JOURNAL OF GEOPHYSICAL RESEARCH, VOL. 112, C05020, doi:10.1029/2006JC003833, 2007

Which near–surface atmospheric variable drives air–sea temperature differences over the global ocean?

A. B. Kara,[1] H. E. Hurlburt,[1] and W.-Y. Loh[2]

Received 24 July 2006; revised 1 November 2006; accepted 21 November 2006; published 11 May 2007.

[1] This paper investigates the influence of atmospheric variables (net solar radiation, wind speed, precipitation and vapor mixing ratio, all of which are at or near the sea surface) on the annual and seasonal cycle of near surface air minus sea surface temperature (Tair–Tsst) over the global ocean. The importance order of these variables is discussed using several statistical methods and two global data sets. After demonstrating that neither Tair nor Tsst exhibit any skill in determining difference between the two, a regression tree model (the so-called Generalized, Unbiased, Interaction Detection and Estimation (GUIDE) algorithm) is used to investigate influences of the atmospheric variables mentioned above in regulating Tair–Tsst. Overall, net solar radiation (sum of net shortwave and longwave radiation) at the sea surface is found to be the most important variable in driving the seasonal cycle of Tair–Tsst over the global ocean when the nonlinear relationship between Tair–Tsst and atmospheric variables is taken into account. This is true for both annual and seasonal (May through August) or monthly (November and December) timescales. Similar to the GUIDE results, a simple linear regression analysis also confirms that the net solar radiation explains most of the variance in the seasonal cycle of Tair–Tsst over most (≈50%) of the global ocean. The importance of the net solar radiation in controlling Tair–Tsst is even more significant in the regions surrounding the Kuroshio and the Gulf Stream current systems. The results presented in this paper have various implications for air–sea interaction and ocean mixed layer studies.

Citation: Kara, A. B., H. E. Hurlburt, and W.-Y. Loh (2007), Which near–surface atmospheric variable drives air–sea temperature differences over the global ocean?, *J. Geophys. Res.*, *112*, C05020, doi:10.1029/2006JC003833.

1. Introduction

[2] Differences between air temperature (Tair) near the sea surface (e.g., at 10 m above the sea surface) and sea surface temperature (Tsst) have important implications for climate studies over the global ocean. Oceans exchange energy with the atmosphere via evaporation and turbulent transfer of sensible heat. Tair–Tsst is an important controlling factor in these exchanges [*Kraus and Businger*, 1994; *Yu et al.*, 2004]. Air typically either gains (loses) heat from (to) the ocean depending on the sign of Tair–Tsst through the sensible heat flux [e.g., *Cayan*, 1992; *Fairall et al.*, 2003].

[3] In addition to the sign, the magnitude of Tair–Tsst also plays a major role in maintaining the heating/cooling processes over the ocean surface [e.g., *Send et al.*, 1987; *Soloviev and Lukas*, 1997; *Soloviev et al.*, 2001]. Therefore, heat budget studies for the ocean mixed layer and the atmospheric boundary layer above the sea surface require quantitative analysis of Tair–Tsst and factors affecting this difference. Numerical ocean modeling studies [e.g., *Murtugudde et al.*, 2002; *Barron et al.*, 2004; *Kara et al.*, 2004] generally require knowledge of Tair–Tsst for stability corrections in calculating wind stress, sensible and latent heat fluxes [*Kara et al.*, 2005].

[4] An examination of the climatological monthly means of Tair–Tsst reveals large spatial and temporal variations over the global ocean (Figure 1), but Tsst is typically warmer than Tair. The magnitude of Tair–Tsst can even be <–3°C along the Kuroshio and Gulf Stream pathways. Because this temperature difference varies regionally, in this paper we investigate how different atmospheric variables affect such changes in Tair–Tsst.

[5] As expected, differences between Tair and Tsst are closely related to processes at the air–sea interface. A typical example is that as explained in *Frankignoul* [1985], net surface heat flux, in particular a combination of latent and sensible heat fluxes involving vapor mixing ratio and Tair–Tsst values, is highly correlated with Tair–Tsst, but weakly correlated with Tsst alone over most of the mid–latitudes. This simply indicates that Tair–Tsst cannot be drived from Tair or Tsst by itself, implying the existence of other variables in its regulation.

[6] Given the increasing emphasis placed on studying the ocean's role in climate dynamics, as mentioned above, understanding the relationship between Tair and Tsst is essential. Thus, the major objective of this paper is to

[1]Oceanography Division, Naval Research Laboratory, Stennis Space Center, Mississippi, USA.
[2]Department of Statistics, University of Wisconsin, Madison, Wisconsin, USA.

0148-0227/07/2006JC003833$09.00

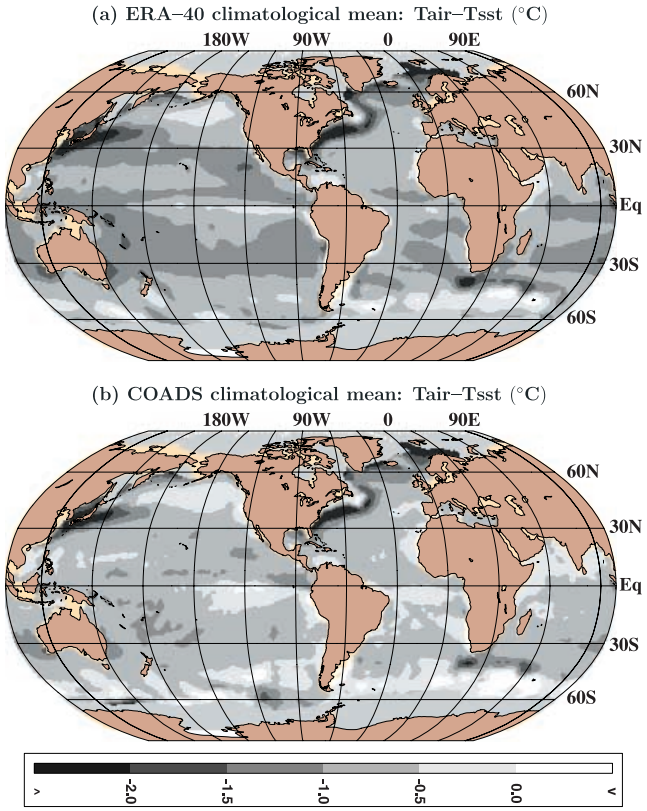

Figure 1. Climatological mean air–sea surface temperature differences over the global ocean as obtained from (a) ERA-40 and (b) COADS. Construction of both data sets are described in section 2. Tair–Tsst values at high latitudes (Arctic and Antarctic) will not be used in this paper due to the existence of ice.

address the question, "which atmospheric variable has the greatest influence on Tair–Tsst?" Possible answers to this question are sought on climatological timescales by analyzing data from two global data sets using various statistical approaches. In particular, we use atmospheric variables at/near the sea surface (net solar radiation, wind speed, vapor mixing ratio and precipitation) to examine the importance order of each one in driving the seasonal cycle of Tair–Tsst.

[7] The paper is divided into six sections. First, data sets and statistical metrics used throughout the paper are described (section 2), followed by an analysis for the relationship between Tair and Tsst (section 3). Next, the influence of atmospheric forcing variables on Tair–Tsst is investigated over the global ocean on climatological timescales (section 4), and important variables that are essential in driving the seasonal cycle of Tair–Tsst are investigated by building a regression tree model that can fit piecewise linear models (section 5). Finally, conclusions of the paper are provided (section 6).

2. Data and Statistical Metrics

[8] The relationship between Tair and Tsst (and between Tair–Tsst and various meteorological variables) is investigated using global monthly mean climatological data from two sources: (1) 1.125° × 1.125° European Centre for Medium–Range Weather Forecasts (ECMWF) 40-year Re-Analysis (ERA-40) climatology formed over the years 1957–2002, and (2) 1/2° × 1/2° Comprehensive Ocean Atmosphere Data Set (COADS) climatology formed over 1945–1989. The latter is the new COADS climatology based on the Atlas of Surface Marine Data, Supplement B (http://www.nodc.noaa.gov/OC5/asmdnew.html). Details of the archived ERA-40 (a numerical model product) and observation–based COADS data set (constructed mainly from ship observations) can be found in *Kållberg et al.* [2004] and *da Silva et al.* [1994], respectively. For consistency, climatological monthly means of Tair and Tsst from ERA-40 and COADS are interpolated to a common grid of 1.0° × 1.0° for the analyses.

[9] We directly obtain monthly mean climatological fields from COADS, while in the case of ERA-40 we construct monthly climatological data of Tair and Tsst based on 6 hourly model output covering the period 1979–2002. The data set from the ERA-40 project covering the full analysis period (1957–2002) is not used because earlier time periods did not include many observational and satellite–based data sets in the re-analysis. Note that Tair from ERA-40 is at 2 m, while that from COADS is at 10 m above the sea surface.

[10] Monthly mean climatologies of Tair–Tsst reveal existence of a strong seasonal cycle in many regions as evident from both data sets (Figure 2). For example, Tair is as much as 5°C colder than Tsst along the western boundaries of ocean basin during February and November, and Tair–Tsst can even be positive (>0°C) in May and August. Substantial seasonal variability is also evident over most of the Indian, Atlantic and Pacific Oceans. Globally, monthly mean Tair–Tsst from both data sets agree with each other reasonably well within 0.3°C (Table 1), indicating their consistency for use in this study.

