Improved diagnostics for NWP verification in the tropics

Tieh-Yong Koh¹,² and Jun Siang Ng¹

Received 20 September 2008; revised 4 February 2009; accepted 23 April 2009; published 17 June 2009.

The root-mean-square error (RMSE) is often used to verify forecasts. However, its strong dependence on the observation variability makes it unsuitable for comparing model performance between regions where observation variability is much different, e.g., across vertical levels or between the midlatitudes and the tropics. The alpha index based on the tensor variance of forecast-observation discrepancy was formulated to improve on RMSE (and the closely related bias-corrected RMSE). An “error ellipse” was used to represent the random error in vector wind, yielding two other diagnostics: eccentricity and orientation. These diagnostics were applied to verify Naval Research Laboratory’s limited-area model, Coupled Ocean/Atmospheric Mesoscale Prediction System (COAMPS), for the first time in Southeast Asia. COAMPS forecasts were verified against radiosonde data from South China Sea Monsoon Experiment (SCSMEX), May–June 1998. Results revealed falling model performance as forecast time increases but little difference between forecasts at 18-km and 54-km resolution. Systematic errors in the model dynamics were suggested from the biases. The alpha indices show that (after bias correction) COAMPS performs best for wind, followed by temperature and then by dew point depression. In this tropical region, 1-day persistence forecasts were only outperformed by the model for wind predictions between 400 mb and 850 mb at forecast times less than 24 hr. The RMSE diagnostic was shown to sometimes yield misleading evaluation of the model’s performance. The wind error ellipses revealed that the random error tended to align more with the background flow than with the model bias, possibly indicating a dynamical reason for its existence.

Citation: Koh, T.-Y., and J. S. Ng (2009), Improved diagnostics for NWP verification in the tropics, J. Geophys. Res., 114, D12102, doi:10.1029/2008JD011179.

1. Introduction

Numerical weather prediction (NWP) is a science that is about a century old, as one traces its history back to the important contributions of Bjerknes [1904], Richardson [1922] and Charney et al. [1950]. In particular, the advent of the electronic computer enabled the fluid dynamical equations for the global atmosphere to be solved quickly and accurately enough for midlatitude operational weather forecast from the 1950s onward [Platzman, 1979]. Enhancement in computational power by orders of magnitude has allowed NWP models to employ higher resolution and more sophisticated physical parameterizations that are necessary for tropical weather research and forecast. Since the 1980s, limited-area mesoscale models that focus on shorter spatio-temporal-scale phenomena have been applied extensively for operation and research [Anthes et al., 1989].

In the development of any NWP model, the issue of verification occupies center-stage attention. By today, NWP skill, as measured by the correlation between 500 hPa-height anomalies between forecasts and analyses, remains high up to about one week in midlatitude regions in both hemispheres [Shapiro and Thorpe, 2004]. Such impressive achievement is understandable as (1) most land-based observation networks are spread across midlatitude North America and Eurasia and satellite remote-sensing footprints are closely tessellated in the extratropics [Kalnay, 2003], and (2) numerical models had been innovated [e.g., Phillips, 1956] and consistently improved to solve for extratropical cyclones and frontal systems which are the dominant extratropical weather. With the success of NWP in the midlatitudes, an important question arises: what is the state-of-the-art NWP skill in the tropics?

NWP verification in the tropics has a few salient features. Tropical weather is dominated by mesoscale moist convection and convective self-organization, quite unlike the potential vorticity (PV) balance dynamics of midlatitude weather [Hoskins et al., 1985]. This means that commonly used verification metrics based on height anomalies cannot be applied in the tropics. Instead, wind and temperature anomalies must be assessed independently, with particular emphasis on wind verification because monsoons [Chang, 2004] and land-sea breeze systems [Hadi et al., 2002] are dominant. Verification of humidity variables is also of primary importance in the tropics because of the pivotal role of moisture in the tropical circulation dynamics.

NWP skill in the tropics is generally poor because of two factors: (1) model imperfection, e.g., uncertain param-
eters in cumulus convection and cloud microphysics, and (2) errors in initial conditions arising from the dearth in observational data which grow exponentially with time [Yoden, 2007]. Poorer forecast skill demands that the verification process extract more information on the model errors. The root-mean-square error (RMSE) is a popular measure of NWP skill among others [e.g., Anthes et al., 1989; White et al., 1999]. However, RMSE has a few drawbacks.

[6] 1. It does not discriminate between systematic and random errors of the model.

[7] 2. As a scalar, it captures less than the full set of error information if the predicted variable is a two-dimensional vector, e.g., wind.

[8] 3. It tends to be larger when the variation in the observation is large (e.g., upper-level wind, lower-level temperature) giving an illusion that model performance is poorer and vice versa.

[9] An improvement to the error diagnostic is necessary in order to provide clearer directions for future model development and to benchmark progress in NWP skills in the tropics. The improved diagnostics can also be applied to mesoscale weather predictions in the extratropics, especially during summer months when convection is prevalent and forecasts are poorer.

[10] This paper has two objectives: (1) to develop an error diagnostic that overcomes the problems of RMSE and (2) to demonstrate the current NWP skill in the tropics by applying the improved diagnostic to the verification of a limited-area model, Coupled Ocean/Atmosphere Mesoscale Prediction System (COAMPS [COAMPS is a registered trademark of US Naval Research Laboratory]), in Southeast Asia.

