | 69 | 31,547 | |

## Figures

Figure 1. Map of forecast locations (black circles). The river channel, significant water bodies and basin boundary are shown in grey outline.

Figure 2. Time series of river height observations (black lines) and forecasts (colored dots) for Luang Prabang (top), Pakse (middle) and Tan Chau (bottom) for 2010-2011. Flood Levels and Alarm Levels are horizontal lines and vertical lines divide the wet and dry seasons. Below each plot of river heights is a plot of forecast errors (forecast – observed).

Figure 3. Error standard deviation (middle) and Coefficient of Persistence (bottom) for locations upstream (left) to downstream (right) for wet-season forecasts from 2000-2012. The top plot shows the period of record standard deviation for the wet-season observations and the observations for August (only complete forecast-observation pairs were included).

Figure 4. Percentage Satisfactory for 1 (top) and 5 (bottom) day-ahead wet-season forecasts by location. Forecasts are evaluated using four different benchmarks (colored lines). The benchmark proposed by this study (black line with large circles) is defined to give a 70% satisfactory rate over the long-term; deviations from 70% are due to the rounding of the benchmark thresholds.

Figure 5. Percentage Satisfactory for all lead-times and locations for each year (x-axis) using four different benchmarks.

# References

Corby, R. J., West Gulf River Forecast Center, Lawrence, W. E., and Arkansas-Red Basin River Forecast Center: A Categorical Flood Forecast Verification System for Southern Region RFC River Forecasts, National Weather Service, Southern Region, 17, 2002.

Cox, R., Bauer, B. L., and Smith, T.: A mesoscale model intercomparison, Bulletin of the American Meteorological Society, 79, 265-283, 1998.

Gandin, L. S., and Murphy, A. H.: Equitable skill scores for categorical forecasts, Monthly Weather Review, 120, 361-370, 1992.

Hapuarachchi, H. A. P., Takeuchi, K., Zhou, M., Kiem, A. S., Georgievski, M., Magome, J., and Ishidaira, H.: Investigation of the Mekong River basin hydrology for 1980–2000 using the YHyM, Hydrological Processes, 22, 1246-1256, 2008.

Johnston, R., and Kummu, M.: Water resource models in the Mekong Basin: A review, Water Resources Management, 26, 429-455, 2012.

Kanning, W., Pich, S., and Pengel, B.: Flood forecasting accuracy for the Mekong River Basin, 6th Annual Mekong Flood Forum Integrated approaches and applicable systems for medium-term flood forecasting and early warning in the Mekong River Basin Phnom Penh, Cambodia, 2008.

Keoduangsine, S., and Goodwin, R.: An Appropriate Flood Warning System in the Context of Developing Countries, International Journal of Innovation, Management and Technology, 3, 213-216, 2012.

Kitanidis, P. K., and Bras, R. L.: Real-time forecasting with a conceptual hydrologic model: 2. Applications and results, Water Resources Research, 16, 1034-1044, 1980.

Malone, T.: Roadmap mission for the development of a flood forecasting system for the Lower Mekong River, Mekong River Commission Flood Management and Mitigation Programme, Technical Component-Main Report, 72, 2006.

Mekong River Commission: Overview of the Hydrology of the Mekong Basin, Mekong River Commission, Vientiane, 73, 2005.

Mekong River Commission: Accuracy analysis of the NOAA’s Satellite Rainfall Estimate (SRE) and Tropical Rainfall Measuring Mission (TRMM) for flood season 2009, Regional Flood Management and Mitigation Centre, Phnom Penh, 28, 2010.

Mekong River Commission: Flood operations policy, Regional Flood Management and Mitigation Centre, Phnom Penh, 37, 2013.

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

Nash, J. E., and Sutcliffe, J. V.: River flow forecasting through conceptual models part I -- A discussion of principles, Journal of Hydrology, 10, 282-290, 1970.

Pech, S., and Sunada, K.: Population growth and natural-resources pressures in the Mekong River Basin, AMBIO: A Journal of the Human Environment, 37, 219-224, 2008.

Pengel, B., Malone, T., Katry, P., Pich, S., and Hartman, M.: Towards a new flood forecasting system for the lower Mekong river basin, 3rd South-East Asia Water Forum, Malaysia, 2007, 1-10,

Pengel, B., Tospornsampan, J., Malone, T., Hartman, M., and Janssen, A.: The Mekong River Flood Forecasting System at the Regional Flood Management and Mitigation Centre, 6th Annual Mekong Flood Forum Proceedings, 2008,

Plate, E. J., and Insisiengmay, T.: Early Warning System for the Lower Mekong River, Water international, 30, 99-107, 10.1080/02508060508691841, 2005.

Plate, E. j., and Lindenmaier, F.: Quality assessment of forecasts, Sixth Annual Flood Forum, Phnom Penh, 2008, 27-28,

Rockwood, D. M.: Application of Streamflow Synthesis and Reservoir Regulation-SSARR-program to the Lower Mekong River, US Army Corps of Engineers, 1968.

Shahzad, M., Lindenmaier, F., Ihringer, J., Plate, E., and Nestmann, F.: Statistical Flood Forecasting for the Mekong River, EGU General Assembly Conference Abstracts, 2009, 4333, Smith, G. F.: Development of Performance Indicators for the new Mekong Flood Forecasting System (FEWS-URBS-ISIS) and Mekong Flash Flood Guidance System (MRC FFG), Regional Flood Management and Mitigation Centre, Phnom Penh, 91, 2009.

Tospornsampan, J., Malone, T., Katry, P., Pengel, B., and An, H. P.: FMMP Component 1 Short And Medium-Term Flood Forecasting At The Regional Flood Management And Mitigation Centre, 7th Annual Mekong Flood Forum, Bangkok, 2009.

Welles, E., Sorooshian, S., Carter, G., and Olsen, B.: Hydrologic Verification: A Call for Action and Collaboration, Bulletin of the American Meteorological Society, 88, 503-511, doi:10.1175/BAMS-88-4-503, 2007.