[11] Given the large variability in Tair–Tsst, we will first investigate the relationship between Tair and Tsst. This is done to determine the source of difference between the two. Time series of Tair and Tsst at each ocean grid point from ERA-40 and COADS are compared using various statistical metrics: mean difference (MD), root-mean–square (RMS) difference, correlation coefficient (R) and non–dimensional skill score (SS). Let X_i (i = 1, 2,$\cdots$, n) be the set of n Tsst (reference) values, and let Y_i (i = 1, 2,$\cdots$, n) be the set of n Tair values. Also let $\bar{X}$ ($\bar{Y}$) and σ_X (σ_Y) be the means and standard deviations of Tsst (Tair) values, respectively.

[12] Following *Wilks* [1995], the statistical metrics used throughout the paper are as follows:

$$\mathrm{MD} = \bar{Y} - \bar{X}, \tag{1}$$

$$\mathrm{RMS} = \left(\frac{1}{n}\sum_{i=1}^{n}(Y_i - X_i)^2\right)^{1/2}, \tag{2}$$

$$R = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})/\sigma_X\sigma_Y), \tag{3}$$

$$\mathrm{SS} = R^2 - \underbrace{[R - (\sigma_Y/\sigma_X)]^2}_{B_{\mathrm{cond}}} - \underbrace{[(\bar{Y} - \bar{X})/\sigma_X]^2}_{B_{\mathrm{uncond}}}, \tag{4}$$

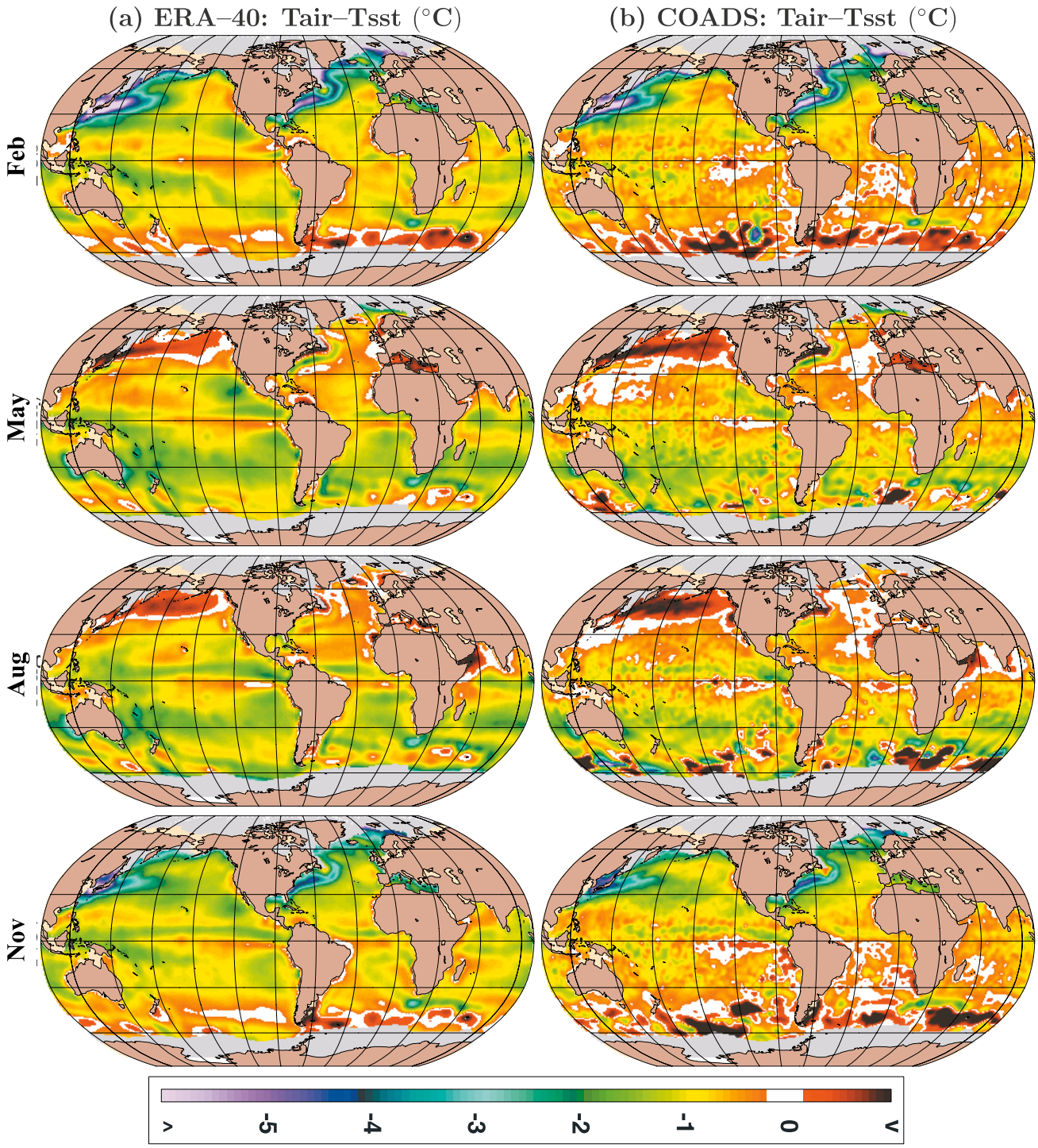

Figure 2. Climatological monthly mean air–sea surface temperature over the global ocean in February, May, August and November. They are obtained from two data sets: (a) ERA-40 and (b) COADS. The regions where ice exists are shown in gray. An ice land mask is used to determine the ice–free regions over the global ocean as explained in the text.

where n is equal to 12 (January through December) at each point of a 1° grid over the global ocean. In particular, MD is the annual mean of Tair–Tsst. RMS can be considered as an absolute measure of the distance between the Tair and Tsst time series. The R value is a measure of the degree of linear association between the two variables.

[13] SS in equation (4) is the fraction of variance in Tair explained by Tsst minus two non-dimensional biases (conditional bias, B_{cond}, and unconditional bias, B_{uncond}) which are not taken into account in the correlation coefficient. B_{uncond} (also called systematic bias) is a non-dimensional measure of the difference between the mean values of the Tair and Tsst time series, and B_{cond} is a

Table 1. Global Average and Standard Deviations of Climatological Mean Tair–Tsst[a]

	Global Mean, °C		Standard Dev., °C	
Month	ERA–40	COADS	ERA-40	COADS
Jan	−1.01	−0.69	1.05	1.06
Feb	−0.99	−0.67	0.94	0.97
Mar	−0.95	−0.65	0.67	0.68
Apr	−0.91	−0.60	0.52	0.58
May	−0.89	−0.61	0.58	0.69
Jun	−0.89	−0.66	0.68	0.82
Jul	−0.90	−0.64	0.70	0.71
Aug	−0.89	−0.58	0.59	0.73
Sep	−0.93	−0.58	0.46	0.60
Oct	−0.95	−0.59	0.50	0.68
Nov	−0.99	−0.61	0.72	0.82
Dec	−1.01	−0.68	0.97	0.99
All	−0.94	−0.63	0.70	0.78

[a]Mean and standard deviation values are calculated only in ice–free regions over the global ocean. The last row (All) denotes values calculated over the seaonal cycle.

measure of the relative amplitude of the variability in the two data sets. An examination of the SS formulation reveals that R^2 is equal to SS only when B_{cond} and B_{uncond} are zero. Because these two biases are never negative, the R value can be considered to be a measure of the "potential" skill in using Tsst to estimate Tair, i.e., the skill that one can obtain when there are no differences between Tair and Tsst. A SS value of 1.0 indicates that Tair and Tsst are identical, and SS can be negative if there is no skill between Tair and Tsst.

3. Statistical Relationship Between Tair and Tsst

[14] Comparisons between Tair and Tsst (Figure 3) are performed using the data sets (ERA-40 and COADS) and statistical metrics, both of which are already described in section 2. Regions where ice is present (e.g., high latitudes) are masked and shown in gray. The ice–free regions over the global ocean are determined from an ice land mask [*Reynolds et al.*, 2002]. The ice land mask is a function of the ice analysis and may change periodically. For this reason, a climatological mean of maximum ice extent for the mask is used in all calculations.

[15] Ignoring high latitudes where sea–ice forms, the mean difference fields are broadly similar to each other with Tsst warmer than Tair (generally by <1°C) nearly everywhere over the global ocean. Tsst is warmer because solar radiation is absorbed more efficiently by the ocean (and land) than by the atmosphere (i.e., troposphere is heated from below). In addition, warm Tsst relative to the subsurface usually gives stable stratification, while the situation is opposite for the atmosphere. Having Tsst warmer than Tair simply explains that average sensible heat flux is almost always cooling (warming) the ocean (atmosphere) on climatological timescales.