[11] The COAMPS model is the operational model of the US Navy and it has been an effective NWP tool in various parts of the world: continental USA [Kong, 2002; Wetzel et al., 2001; C. H. Wash et al., Verification and evaluation of NOGAPS and COAMPS analyses and forecasts for the 24–26 January 2000 East Coast cyclone, paper presented at 18th Conference on Weather Analysis and Forecasting, American Meteorology Society, Ft. Lauderdale, Florida, 29 Jul.–2 Aug., 2001], Northwestern Pacific (M. S. Peng et al., Performance evaluation of the Naval Research Laboratory COAMPS on the forecast of Typhoon Herb in the western Pacific in 1996, paper presented at 22nd Conference on Hurricane and Tropical Meteorology, American Meteorology Society, Boston, Massachusetts, 19–23 May, 1997), the Mediterranean (J. E. Nachamkin and R. M. Hodur, Verification of short-term forecasts from the Navy COAMPS over the Mediterranean, paper presented at 15th Conference on Probability and Statistics in the Atmospheric Sciences, American Meteorology Society, Asheville, North Carolina, 8–11 May, 2000) and the Middle East [Liu et al., 2007]. However, its NWP skill in Southeast Asia has not been evaluated before. Thus COAMPS is the model of choice for this study. This work is part of a longer-term research program in Singapore to improve NWP in Southeast Asia.

[12] The rest of the paper is organized as follows: section 2 develops an error diagnostic that overcomes the problems associated with RMSE and extracts more information on model wind errors; section 3 describes the verification of COAMPS in Southeast Asia; section 4 shows the results of the verification; section 5 contains some discussion and the conclusions.

2. Vector Error Diagnostic

2.1. Bias, RMSE and Tensor Variance

[13] The discrepancy $\mathbf{D}$ between the forecast $\mathbf{F}$ and observation $\mathbf{O}$ of a vector is defined as

$$ \mathbf{D} = \mathbf{F} - \mathbf{O} $$  \hfill (1)

(NB, a scalar is a one-dimensional vector). Strictly speaking, $\mathbf{D}$ is the sum of errors in the model forecast and observation. For conceptual simplicity, we first assume that observations are perfect so that $\mathbf{D}$ is only caused by model error. The more general case is treated at the end of this section.

[14] The bias and RMSE are common diagnostics of NWP skill.

$$ \text{bias} = \mathbf{D} = \mathbf{F} - \mathbf{O} $$

$$ \text{RMSE} = \sqrt{\mathbf{D}^T \mathbf{D}} $$  \hfill (2)

For a large data set, the bias indicates the model’s systematic error. The RMSE and bias are not independent as RMSE may be decomposed into systematic and random error contributions [see, e.g., Wilks, 2006]. To represent the model’s random error, the standard deviation of the discrepancy could be calculated from a large data set.

$$ \text{std}(\mathbf{D}) = \sqrt{\frac{N}{N-1} \mathbf{D}^T \mathbf{D}} = \sqrt{\frac{N}{N-1} (\text{RMSE}^2 - \text{bias}^2)} $$  \hfill (3)

where $N$ is the number of data points. For large $N$, $\text{std}(\mathbf{D})$ is the RMSE of a bias-corrected forecast. For vector variables in general, the full “spread” of the model’s random error is measured by the tensor variance, $\text{var}(\mathbf{D})$. The square root of $\text{tr}[\text{var}(\mathbf{D})]$ is the scalar measure, $\text{std}(\mathbf{D})$. Appendix A1 reviews some basic definitions and Appendix A2 shows that

$$ \text{var}(\mathbf{D}) = \text{var}(\mathbf{F}) + \text{var}(\mathbf{O}) - \text{cov}(\mathbf{F}, \mathbf{O}) - [\text{cov}(\mathbf{F}, \mathbf{O})]^T $$  \hfill (4)

Equation (4) applied to scalars is a special case of the more general formula by Baird [1962] for the variance of a scalar function $f(x, y)$.

[15] Since $\text{var}(\mathbf{D})$ is a symmetric tensor, for a two-dimensional discrepancy vector such as (horizontal) wind, it holds altogether 3 independent pieces of information. The verification for wind using RMSE is usually carried in one of following ways: (1) to compute the RMSE of the wind vector as in equation (2) [e.g., Anthes et al., 1989], (2) to treat the east-west and north-south wind as scalars and compute their RMSE separately [e.g., Qian et al., 2003], and (3) to treat the wind speed and direction as scalars and compute their RMSE separately [e.g., Hanna and Yang, 2001; Hogrefe et al., 2001]. Methods (1), (2) and (3) do not capture the full set of error information contained in $\text{var}(\mathbf{D})$. It is also not possible to ascribe a Gaussian error model for wind in method (c) for cases of large error variations because wind direction is periodic over 360 degrees and wind speed is nonnegative.

[16] In the subfield of boundary-layer meteorology, other methods are used to estimate the variance of the direction of
Verrall is defined as the difference between the forecast \( E \) and the observation \( O \)
\[
D = E - O
\]
and the model error \( A \) is separately independent of the model error.
\[
A_{\text{tr var}} = 1/C_0 = 1/[2006] \quad \text{(cf. principal component analysis theory in the work of Ackermann, 1984, and Williams, 1983; Yamartino, 1984).}
\]

However, it is of limited use to adapt these methods to analyze wind discrepancy for verification purposes because they would suffer the same drawbacks as method (3) above.

### 2.2. \( \alpha \)-Index, Eccentricity \( \beta \), and Orientation \( \theta \)

To capture the random fluctuations of any two-dimensional vector \( A \), the spread of the fluctuations can be represented schematically as an “error ellipse” (Figure 1) with semimajor axis \( a \) and semiminor axis \( b \) equivalent to the square roots of the eigenvalues of \( \text{var}(A) \). The major and minor axes are aligned parallel to the corresponding eigenvectors which are the principal components of the discrepancy vector \( A \) (cf. principal component analysis theory in the work of Wilks [2006]). The simplest error model that may be ascribed to such a representation of the tensor variance is the elliptical Gaussian distribution on a plane, but such an error model need not be associated to the representation as in fact, distributions of meteorological variables often exhibit skewness which is not captured by \( \text{var}(D) \) or the Gaussian model.

In general, \( A \) can be the discrepancy, forecast or observation.

For \( A \) being the discrepancy \( D \), the size of the error ellipse denotes the magnitude of the random error and is measured by \( \text{std}(D) \) (cf. equation (3)).

\[
\text{std}(D) = \sqrt{\text{tr}[\text{var}(D)]} = \sqrt{a^2 + b^2} \quad \text{(5)}
\]

From equations (4) and (5), larger variations in the observation or forecast will cause \( \text{std}(D) \) to increase. To eliminate such influences, a normalized scalar measure of random error magnitude \( \alpha \) is defined as

\[
\alpha = \frac{\text{tr}[\text{var}(D)]}{\text{tr}[\text{var}(F)] + \text{tr}[\text{var}(O)]} \quad \text{(6)}
\]

where the inequality (7) is proven in Appendix A3.

\[
0 \leq \alpha \leq 2
\]

The \( \alpha \)-index quantifies prediction skill with regards to the model’s random error.

1. \( \alpha \) close to 0 denotes small random error and good agreement between the bias-corrected prediction and observation, e.g., \( F - F = O - \bar{O} \).
2. \( \alpha \) close to 1 denotes large random errors and poor agreement between the prediction and observation even after bias removal, e.g., \( \text{tr}[\text{cov}(F, O)] \approx 0 \).
3. \( \alpha \) close to 2 denotes small random errors but poor agreement between the bias-corrected prediction and observation as they tend to be opposite, e.g., \( F - F = -(O - \bar{O}) \).

For models to be physically reliable, \( \alpha \) should be substantially less than 1.

The eccentricity \( \beta \) of the ellipse representing the spread is defined as

\[
\beta = \frac{a - b}{a + b} ; \quad 0 \leq \beta \leq 1
\]

The orientation of the ellipse is specified by the angle \( \theta \) between the major axis of the ellipse and a reference direction (such as the cardinal north for wind diagnostics), measured clockwise from 0° to 180°. If the eccentricity is close to 0, the spread is nearly circular and the precise orientation of the ellipse is insignificant. Otherwise, the orientation \( \theta \) represents a tendency for the random error to align in that direction. Thus for example, if wind components \((u, v)\) were taken as independent, the axes of the ellipse would align in the north-south and east-west directions, which is evidently not true in general.

25. \( \alpha \), \( \beta \) and \( \theta \) capture all 3 pieces of information inherent in \( \text{var}(D) \) for a two-dimensional vector. For a scalar (i.e., one-dimensional vector), \( \beta \) and \( \theta \) are undefined while the expression for the \( \alpha \)-index simplifies to

\[
\alpha = \frac{\text{var}(D)}{\text{var}(O) + \text{var}(F)} \quad \text{(8)}
\]

where inequality (7) applies as before.

26. The \( \alpha \)-index for a scalar and Pearson’s correlation \( r \) are independent measures and are related by

\[
|1 - \alpha| \leq |r| \quad \text{(9)}
\]

where \((1 - \alpha)\) and \( r \) always bear the same sign (cf. Appendix A4 for the proof). Thus \( \alpha \) is more stringent than \( r \) as a measure of the magnitude of random errors because \( r \) may be close to ±1 but \( \alpha \) may still be far from 0 or 2.

### 2.3. Imperfect Observations

Now, the general case where there is an error \( E \) in the observation is considered, i.e., \( O = O_t + E \) where \( O_t \) is the true observation. The true discrepancy \( D_t \), representing only the model error is

\[
D_t = F - O_t = D + E
\]

where \( D \) is as defined as the difference between the forecast and (erroneous) observation as before. Note that in practice, \( D \) can be computed but \( E \) and \( D_t \) may be unknown.

28. Given \( D = D_t - E \) and \( O = O_t + E \), assuming the observation error \( E \) is separately independent of the model.
error $D_t$ and the true observation $O_t$, it can be shown (analogously with equation (4)) that
\[
\text{tr}[\text{var}(D_t)] = \text{tr}[\text{var}(D_t)] + \text{tr}[\text{var}(E_t)] \\
\geq \text{tr}[\text{var}(D_t)] \\
\text{tr}[\text{var}(O_t)] = \text{tr}[\text{var}(O_t)] + \text{tr}[\text{var}(E_t)] \\
\geq \text{tr}[\text{var}(O_t)]
\]

Inequalities (11) and (13) set the upper bounds on the magnitude of the model’s random error $\sqrt{\text{tr}[\text{var}(D_t)]}$ and the true variability of the observation $\sqrt{\text{tr}[\text{var}(O_t)]}$. For $\alpha \leq 1$, equations (6), (10) and (12) imply that the true $\alpha$-index is
\[
\alpha_t = \frac{\text{tr}[\text{var}(D_t)]}{\text{tr}[\text{var}(F)] + \text{tr}[\text{var}(O_t)]} \leq \frac{\text{tr}[\text{var}(D)]}{\text{tr}[\text{var}(F)] + \text{tr}[\text{var}(O)]} = \alpha
\]

Thus for $\alpha \leq 1$, $\alpha$ is an upper bound on the normalized random error of the model.

[29] The computed bias $\overline{D}$ is only one component of the true model bias $D$, since
\[
\overline{D} = D + E
\]

Subtracting $\overline{D}$ from the prediction $F$ still leaves a residual bias equal to the observation bias $E$. Thus as long as the observation bias $E$ is of smaller magnitude than the model bias $D_t$, it is worthwhile to correct for the bias $\overline{D}$.

3. NWP Verification in Southeast Asia

3.1. Model Description


[31] Twice daily COAMPS assimilation-forecast cycles were started from 00 UTC, 30 April 1998 to 12 UTC, 30 June 1998 at 54-km and 18-km resolution. Navy Operational Global Atmospheric Prediction System (NOGAPS) [Louis, 1992] provide global fields for the initial cold-start and boundary conditions for the 54-km domain. Boundary conditions for the 18-km domain were taken from the 54-km domain. The 18-km domain is shown in Figure 2. There are 61 full $\sigma_z$-levels in both domains, of which 13 lie within 1 km from the surface. 6, 12, 18 and 24 h-forecasts were made from 00 and 12 UTC daily. The 12-hr forecasts were used to warm-start the next assimilation-forecast cycle. Only the results from the 18-km domain are presented below; the results from the 54-km grid are very similar.