[16] A striking feature of Figure 3 is that relatively large Tair–Tsst values (even as large as −5°C) do exist in mid-latitudes along the western boundaries (Kuroshio and Gulf Stream pathways), where the RMS difference between the two is generally >3°C. Atmospheric advection of Tair and oceanic advection of Tsst play an important role in determining Tair–Tsst in these regions [*Yasuda et al.*, 2000; *Qu et al.*, 2004; *Dong and Kelly*, 2004]. Overall, the results from ERA-40 and COADS are similar except for differences in some regions, such as the southwestern Pacific, including some regions of the Indian Ocean and high southern latitudes. Such discrepancies are generally seen from maps of RMS, SS and B_{cond}. Not surprisingly, the discrepancies tend to occur in regions of sparse observations especially at high latitudes.

[17] There is a close relationship (large R) between Tair and Tsst over the annual cycle in most regions (Figure 3). However, there is almost no skill between the two in three major regions: (i) most of the tropics, extending even to mid–latitudes in some places, (ii) along the western boundaries, and (iii) at high southern and North Atlantic latitudes. The low skill in all three regions is due mainly to large differences in the mean Tair and Tsst values (i.e., large B_{uncond} over the seasonal cycle). For COADS, B_{cond} and low or even negative R values play a substantial role in giving low or negative skill in the western equatorial Pacific warm pool and at high southern latitudes. Overall, R values are generally very high (>0.9) over most of the global ocean. However, further analysis reveals that when Tair >27°C (i.e., regions around the equator) and Tair <5°C (i.e., high southern latitudes), R values are typically <0.5, especially for COADS.

[18] Based on the zonally averaged statistical metrics between Tair and Tsst (Figure 4), it is further confirmed that ERA-40 and COADS give similar results over most of the global ocean. However, noticeable differences do show up in SS values south of 40°S. As mentioned earlier, this is due partly to the fact that the seasonal cycle of Tair and Tsst (i.e., R) is quite different between the two data sets at those latitude bands. The combination of different R and B_{cond} values in the two data sets results in the large differences in skill score south of 40°S.

[19] We also present scatter diagrams for Tair versus Tsst, Tair versus Tair–Tsst, and Tsst versus Tair–Tsst (Figure 5). This is done to further examine the relationship between Tair and Tsst and decide which one (Tair or Tsst) controls Tair–Tsst. There is a strong linear relationship between Tair and Tsst with a R value >0.99 for both ERA-40 and COADS over the global ocean. Unlike Tair versus Tsst, there is no linear relationship between Tair (or Tsst) and Tair–Tsst ($R \approx 0$), suggesting that neither Tair nor Tsst modulates Tair–Tsst over the global ocean. This simply indicates that, as expected, there must be other factors that control Tair–Tsst, at least in some regions of the global ocean.

4. Effects of Atmospheric Variables on Tair–Tsst

[20] The results in the preceding section demonstrate that the climatological mean of Tair–Tsst must be controlled by variables other than Tair or Tsst itself. Thus, our focus here is to examine the possible effects of near–surface atmospheric variables in driving the seasonal cycle of Tair–Tsst. We consider several scalar atmospheric variables: wind speed at 10 m above the sea surface, air mixing ratio at 10 m above the sea surface, net radiation (the total of net shortwave and net longwave radiation) at the sea surface, Tair, Tsst and precipitation at the sea

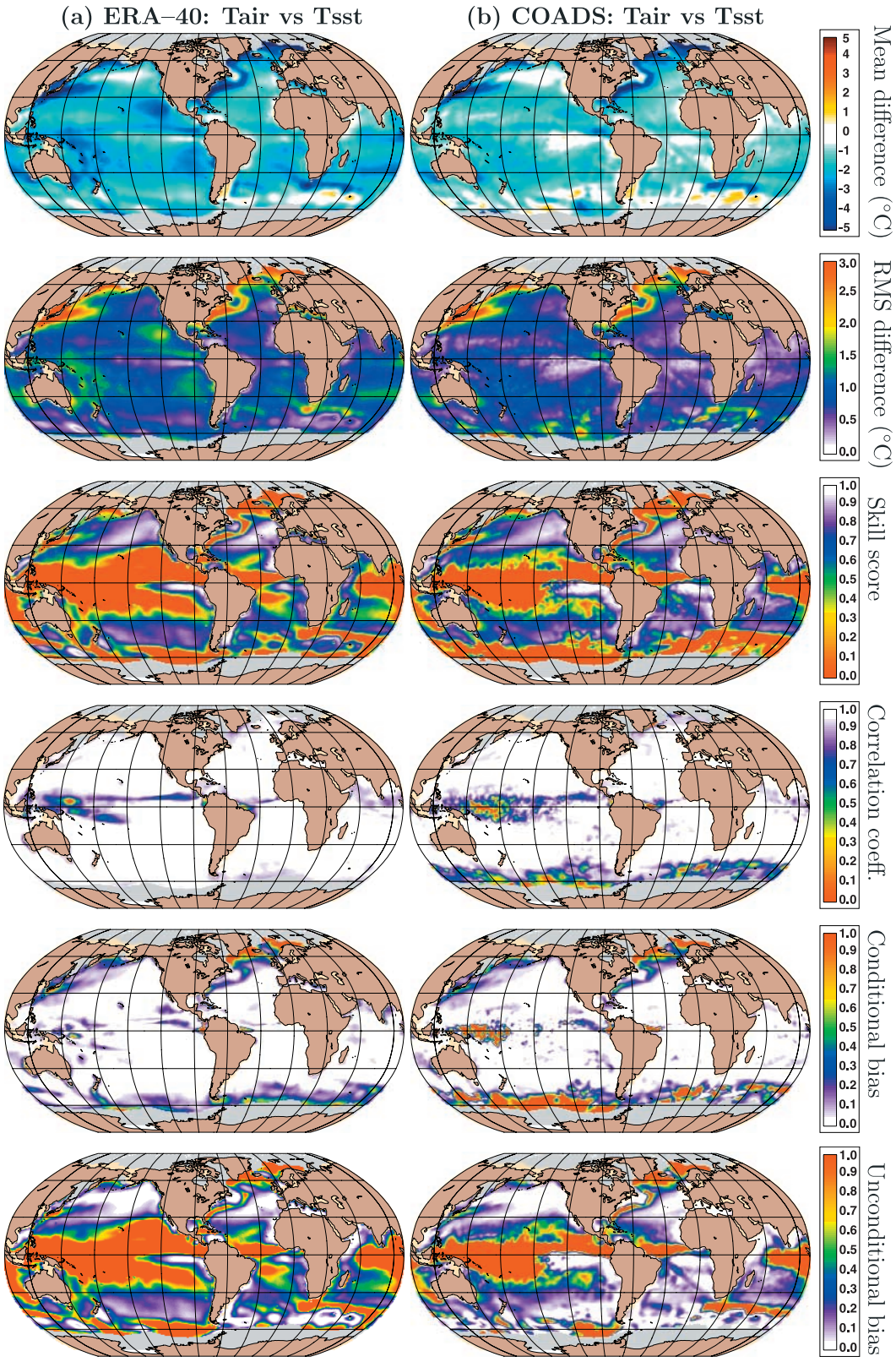

Figure 3. Spatial maps of statistical metrics (see section 2) calculated between climatological monthly means of Tair and Tsst over the global ocean. In the maps, except for the mean difference, white (red) is intended to represent a tendency for good (poor) relationship between the two variables.

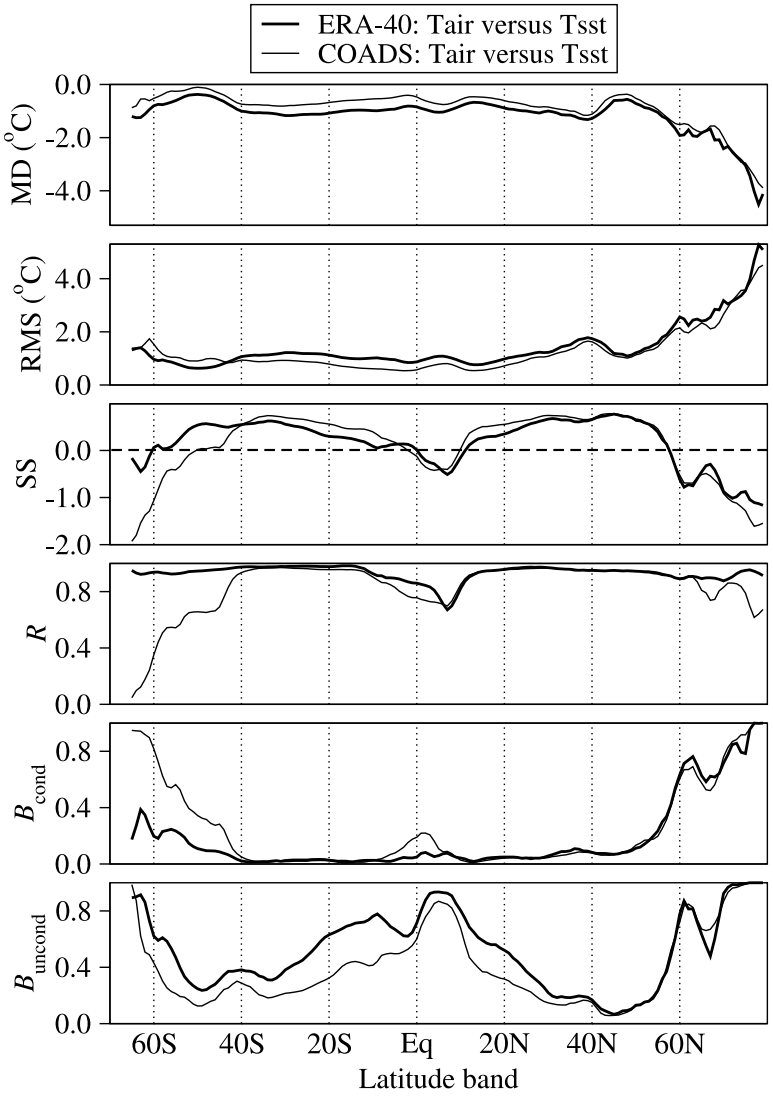

Figure 4. Zonal averages of statistical metrics shown in Figure 3. Zonal averaging was performed at each 1° latitude belt over the global ocean.

surface. These variables are specifically chosen because they are typically used as atmospheric forcing for coupled ocean–atmosphere and ocean general circulations models [e.g., *Haidvogel and Bryan*, 1992]. Monthly mean of these variables obtained from the COADS and ERA-40 data sets are interpolated to a common 1.0° × 1.0° global grid.