3.2. Observation Data

[32] The model simulation period was selected to coincide with the field phase of South China Sea Monsoon Experiment (SCSMEX) from 1 May to 30 June 1998 [Lau et al., 2000; Ding et al., 2004]. Up to four times daily radiosonde data from 20 stations were selected (Figure 2) from SCSMEX Data Archive, Colorado State University (http://tornado.atmos.colostate.edu/scsmex/), with each station providing temperature, dew point depression, wind speed and wind bearing. The spatial density of data represented by these stations implies that only predictions of synoptic-scale weather could be assessed. Data were extracted at 9 World Meteorological Organization (WMO) stipulated mandatory pressure levels (in mb): 1000, 925, 850 (lower troposphere); 700, 500, 400 (middle troposphere); 300, 200, and 100 (upper troposphere). Dew point depression at 100 mb and 200 mb was not used since instrument error was believed to be larger than forecast errors because of very low specific humidity.

[33] Only data that were flagged “good” or “interpolated” were used in the verification. Linear interpolation was further carried out to fill in missing data if data existed less than 25 mb above and below the mandatory level. To maintain a minimum level of statistical significance, a set of observations for one variable at one pressure level of one station was rejected if it contained less than 30 data points in the 61-day observation window. The data availability for a station for each variable shown in Table 1 was computed as the average number of data points per day per pressure level, rounded up to the nearest integer. Stations 3 and 14 have much less data while stations 15, 16 and 17 have much...
3.3. Additional Considerations

Model forecasts were first linearly interpolated from \(\sigma_z\) coordinates onto the mandatory pressure levels and then bilinearly interpolated onto the station locations. The interpolation errors are subsumed as part of forecast error.

A forecast corresponding to a missing observation was ignored in the estimation of the mean and variance while the corresponding discrepancy was undefined and ignored in the estimation of the bias and spread.

3.4. Verification Method

The mean and variance of the forecasts, observations and discrepancies for each variable at each pressure level were calculated over all stations in the verification period. For the wind vector \((u, v)\), where \(u\) and \(v\) are the east-west and north-south components, respectively, the eigenvalues and eigenvectors of the tensor variance were calculated to derive \(\alpha, \beta, \theta\), where is measured \(\theta\) clockwise from the cardinal north. The \(\alpha\)-indices for temperature and dew point depression were also computed for comparison. As the SCSMEX database had undergone rigorous quality control, it was assumed that the observation bias was much smaller than the model bias.

To estimate the uncertainty in the statistics computed above, the data was divided into two subsets separately comprising the odd- and even-numbered days of the month. The same statistics were computed from the two data subsets and the uncertainty is estimated to be half the difference between the two sets of statistics.

4. Results

As a caveat, the results for 12-hr and 24-hr forecasts should not be compared to those for 6-hr and 18-hr forecasts because the observation sample sizes were very different: undersampling could be an issue at for 6-hr and 18-hr forecast verification. Results for wind at 100 mb were also different from the other levels for the same reason. All figures in this section show results from the 18-km grid as the results for the 54-km grid differ only a little. The estimated uncertainties are mostly stated in the figure captions and only mentioned in the text where they appear important to the results.

Table 2. Number of Data Points Used (Max Possible: 2440) in the Verification of the Forecasts at Different Pressure Levels

<table>
<thead>
<tr>
<th>Variable</th>
<th>6-hr/18-hr Forecasts</th>
<th>12-hr/24-hr Forecasts</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1000–300 mb</td>
<td>200 mb</td>
</tr>
<tr>
<td>Temperature</td>
<td>732–913</td>
<td>865</td>
</tr>
<tr>
<td>Dew point depression</td>
<td>732–913</td>
<td></td>
</tr>
<tr>
<td>Wind</td>
<td>655–768</td>
<td>659</td>
</tr>
</tbody>
</table>

A range of values is given for the levels between 1000 mb and 300 mb.
part to the cold bias at 1000 mb by depressing surface temperature and reducing upward sensible heat flux; (II) may be the systematic error that caused in part the alternate warm and cold bias at 300 and 400 mb (cf. Figure 3b), through excessive emission of long-wave radiation by water vapor at 400 mb resulting in enhanced radiative cooling at 400 mb and enhanced radiative warming just above it. (Related excessive upward transport of low-entropy air from lower levels would also favor a cold bias at upper levels.) The moisture bias tends to be larger at 24 hr and 18 hr compared to 12 hr and 6 hr, respectively, consistent with the temperature bias tendency and hence corroborating with the idea that the moisture bias is influencing the temperature bias.

[42] In Figure 5a, weak south-westerlies and strong northeasterlies are observed in the mean at lower and upper levels, respectively, as expected for during the South China Sea summer monsoon. Tropospheric mean wind at 100 mb and 200 mb is stronger for 06/18 UTC than for 06/18 UTC, while lower and midtropospheric mean wind show little difference in strength. Figure 5b shows that the wind bias is substantial (up to ~50% at 100 mb and up to ~25% elsewhere) compared to the mean wind. At 100 mb, the bias is 1.5 to 9 times larger than elsewhere possibly because the data sample is much smaller up there. The direction of the 100 mb-wind bias is close to the direction of the observed mean wind implying the mean forecast has larger magnitude but the same direction as the mean observation. The magnitude of this bias tends to be smaller for 18-hr and 24-hr forecasts than for 6-hr and 12-hr forecasts, respectively, opposite to the case for humidity and temperature. It was found that the 100-mb wind analyses at 0 hr also had smaller bias than 6-hr and 12-hr forecasts. Thus it is unlikely that the lack of a physical process is the reason...
for the 100-mb wind bias; the underlying cause remains to be found. No clear trend with forecast time was discerned from the wind bias at other levels where mean winds are all weaker than 5 m/s and biases are all smaller than 2 m/s.

4.2. Orientation, Eccentricity and α-index

Figure 6 shows the measures of random error computed from the COAMPS forecasts and SCSMEX observations. From Figures 6a and 6b, the random error tends to align east-west (e.g., see 24-hr forecasts at 500 mb, 700 mb and 850 mb) or northeast-southwest (e.g., see 24-hr forecasts at 100 mb and 200 mb) when the eccentricity of spread $\beta$ is significantly larger ($\geq 0.06$) than the root-mean-square (r.m.s.) uncertainty of eccentricity across all vertical levels ($= 0.02$). The r.m.s. uncertainty of orientation angle $\theta$ over the levels (100 mb, 200 mb, 500 mb, 700 mb and 850 mb) where $\beta$ is significantly large is $9^\circ$ and $17^\circ$ for the 12/24-hr and 6/18-hr forecasts, respectively. Therefore the anisotropy in the spread of random error is small but clearly organized at certain levels.

One question that arises is whether the preferred direction of random error aligns more closely with the mean observed wind (i.e., seasonal background) or with the model bias (i.e., systematic error). To quantify the alignment between two sets of angles $A_n$ and $B_m$, the angles were first mapped onto complex numbers of unit magnitude, $\exp(2iA_n)$ and $\exp(2iB_m)$, to ensure the representations of the eigens directions of variance are $\pi$-periodic. The square magnitude of their complex difference, $|\exp(2iA_n) - \exp(2iB_m)|^2$, is a smoothly varying and even function of $(A_n - B_m)$ and so was chosen as a simple measure of the misalignment between $A_n$ and $B_m$. The measure was normalized by 1/4 so that it is bounded by 1 and averaged over all $n$ vertical levels with weights $w_n$ determined by eccentricity $\beta_n$ (Figure 6b) and sample size $S_n$ (Table 2). The ad-hoc quadratic form $w_n \sim (\beta_n S_n)^2$ was consistently applied to emphasize the levels which were considered to yield physically and statistically significant results, i.e., where eccentricity and sample size were large, respectively. Hence the following measure $\delta$ was finally adopted to quantify the misalignment between $A_n$ and $B_n$.

$$\sin^2 \delta = \sum_{n=1}^{N} w_n \cdot \frac{1}{4} |\exp(2iA_n) - \exp(2iB_n)|^2$$

$$= \sum_{n=1}^{N} w_n \sin^2 (A_n - B_n) ; \quad \delta \in [0, \frac{\pi}{2}] \quad (15)$$

$$w_n = w_0 \cdot (\beta_n S_n)^2 ; \quad w_0 = \sum_{n=1}^{N} (\beta_n S_n)^2$$

$\delta$ was computed variably between the mean observed wind, model bias and major axis of spread and Table 3 summarizes the result from 8 data sets (6-hr, 12-hr, 18-hr and 24-hr forecasts separately at 54-km and 18-km resolutions). The table shows that (1) the bias and mean observed wind were the least consistently aligned pair which indicates that the two are largely independent (except possibly at the tropopause level as noted earlier); and (2) the major axis of spread aligned more closely and consistently with the mean observed wind than with the model bias. Thus the random error in wind prediction tended to align with the background wind rather than with the systematic error (wherever anisotropy was significant).

The $\alpha$-indices in Figures 6c to 6e are all substantially smaller than 1 confirming that the model is physically sound and the random errors do not overwhelm COAMPS model predictions altogether. The $\alpha$-indices are about 0.3 for wind, 0.4 for temperature, and 0.6 for dew point depression. This implies that after bias correction, COAMPS performs best for wind and worst for dew point depression. Apart from the
diurnal cycle, tropical weather from one day to the next tends to be similar. Therefore a tougher test of the model performance is to compare it with 1-day persistence forecast (cf. the circles in Figures 6c to 6e). Beyond the limits of uncertainty, the 6-hr, 12-hr and 18-hr forecasts for wind outperform 1-day persistence from 400 mb to 850 mb while the forecasts at other levels and the 24-hr forecast show comparable performance with 1-day persistence. This demonstrates the model’s practically useful predictability limit for wind to be at most 24 hours. Overall, 1-day persistence performs comparably with or outperforms COAMPS for temperature and dew point depression as these variables have even less day-to-day variability than wind in Southeast Asia.

Comparing ω-indices across pressure levels, the normalized random error in wind prediction is largest near the surface, presumably because of local effects of terrain and surface roughness not captured by the model. The random error in temperature prediction maximizes between 500 mb and 700 mb while the forecasts at other levels and the 24-hr forecast show comparable performance with 1-day persistence. This demonstrates the model’s practically useful predictability limit for wind to be at most 24 hours. Overall, 1-day persistence performs comparably with or outperforms COAMPS for temperature and dew point depression as these variables have even less day-to-day variability than wind in Southeast Asia.

Comparing ω-indices across pressure levels, the normalized random error in wind prediction is largest near the surface, presumably because of local effects of terrain and surface roughness not captured by the model. The random error in temperature prediction maximizes between 500 mb and 700 mb while the forecasts at other levels and the 24-hr forecast show comparable performance with 1-day persistence. This demonstrates the model’s practically useful predictability limit for wind to be at most 24 hours. Overall, 1-day persistence performs comparably with or outperforms COAMPS for temperature and dew point depression as these variables have even less day-to-day variability than wind in Southeast Asia.

The ω-indices for all variables tend to be higher for 18-hr and 24-hr forecasts compared to 6-hr and 12-hr forecasts, respectively. This denotes that random errors in the model grow with integration time. For wind, the random error grows faster in the direction where it is larger, causing the eccentricity to increase with model integration time (i.e., 6 hr versus 18 hr and 12 hr versus 24 hr).