[21] Linear correlation coefficients between Tair–Tsst and atmospheric forcing variables mentioned above are calculated at each 1.0° × 1.0° grid box. They are then mapped over the global ocean (Figure 6). Correlation values for the seasonal cycle based on equation (3) in section 2 reveal a strong (or weak) positive (or negative) relationship between Tair–Tsst and other variables. There is a strong and positive relationship between Tair–Tsst and net solar radiation at the sea surface, especially from the mid– to high latitudes. Note that one must have at least an R value of ±0.53 for it to be statistically different from a zero correlation ($R = 0$) at a 95% confidence level based on the 12 monthly values at each grid point over the global ocean.

[22] While there are relatively large differences in R values from ERA-40 and COADS in some regions (e.g., southern hemisphere), the fields are broadly similar over most of the global ocean (Figure 6). In some regions, R values are high even when there are noticeable differences in the magnitudes of atmospheric variables from the two data sets. This is due to the fact that the shape and phase of the seasonal cycles for all of the atmospheric variables, except precipitation, are very similar illustrated for the latitude belt at 30°N in (Figure 7).

[23] The lowest and statistically insignificant R values are generally found in the equatorial regions. These statistically insignificant R values in comparison to those at other latitudes are also evident from the zonally–averaged correlation values (Figure 8). This is true regardless of the atmospheric variable correlated with Tair–Tsst. Therefore, it appears that none of the atmospheric variables modulate Tair–Tsst in that region. On the contrary, Tair–Tsst is generally driven by combination of all atmospheric variables which may be linearly

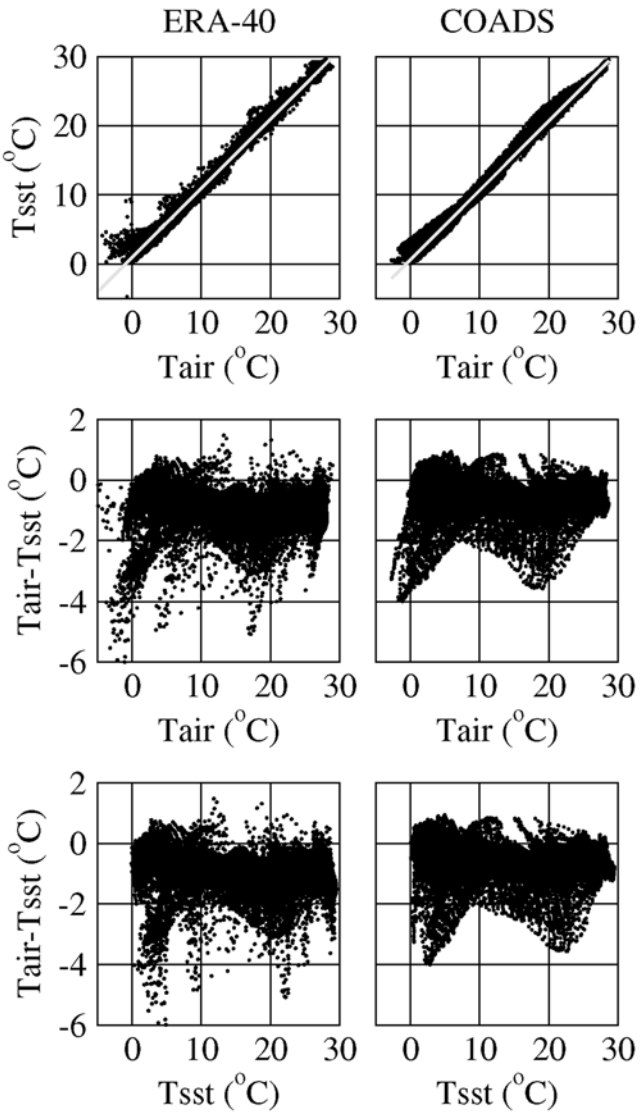

Figure 5. Scatterplots for Tair versus Tsst, Tair versus Tair−Tsst and Tsst versus Tair−Tsst based on annual mean values at each 1° bin over the global ocean. A linear regression model for Tair−Tsst, fitted to the 31671 points, gives constant slope of 1.0 for both data sets. Note that the large spread between Tair and Tsst is mostly due to the values along the western boundary currents. Linear regression equations are Tsst = 0.93 + 1.00 Tair for ERA-40 and Tsst = 0.63 + 1.00 Tair for COADS.

dependent themselves along the western boundaries including the Gulf Stream and Kuroshio current systems as evident from the significantly large R values close to 1. Wind speed at 10 m above the sea surface and net solar radiation at the sea surface are the two main variables tracked by Tair−Tsst at the subtropical northern and southern latitudes.

[24] An interesting feature that is evident from Figure 6 is that unlike other variables, wind speed and precipitation have strong negative correlations with Tair−Tsst over most of the global ocean. Figure 9 shows the cumulative frequency of correlations between Tair−Tsst and all variables. As mentioned previously, R is calculated over the seasonal cycle. Overall, 38% (28%) of R values between Tair−Tsst and wind speed are <-0.6 for ERA-40 (COADS), and similarly 36% (28%) of R values between Tair−Tsst and precipitation are <-0.6 for ERA-40 (COADS) over the global ocean (Table 2).

[25] Median values are calculated to obtain a quantitative analysis of correlations between Tair−Tsst and other variables over the global ocean. All R values are ordered from lowest to highest value, and the middle value corresponding to 50% is picked to find the median correlation for each case. Median R using data from ERA-40 (COADS) is the highest with a value of 0.84 (0.85) when Tair−Tsst is correlated to net solar radiation at the sea surface (Table 3). Since the median R values are not that large between Tair−Tsst and all other variables, net solar radiation at the sea surface plays an important role in maintaining the seasonal cycle of Tair−Tsst over the global ocean. Regardless of which data set (ERA-40 or COADS) is used for calculating R, the median values are generally very close to each other, except for wind speed, and global averages of R are almost same (Table 3). This confirms the robustness and consistency of the results.

[26] Using R values presented in Figure 6, we calculate the overall percentage of variance explained by each atmospheric variable. This is done to quantitatively identify the strongest of these predictors. Square of correlation values (i.e., R^2) between each atmospheric variable and Tair−Tsst are first obtained at each grid point. The maximum of them is then determined. This process is repeated at each grid point over the global ocean, yielding a map of the most important variables (Figure 10). Overall, most of the variance in the seasonal cycle of Tair−Tsst is again explained by the net solar radiation at the sea surface for 50.8% (57.0%) of the global ocean, when ERA-40 (COADS) data are used. There are some regional differences though. For example, wind speed is the most effective of the variables over 18.7% of the global ocean when using ERA-40, while the percentage amount drops to less than half that value (9.1%) for the COADS data set. Other regional differences exist in both data sets, especially at high southern latitudes where both data sets suffer from a lack of quality observational data.

5. What Controls the Climatological Mean of Tair−Tsst?

[27] As explained in section 4, in addition to net solar radiation at the sea surface, Tair−Tsst is strongly correlated with wind speed and air mixing ratio at 10 m, Tair and Tsst depending on the region of the global ocean (Figure 8). In most cases, there is more than one variable that has a direct effect (or a linearly dependent effect) on Tair−Tsst because the linear correlation values between Tair−Tsst and two or more variables can be quite high for a given grid point over most of the global ocean (see Figure 6).