### 4.3. Detailed Comparisons

Only examples of 24-hr forecasts at selected pressure levels are presented for illustration. The same results could be seen from other forecast hours and pressure levels. All

| Table 3. Minimum, Maximum and Mean Values of δ (in Degrees) Derived From 8 Data Sets (6-hr, 12-hr, 18-hr, and 24-hr Forecasts Separately at 54-km and 18-km Resolutions), for Different Pairs of the Direction Angles of the Mean Observed Wind, Bias, and Major Axis of Spread |
|-----------------|-----------------|-----------------|-----------------|
|                 | Bias, Mean Obs  | Spread, Mean Obs| Spread, Bias    |
| Min             | 3.0             | 9.0             | 8.0             |
| Max             | 12.7            | 11.4            | 14.2            |
| Mean            | 8.8             | 10.3            | 10.8            |
comparisons are valid beyond the limits of uncertainty at those levels.

Figures 7 and 8 show the values of model bias and standard deviations of discrepancy, forecast and observation of temperature and dew point depression, respectively. Figure 9 illustrates the analogous information for the vector wind. Biases in all 3 variables are smaller than the variability in observation or in prediction, denoting small systematic errors. As for wind, except near the surface, the same result is implied as forecast ellipses are smaller than observation ellipses. The model’s random error std(D) is smaller than observation variability std(O) for temperature and wind away from the surface, but the reverse is true for dew point depression. For near-surface wind, the two quantities are comparable.

In Figure 9, the random error in wind shows clear anisotropy at 200 mb, with the major axis aligned along the northeast-southwest direction similar to the prevailing north-easterly wind at that level (cf. Figure 5a). At 1000 mb, the random error is practically isotropic, which may be understood as follows: Mori [1986] showed from more than 700 observation samples that boundary-layer wind possesses an isotropic Gaussian distribution when the wind persistence is small; Ibarra [1995] reasoned that boundary-layer factors like differential heating/cooling over surfaces of disparate roughness, buoyancy of overheated air bubbles and low-frequency gravity waves all tend to cause maximum randomness which favors a circular Gaussian wind distribution. Supposing that COAMPS does not capture these randomizing influences on the boundary-layer wind, the random error arising between forecast and observation would isotropic in the boundary layer, just as in Figure 9b.

4.4. RMSE, std(D) and Observation Variability

To better appreciate the advantage of the above diagnostics over RMSE and std(D) (which is essentially the “bias-corrected RMSE”), the latter statistics for COAMPS and persistence forecasts are shown together in Figure 10 with the observation variability, std(O) for wind, temperature and dew point depression. Variability in wind and dew point depression increases with height, whereas temperature variability is minimum and roughly constant with height in the mid troposphere. Figure 10 shows that the vertical trend of RMSE is dominated by that of observation variability: from equations (3), (5) and (6)

\[
\text{RMSE}^2 = \text{bias}^2 + \frac{N-1}{N} \text{std(D)}^2
\]

\[
= \text{bias}^2 + \frac{N-1}{N} \alpha \cdot \{\text{tr[var(F)]} + \text{tr[var(O)]}\}
\]

[52] For any reasonable model, forecast variability closely follows observation variability while the systematic error is substantially smaller than the random error. Thus from equation (16)

\[
\text{RMSE}^2 \approx \text{std(D)}^2 \propto \alpha \cdot \text{tr[O]} \]

\[
\Rightarrow \frac{\delta(\text{RMSE})}{\text{RMSE}} \approx \frac{\delta(\text{std(D)})}{\text{std(D)}} \approx \frac{1}{2} \frac{\text{std(O)}}{\text{std(D)}} \]

\[
\approx \frac{\delta(\text{std(O)})}{\text{std(O)}}
\]

Figure 7. Temperature bias in °C and other measures for 24-hr forecast at 18-km resolution. “Std” denotes standard deviation; D, F, and O represent discrepancy, forecast, and observation, respectively.

Figure 8. Same as in Figure 7 but for dew point depression.
where the last approximation used the fact that the fractional variation of $a$ is considerably smaller that of the standard deviation of observation (cf. Figures 6 and 10). Thus RMSE and $\text{std}(D)$ have similar values (cf. thick and thin solid lines in Figure 10) and the fractional fluctuation of both measures follow mainly that of $\text{std}(O)$ (cf. dashed line in Figure 10) beyond the limits of uncertainty in all 3 measures. Hence the use of RMSE or $\text{std}(D)$ is potentially misleading when comparing model errors where observation variability changes significantly, e.g., across vertical levels or between the tropics and midlatitudes. In fact, beyond the limits of uncertainty, Figure 6 shows that the (normalized) random error for wind and temperature is maximal at low and midlevels, respectively, while Figure 10 shows the opposite:

**Figure 9.** Wind bias and error ellipses (centered at the origin) for 24-hr forecast at 18-km resolution. The bias is denoted by the displacement of the cross from the origin. The ellipses for discrepancy (D), observation (O), and forecast (F) are in thick, thin, and dashed lines, respectively.

**Figure 10.** RMSE (thick solid lines) and $\text{std}(D)$ (thin solid lines) for wind (left), temperature (center), and dew point depression (right) of 24-hr forecast on the 18-km grid, as compared to the standard deviation of observation (dashed lines). Circles denote both RMSE and $\text{std}(D)$ for 1-day persistence forecast because the two are indistinguishable at the scale depicted. The thick and thin solid lines are virtually coincident in the left. The maximum uncertainties over all levels in $\text{std}(O)$ model RMSE, model $\text{std}(D)$, and persistence RMSE (and $\text{std}(D)$) are, respectively, 0.2 m/s, 0.4 m/s, 0.8°C, and 0.2 m/s for wind; 0.05°C, 0.05°C, 0.08°C, and 0.04°C for temperature; and 0.3°C, 0.4°C, 0.4°C, and 0.1°C for dew point depression.
RMSE and std(D) are minimal for these variables at these levels!

[53] There is another more subtle error in the use of RMSE: in Figure 6, 1-day persistence outperforms COAMPS 24-hr forecast at 700 mb for temperature and 500 mb for dew point depression. However, the RMSE and std(D) diagnostics in Figure 10 indicate the reverse! Table 4 confirms the quantitative comparison beyond the limits of uncertainty at those levels. The reason could be traced to the fact that COAMPS does not fully capture the observed variance (std(F) < std(O)) while the variance in persistence forecast is equal to the observed variance. From equation (16), at levels where the bias is small (cf. Figures 3 and 4), the smaller std(F) overcomes the larger α and leads to the smaller std(D) and hence RMSE for COAMPS than for persistence. Thus the RMSE and std(D) diagnostics ironically yield better performance for COAMPS than for persistence because COAMPS fails to fully capture atmospheric variability. Because NWP for temperature and dew point depression is particularly challenging in tropical climes, the evaluation of a model can be very sensitive to the choice of diagnostics in this region.

5. Conclusions

[54] The preceding exercise in the verification of COAMPS model in Southeast Asia demonstrated several advantages of the vector error diagnostics, chief among which is all 3 independent pieces of information on the error variance of vectors are preserved as α, β and θ: α is a normalized measure of random error (cf. equation (6)); for vector variables, β is the eccentricity or anisotropy of the spread of random error; and θ is the orientation angle of the largest principal component of random error. Error ellipses (Figures 1 and 9) are useful tools to show that the principal components of the vector discrepancies are not necessarily parallel to the axes of Cartesian coordinates (east-west wind u, north-south wind v) or polar coordinates (wind speed |w|, wind direction φ). Consequently, widely used analyses treating wind components the same way as scalars are strictly speaking incorrect as they ignore the significant covariance of the wind components.

[55] Two new but preliminary results are reported from the analysis of COAMPS model performance during the SCSMEX period: (1) the dominant principal component of random error in wind (i.e., major axis of the error ellipse represented by θ) depends more on the background atmospheric flow than on the model bias; and (2) the random error in wind is isotropic (β = 0) in the boundary layer. The background flow influencing θ may be part of the monsoon system but the influence of land-sea breezes should not be discounted since models generally do not represent such local features well. Thus future efforts may be directed toward investigating how the orientation of coastlines near observation stations may impact θ. System dynamics influencing random error growth, such as the alignment of the strain axis with the mean flow, and parameterization of boundary-layer processes to minimize anisotropy in random error are other possible avenues of research.

[56] The α-index is an improvement on the RMSE (and the “bias-corrected RMSE” or std(D)) diagnostic. With its normalized scale between 0 and 2, α measures the worth of a (bias-corrected) prediction: α should be much less than 1 for a good prediction; α more than 1 indicates wrong model physics. Thus the COAMPS model was shown to yield useful predictions for wind, temperature and dew point depression. The RMSE diagnostic is disadvantaged in that RMSE depends greatly on the variability in the observation and prediction. Hence the RMSE gives a seemingly bad score where there is large natural variability (e.g., for wind at upper levels) while it gives a misleadingly good score in more constant regimes (e.g., for temperature in the equatorial tropics). The α-index is a normalized error measure that is free of the influence of observation and forecast variability. It is a more suitable diagnostic than RMSE when evaluating tropical forecasts of NWP models that have been successfully verified in the midlatitudes.

[57] The fact that α is nondimensionalized by a natural scale of variability also enables the comparison of model performance between different variables including scalars and vectors. Therefore it is quantifiably meaningful to conclude that in Southeast Asia, COAMPS predicts decreasingly well from wind to temperature to dew point depression. A 24-hr practical predictability limit for wind was tentatively obtained in this tropical region when compared to 1-day persistence forecasts. As for temperature and dew point depression, it seems like model random errors must be reduced for this model to be practically useful, because at the same time of the day, temperature and humidity changes little from one day to the next in the tropics. COAMPS is currently being verified over a longer time base to confirm these findings.

[58] For COAMPS, wind bias is generally small except possibly at the tropopause level. Temperature and dew point depression biases are plausibly related through surface evaporation, vertical moisture and heat transport and long-wave radiative interactions: the model is excessively cold and moist near the surface, too dry with little temperature...
bias in the lower troposphere and too moist with alternate warm and cold bias in the upper troposphere. The model consistently fails to capture the full variability observed for all 3 variables assessed here. The bias and α-index show little improvement from the 54-km grid to 18-km grid, suggesting that in the tropics, limited resolution may not be the cause of systematic and random errors in state-of-the-art NWP with parameterized convection.

[59] Equipped with the bias and spread derived from a large calibration data set for an NWP model, it is possible to

\[
\text{var}(\mathbf{D}) = \frac{1}{N-1} \sum_{i=1}^{N} \begin{bmatrix}
(\Delta D_i^1)^2 & \Delta D_i^1 \Delta D_i^2 \\
\Delta D_i^1 \Delta D_i^2 & (\Delta D_i^2)^2
\end{bmatrix} = \frac{1}{N-1} \sum_{i=1}^{N} \begin{bmatrix}
(\Delta F_i^1 - \Delta O_i^1)^2 & (\Delta F_i^1 - \Delta O_i^1)(\Delta F_i^2 - \Delta O_i^2) \\
(\Delta F_i^1 - \Delta O_i^1)(\Delta F_i^2 - \Delta O_i^2) & (\Delta F_i^2 - \Delta O_i^2)^2
\end{bmatrix}
\]

\[
= \frac{1}{N-1} \sum_{i=1}^{N} \begin{bmatrix}
(\Delta F_i^1)^2 & \Delta F_i^1 \Delta F_i^2 \\
\Delta F_i^1 \Delta F_i^2 & (\Delta F_i^2)^2
\end{bmatrix} + \begin{bmatrix}
(\Delta O_i^1)^2 & \Delta O_i^1 \Delta O_i^2 \\
\Delta O_i^1 \Delta O_i^2 & (\Delta O_i^2)^2
\end{bmatrix}
\]

\[
= \text{var}(\mathbf{F}) + \text{var}(\mathbf{O}) - \text{cov}(\mathbf{F}, \mathbf{O}) - \text{cov}(\mathbf{F}, \mathbf{O})^T
\]

(1) improve a new prediction by subtracting the bias; and (2) estimate the uncertainty by putting upper and lower bounds ±std(\(\mathbf{D}\)) for scalars, or superposing the error ellipse representing var(\(\mathbf{D}\)) for vectors. The bias \(\mathbf{D}\) and spread var(\(\mathbf{D}\)) represent the complete set of error information based on underlying first and second-moment error statistics.