[28] Given the results in section 4, two main questions arise: (1) which atmospheric forcing variable is the most important one driving the seasonal cycle of Tair−Tsst over the global ocean?, and (2) what is the importance order of each variable in affecting Tair−Tsst? To answer these questions, we will build a prediction algorithm for Tair−

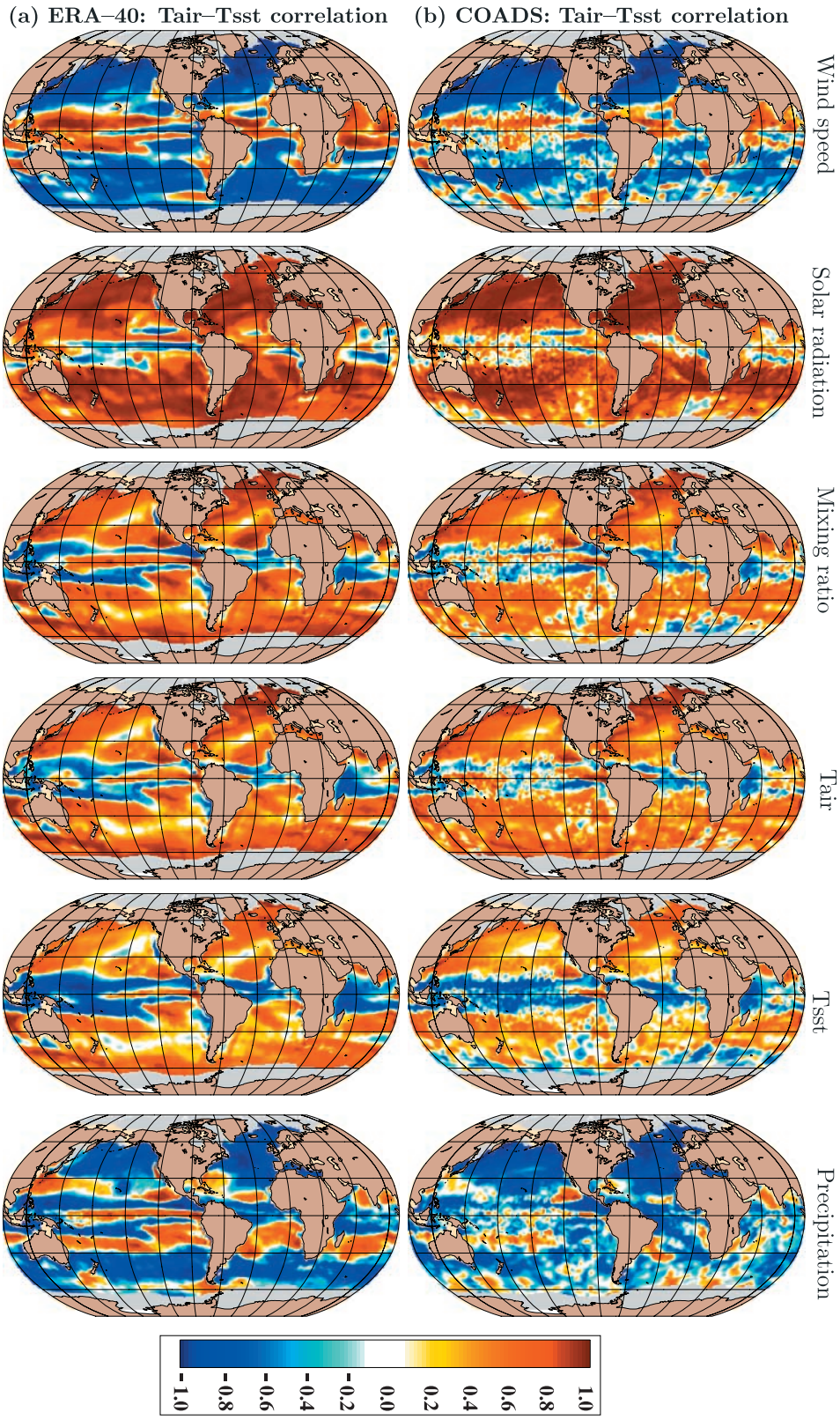

Figure 6. Correlation coefficients between Tair–Tsst and atmospheric variables calculated over the seasonal cycle. Positive (negative) correlations are in red (blue).

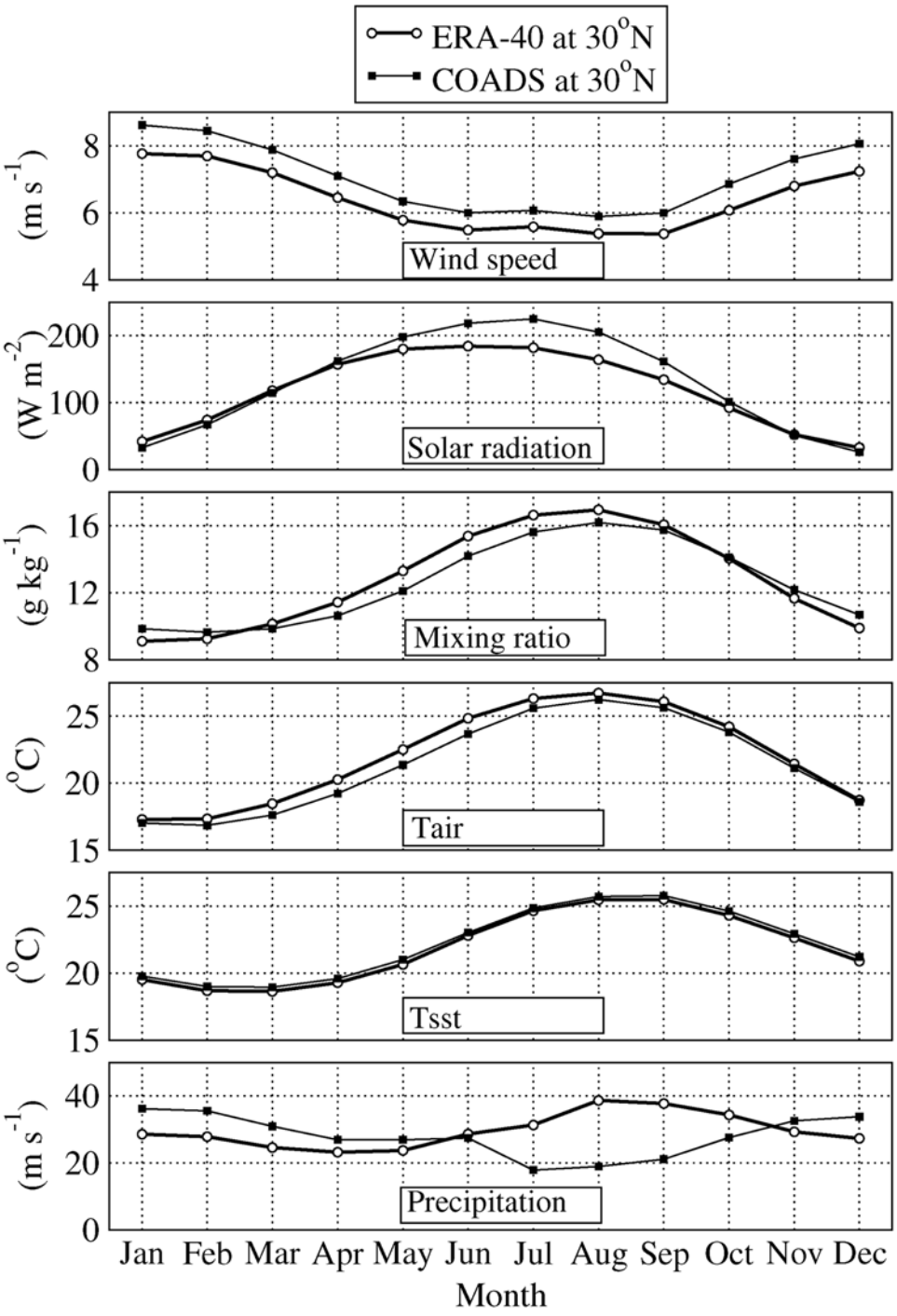

Figure 7. Climatological monthly mean time series for atmospheric forcing variables averaged over the latitude belt of 30°N from top to bottom: wind speed at 10 m above the sea surface, net solar radiation at the sea surface, vapor mixing ratio at 10 m above the sea surface, Tair, Tsst, and precipitation ($\times 10^9$) at the sea surface. Results for ERA-40 (open circles) and COADS (filled squares) are shown separately.

Tsst (section 5.1), and discuss importance order for each atmospheric variable (section 5.2).

[29] The six variables discussed in section 4 are considered as potential predictors in this proposed prediction algorithm. The methodology should have several characteristics that ensure its usefulness and validity. Among the most important of these considerations is that the methodology allow for statistical significance testing by way of cross validation and nonlinear relationships between predictors and Tair–Tsst. The methodology also needs to provide useful and interpretable results. Methods such as linear programming do not allow for validation of the results, while purely statistical methods, such as regression

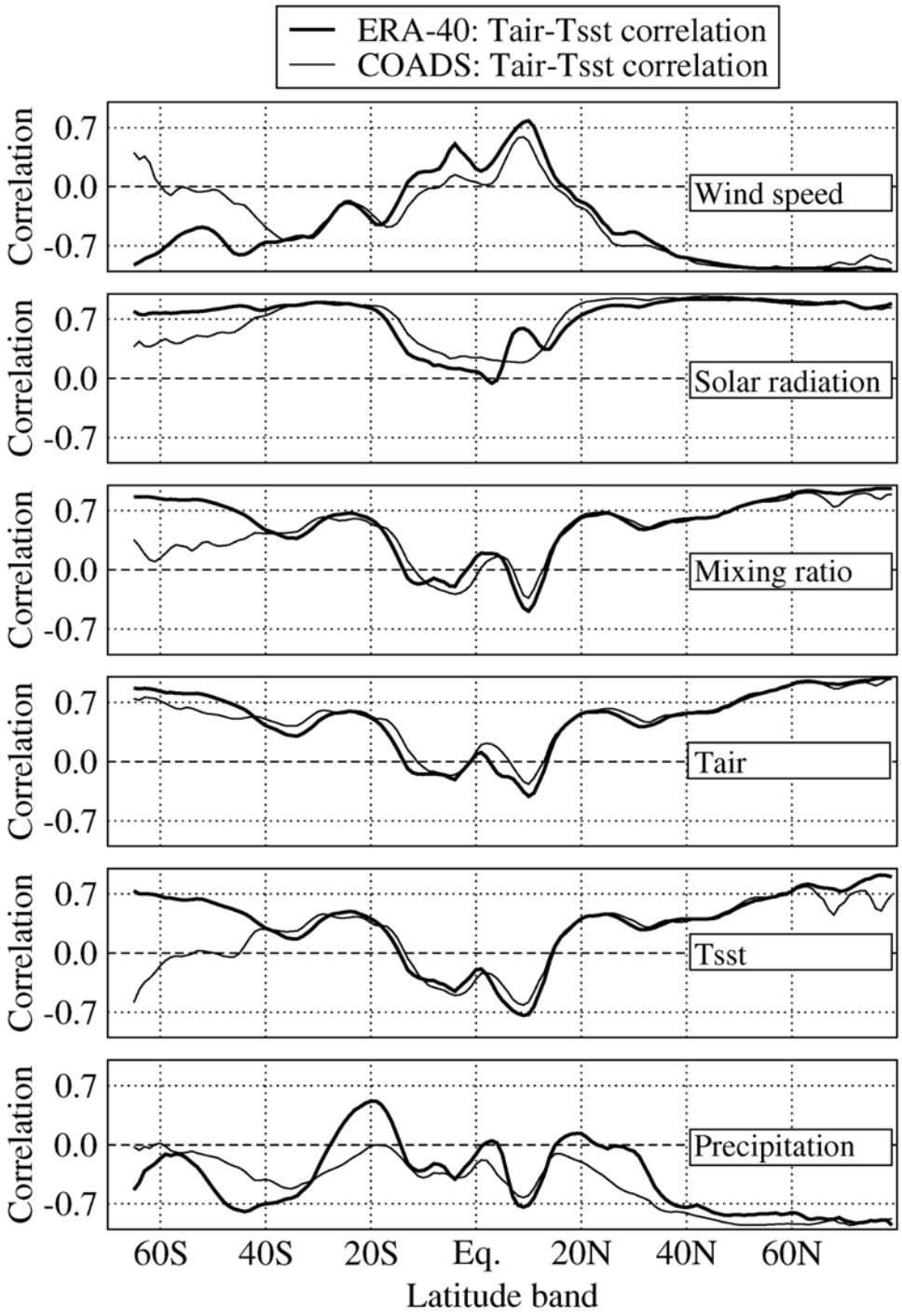

Figure 8. Zonal averages of correlation coefficients shown in Figure 6. Zonal averaging was performed at each 1° latitude belt over the global ocean.

and discriminant analysis do not easily allow for nonlinear relationships [*Breiman et al.*, 1984].

5.1. Prediction Methodology

[30] The importance order of atmospheric variables in driving the seasonal cycle of Tair–Tsst is examined using Generalized, Unbiased, Interaction Detection and Estimation (GUIDE), which is a regression tree model [*Loh*, 2002]. A brief description of GUIDE is given in Appendix A. The goal of a regression tree is to predict or explain the effect of one or more variables on a dependent variable. GUIDE can fit piecewise linear models for Tair–Tsst based on the atmospheric variables. In essence, a regression tree is a piecewise constant or piecewise linear estimate of a regression function, constructed by recursively partitioning the data set and sample space.

[31] In the GUIDE analysis as applied to this investigation, we use climatological monthly means of the dependent variable (Tair–Tsst) and six predictors (wind speed at 10 m above the sea surface, net solar radiation at

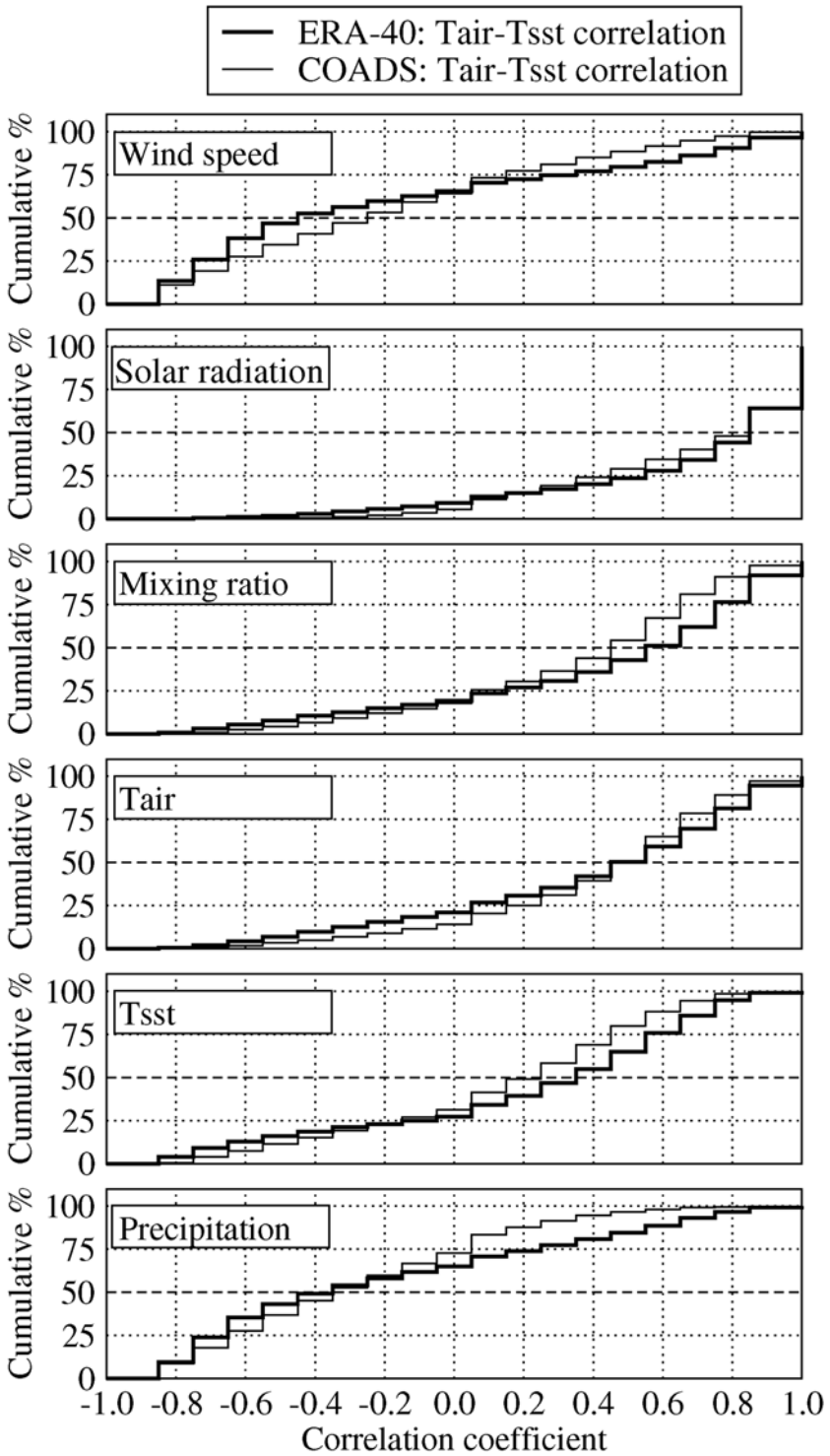

Figure 9. Cumulative percentage of correlation coefficients between Tair–Tsst and other atmospheric variables using monthly mean climatological data from ERA-40 and COADS. The median value is the point intersecting the 50% line.

the sea surface, air mixing ratio at 10 m, Tair, Tsst and precipitation). These mean values are extracted from ERA-40 and COADS at $1^\circ \times 1^\circ$ grid boxes over the global ocean. This is done for each month separately. We also combine all the mean monthly data to examine atmospheric variables which control the seasonal cycle of Tair–Tsst.