### Appendix A: Mathematical Details

#### A1. Basic Definitions

[60] The basic definitions are reviewed here for completeness. See the work of Feller [1968] for further elaboration and proofs.

[61] Suppose a two-dimensional real vector \(\mathbf{A}\) has \(N\) realizations.

\[\mathbf{A}_i = A_i^1 \mathbf{i} + A_i^2 \mathbf{j} \quad i = 1, 2, \ldots, N\]

where \(\mathbf{i}\) and \(\mathbf{j}\) are independent unit vectors. The mean \(\bar{\mathbf{A}}\) and deviation \(\Delta \mathbf{A}_i\) are defined as

\[\bar{\mathbf{A}} = \frac{1}{N} \sum_{i=1}^{N} \mathbf{A}_i, \quad \Delta \mathbf{A}_i = \mathbf{A}_i - \bar{\mathbf{A}}\]

[62] The variance tensor of the vector \(\mathbf{A}\) is defined in terms of the variance and covariance of the vector components as follows.

\[\text{var}(\mathbf{A}) = \begin{bmatrix}
\text{var}(A^1) & \text{cov}(A^1, A^2) \\
\text{cov}(A^1, A^2) & \text{var}(A^2)
\end{bmatrix} = \frac{1}{N-1} \sum_{i=1}^{N} \begin{bmatrix}
(\Delta A_i^1)^2 & (\Delta A_i^1)(\Delta A_i^2) \\
(\Delta A_i^1)(\Delta A_i^2) & (\Delta A_i^2)^2
\end{bmatrix}\]

\[\text{var}(\mathbf{A})\text{ is a symmetric tensor and so it has real eigenvalues and orthogonal eigenvectors. Another important property is that it is positive semidefinite, i.e., all its eigenvalues are positive or zero [Feller, 1968].}\]

[63] The covariance tensor of two vectors \(\mathbf{A}\) and \(\mathbf{B}\) is defined as

\[
\text{cov}(\mathbf{A}, \mathbf{B}) = \begin{bmatrix}
\text{cov}(A^1, B^1) & \text{cov}(A^1, B^2) \\
\text{cov}(A^2, B^1) & \text{cov}(A^2, B^2)
\end{bmatrix}\]

#### A2. Proof of Equation (4)

[64] For two real numbers, \(p\) and \(q\),

\[(p \pm q)^2 \geq 0 \quad \Rightarrow \quad |2pq| \leq p^2 + q^2\]  

\[(A3)\]

[65] Putting in different values for \(p\) and \(q\),

\[|2\Delta F_i^1 |\Delta O_i^1| \leq (\Delta F_i^1)^2 + (\Delta O_i^1)^2\]  

\[(A4)\]

\[2|\Delta F_i^1 |\Delta O_i^1| \leq (\Delta F_i^1)^2 + (\Delta O_i^1)^2\]  

\[(A5)\]

[66] Adding equations (A4) and (A5) together, summing the result for values of \(i\) from 1 to \(N\), using the triangle inequality and dividing by \(N - 1\),

\[
\frac{2\text{tr}[\text{cov}(\mathbf{F}, \mathbf{O})]}{\text{tr}[\text{var}(\mathbf{F})] + \text{tr}[\text{var}(\mathbf{O})]} \leq 1
\]

\[(A6)\]

[67] From the trace of equation (4) and the definition (6),

\[\alpha = 1 - \frac{2\text{tr}[\text{cov}(\mathbf{F}, \mathbf{O})]}{\text{tr}[\text{var}(\mathbf{F})] + \text{tr}[\text{var}(\mathbf{O})]}\]

\[(A7)\]

(A6) and (A7) imply the inequality (7).

#### A3. Proof of Inequality (7)

[68] Pearson’s correlation \(r\) is defined for scalars \(F\) and \(O\) as

\[r = \frac{\text{cov}(F, O)}{\sqrt{\text{var}(F) \cdot \text{var}(O)}}\]

\[(A8)\]
[70] Using the scalar form of equation (A7),

\[ 1 - \alpha = \frac{\text{cov}(F, O)}{2[\text{var}(F) + \text{var}(O)]} \]  

(A9)

[71] Equations (A8) and (A9) show that \( r \) and \( 1 - \alpha \) always have the same sign and are the same covariance term normalized to the geometric and arithmetic means of \( \text{var}(F) \) and \( \text{var}(O) \) respectively. Since the sum and the product of two numbers are independent, \( \alpha \) and \( r \) are independent measures insofar as forecast and observation variability are not a priori related.

[72] The last statement does not imply that \( \alpha \) and \( r \) are unrelated. Dividing equation (A9) by equation (A8) and expressing in terms of standard deviations,

\[ 1 - \alpha = \frac{2 \text{std}(F) \text{std}(O)}{r [\text{std}(F)]^2 + [\text{std}(O)]^2} \]

With the help of inequality (A3), inequality (7) is proven.

[73] Acknowledgments. The authors would like to acknowledge the contribution of George Yee Chun Wei in helping with the data processing and the data support received from Sue Chen of Naval Research Laboratory. We are also indebted to Colorado State University for access to the SCSMEX data set.

References


T.-Y. Koh and J. S. Ng, Division of Physics and Applied Physics, School of Physical and Mathematical Sciences, Nanyang Technological University, 1 Nanyang Walk, Block 3, Level 3, Singapore 637616, Singapore. (kohthy@ntu.edu.sg)