[32] The values in each 1° grid bin are just the sum of the values at grid points in the bin divided by the number of such grid points. Thus, there is no areal averaging. In other words, the assumption is that individual bins are small

Table 2. Percentages of Correlation Coefficients Shown in Figure 6[a]

Correlation Class Intervals	Product	Wind Speed	Solar Rad.	Mixing Ratio	Tair	Tsst	Precipitation
$-1.0 \le R < -0.9$	ERA–40	0.0	0.0	0.0	0.0	0.0	0.0
	COADS	0.0	0.0	0.0	0.0	0.0	0.0
$-0.9 \le R < -0.8$	ERA–40	**13.3**	0.1	0.9	0.5	3.8	9.7
	COADS	**11.1**	0.0	0.0	0.2	0.8	8.7
$-0.8 \le R < -0.7$	ERA–40	12.3	0.4	2.4	1.6	5.0	**14.5**
	COADS	8.1	0.1	0.8	0.6	3.2	9.1
$-0.7 \le R < -0.6$	ERA–40	12.4	0.6	2.2	2.2	3.8	11.3
	COADS	8.3	0.1	1.5	1.1	3.5	9.8
$-0.6 \le R < -0.5$	ERA–40	8.5	0.6	2.4	2.7	3.2	7.8
	COADS	7.0	0.2	1.9	1.5	3.9	9.0
$-0.5 \le R < -0.4$	ERA–40	5.8	1.1	2.7	2.8	2.7	6.0
	COADS	6.2	0.3	2.3	1.7	3.9	8.4
$-0.4 \le R < -0.3$	ERA–40	3.9	1.4	2.2	2.9	2.4	4.9
	COADS	6.4	0.6	2.6	2.0	3.9	7.7
$-0.3 \le R < -0.2$	ERA–40	3.3	1.5	2.1	3.0	2.0	4.1
	COADS	5.9	0.9	2.8	2.1	3.9	7.2
$-0.2 \le R < -0.1$	ERA–40	2.9	1.6	2.1	2.8	1.9	3.6
	COADS	6.0	1.4	2.9	2.3	4.0	6.7
$-0.1 \le R < 0.0$	ERA–40	2.9	1.9	2.1	2.4	2.3	3.2
	COADS	5.1	1.9	3.2	2.6	4.1	6.2
$0.0 \le R < 0.1$	ERA–40	4.9	3.8	4.5	5.9	6.8	5.8
	COADS	9.0	5.9	7.8	6.5	10.2	**10.6**
$0.1 \le R < 0.2$	ERA–40	2.2	2.0	3.3	3.9	5.2	3.1
	COADS	4.0	3.6	4.8	4.5	7.6	4.4
$0.2 \le R < 0.3$	ERA–40	2.2	2.3	3.9	4.8	7.5	3.5
	COADS	4.0	4.3	6.1	6.1	9.3	3.6
$0.3 \le R < 0.4$	ERA–40	2.3	2.8	5.1	6.5	8.2	3.4
	COADS	3.9	4.9	7.5	8.4	10.6	3.1
$0.4 \le R < 0.5$	ERA–40	2.6	3.5	6.8	8.3	10.0	3.7
	COADS	3.5	5.2	10.3	11.2	**10.8**	2.2
$0.5 \le R < 0.6$	ERA–40	3.0	4.5	8.4	8.9	**10.9**	4.2
	COADS	3.3	5.4	13.0	**14.2**	8.5	1.5
$0.6 \le R < 0.7$	ERA–40	3.6	6.3	10.9	10.3	10.2	4.4
	COADS	2.9	5.9	**13.7**	13.6	6.1	1.0
$0.7 \le R < 0.8$	ERA–40	4.4	9.8	14.2	11.9	8.9	3.7
	COADS	2.8	7.7	10.1	10.8	4.2	0.5
$0.8 \le R < 0.9$	ERA–40	5.7	20.0	**15.8**	13.2	4.2	2.5
	COADS	2.1	15.5	6.8	7.9	1.2	0.2
$0.9 \le R < 1.0$	ERA–40	3.8	**35.8**	8.0	5.4	1.0	0.6
	COADS	0.4	**36.1**	1.9	2.7	0.3	0.1

[a]Correlation coefficients are between Tair–Tsst and other atmospheric variables. The class intervals are 0.1 wide and range from −1.0 through 1.0. The highest percentage value is printed in boldface.

enough that areal weighting is not needed within the bin. However, the width (in longitude) of the bins is adjusted so that the size of the bin in m^2 is approximately constant. Thus, at the equator each 1° bin is 1° by 1°, but starting at 60° it becomes 1° in latitude by 3° in longitude. At the pole, essentially all the grid points would be one big bin (in longitude). Since we masked values (i.e., did not use them) at very high latitudes due to ice, weighting the data points by areal coverage, i.e., having larger (smaller) values at the equator (near the poles) is not a concern in this study.

[33] An example of how GUIDE proceeds is illustrated in Figure 11. It shows a regression tree obtained using data from ERA-40 in January. The tree is built on **x** and y, where y is the variable Tair–Tsst, and the predictor variables are **x** = (radflx, vapmix, wndspd, precip). The abbreviations are defined in the figure caption. In this particular example, the regression tree first partitions the results based on vapmix. Observations with vapmix $\le$0.0048 kg kg^{-1} (4.8 g kg^{-1}) go to the left branch and otherwise to the right branch. A least squares function linear in the four predictor variables is fitted to the data in each leaf node of the tree. The average value of Tair–Tsst in each leaf node is in italics (Figure 11). The fact that the GUIDE tree first splits on vapmix (i.e., vapor mixing ratio at the sea surface) implies that the latter has the largest nonlinear effect on the seasonal cycle of Tair–Tsst in January. If vapmix is $\le$4.8 g kg^{-1}, the variable with the next largest nonlinear effect is wndspd (near-surface wind speed). On the other hand, when vapmix is

Table 3. Mean and Median Correlation Coefficients Over the Global Ocean[a]

	Global Mean		Global Median	
Variable	ERA–40	COADS	ERA–40	COADS
Wind speed	−0.27	−0.26	−0.48	−0.29
Solar rad.	0.65	0.64	0.84	0.85
Mixing ratio	0.40	0.33	0.58	0.44
Tair	0.34	0.38	0.51	0.50
Tsst	0.17	0.10	0.38	0.21
Precipitation	−0.28	−0.37	−0.41	−0.38

[a]Correlation values are obtained from a least squares analysis with Tair–Tsst as the dependent variable and six predictor variables, calculated in ice–free regions over the global ocean.

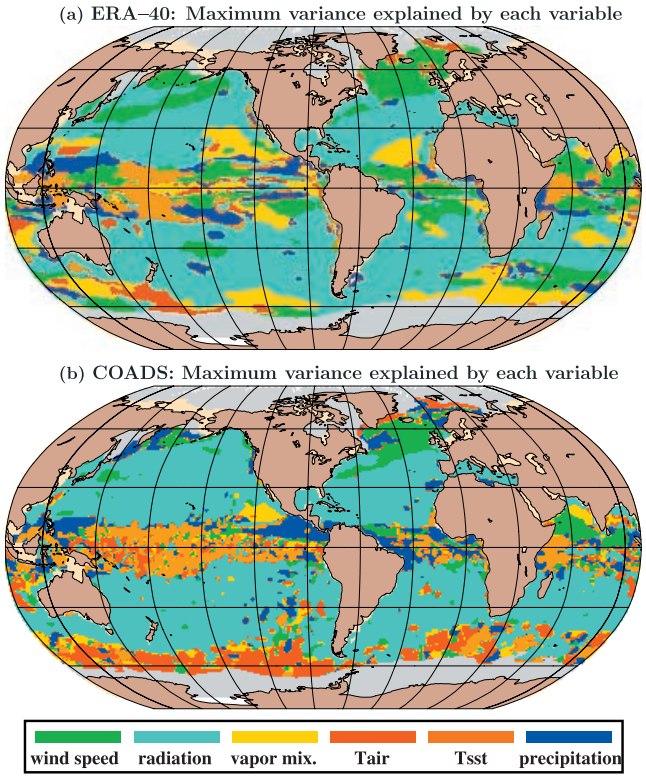

Figure 10. Regions showing which atmospheric variable explains the largest percentage of the variance in Tair–Tsst over the global ocean. They are determined from correlation values shown in Figure 6, as explained in the text. Areal percentage over the global ocean where the maximum variance is explained by each atmospheric variable (wind speed, net solar radiation, vapor mixing ratio, Tair, Tsst and precipitation) is as follows: 18.7 (9.1), 50.8 (57.0), 12.2 (3.8), 3.2 (9.3), 6.5 (10.1), 10.7% (8.6%) for the ERA-40 (COADS) data set, respectively.

>4.8 g kg^{-1}, radflx (net solar radiation) has the next largest effect.

[34] The estimation procedure in Figure 11 begins by creating an initial decision node and then adding further nodes as constrained by the tree growth parameters. Because it is always possible to obtain zero apparent prediction error by partitioning the predictor space so finely that each node contains just enough observations to fit a multiple linear model perfectly, a criterion is necessary to determine the optimal tree size. This is achieved by first constructing an overly large tree and then pruning it to maximize a cross–validated estimate of expected square prediction error. Here pruning refers to an objective tree selection procedure that finds the subtree having the best estimated predictive accuracy.

[35] The regression tree in Figure 11 is obtained after pruning. It is constructed recursively as follows. At each node of the tree, a multiple linear regression model is fitted to the data there. For each observation y, the model gives a predicted value $\hat{y}$. Define the "residual" associated with y by $y - \hat{y}$. If the model fits the data satisfactorily, a plot of the residuals versus any predictor variable should look like random noise.

[36] The tree in Figure 11 also provides empirical evidence for nonlinearity between between Tair–Tsst and atmospheric variables. The GUIDE algorithm fits a multiple linear regression to the data in each node of the tree. Therefore all predictor variables are treated equally, indicating that there is no leading variable. If there is no nonlinearity, a single multiple linear model would be sufficient. This would give a tree with a single terminal node (i.e., no splits). The fact the the tree has so many branches indicates that a multiple linear model is inadequate. Each time the algorithm detects nonlinearity, it would split the data into two subsets and try to fit a linear model separately to each subset. If GUIDE is forced to use a single predictor to fit each node (i.e., simple linear model), the tree gets bigger. Not only are the slopes different in each node (or segment), but even the selected predictors are different.

[37] Figure 12 shows the plots of residuals for each predictor variable for the observations in the top node of the tree. Vapor mixing ratio has the most significant curvature test in terms of the chi–square test. Therefore, it is selected to split the top node in the regression tree.

5.2. Importance of Atmospheric Forcing Variables

[38] The question of "which variable is the most important for determining the magnitude of Tair–Tsst" can be posed in two ways: (a) which variable has the smallest prediction error if each variable is used singly to predict Tair–Tsst?, and (b) which variable whose absence from a model using all the predictors produces the largest increase in prediction error? In (a) Tair–Tsst is considered to be a function of only one variable (e.g., wind speed, net solar radiation, etc., separately), while in (b) the dependence of all variables except one on Tair–Tsst is taken into account. In both cases, we use GUIDE to fit the necessary models and compare its cross–validation estimates of prediction mean square errors to rank the variables in order of importance with respect to their effect on Tair–Tsst. Confidence intervals for the estimated errors are used to determine the statistical significance of the ranks.

[39] We first examine the importance of each variable when it is used singly in predicting Tair–Tsst over the seasonal cycle on climatological timescales. This is done by combining all monthly mean data (from January through December) for ERA-40 and COADS, separately. The analyses for individual months are reported later.

[40] Table 4 gives the cross–validation estimates of prediction mean squared error for each variable when it is used singly to predict Tair–Tsst as mentioned in (a). Net solar radiation at the sea surface is the most important predictor for Tair–Tsst, because it yields the smallest prediction mean squared error estimate of 0.40 for ERA-40 and 0.37 for COADS. It is followed by vapor mixing ratio, Tair, Tsst, precipitation and wind speed when using the ERA-40 data set. Since the error estimates for two variables are not statistically significant if their 95% confidence intervals overlap, we conclude from the table that net solar radiation is the most important predictor and that the other variables are less important, and they are tied among themselves.

[41] To address question (b), we first fit a GUIDE model using solar radiation, vapor mixing ratio, precipitation, and wind speed as predictor variables. Then, a separate GUIDE

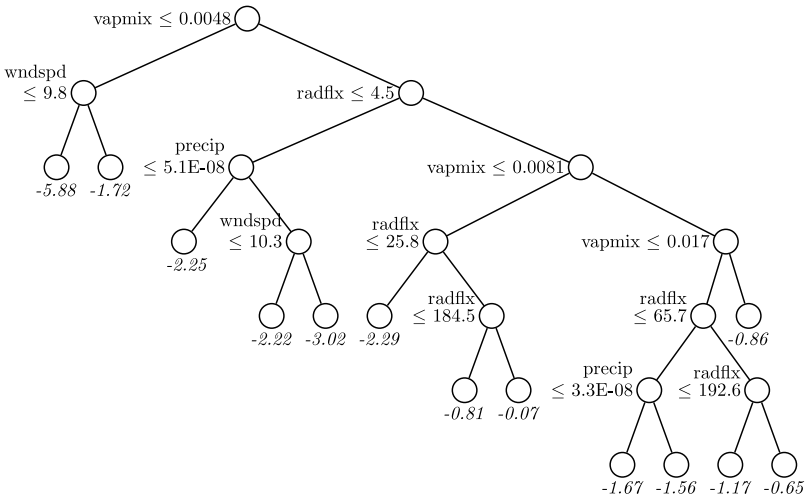

Figure 11. Piecewise linear GUIDE regression tree model for Tair−Tsst based on the January data from ERA-40. At each branch, an observation goes to the left if the stated condition is satisfied; otherwise it goes to the right. The number (in italics) beneath each leaf node is the average value of Tair−Tsst. The predictor variables are wndspd (wind speed at 10 m above the sea surface in m s^{-1}), radflx (net solar radiation at the sea surface in W m^{-2}), vapmix (vapor mixing ratio at 10 m above the sea surface in kg kg^{-1}), and precip (precipitation in m s^{-1}). Note that for this particular example, the predictor picked as most important for prediction of Tair−Tsst is the vapor mixing ratio at 10 m above the sea surface.

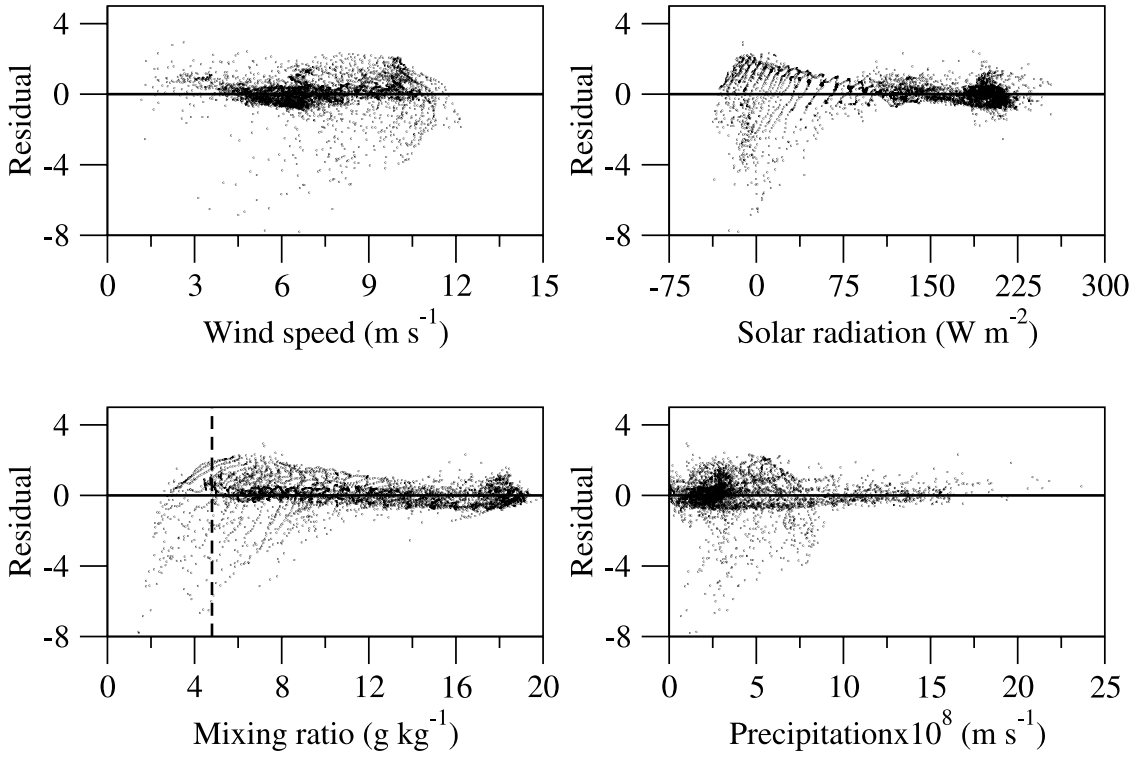

Figure 12. Plots of residuals versus each predictor variable. The residuals are obtained from a piecewise multiple linear GUIDE model shown in Figure 11. Based on the chi-squared test, vapor mixing ratio has the most significant curvature, followed by the net solar radiation. Thus, vapor mixing ratio and net mixing ratio are primary and secondary important variables for the prediction of Tair−Tsst. The other two variables (wind speed and precipitation) have relatively less curvature, thereby less important.

Table 4. Prediction of Mean Squared Error Using a Single Variable in GUIDE[a]

	Results for ERA–40			Results for COADS		
Deleted Variable	Estimated Error	Confidence Interval	Statistically Significant?	Estimated Error	Confidence Interval	Statistically Significant?
Wind speed	0.56	(0.54, 0.58)	No	0.52	(0.50, 0.54)	No
Solar radiation	0.40	(0.38, 0.42)	Yes	0.37	(0.35, 0.39)	Yes
Mixing ratio	0.53	(0.51, 0.55)	No	0.53	(0.51, 0.55)	No
Tair	0.54	(0.52, 0.56)	No	0.53	(0.51, 0.55)	No
Tsst	0.55	(0.53, 0.57)	No	0.53	(0.51, 0.55)	No
Precipitation	0.55	(0.53, 0.57)	No	0.50	(0.48, 0.52)	No

[a]Cross-validation estimates of mean squared error are obtained when each variable is used as a sole predictor of Tair–Tsst in the GUIDE models. Estimated errors are provided along with 95% confidence intervals for each predictor. The results for ERA-40 and COADS are given separately.

model is fitted for each subset of three variables that leaves out one. The reason for not using Tair and Tsst as predictors is because Tair–Tsst is a linear function of these two variables. Thus, Tair and Tsst would predict Tair–Tsst perfectly. Since the prediction error for a model based on a subset of variables is likely to be worse than that for a model based on the whole set, we can use the increase in prediction error due to variable exclusion to rank the variables in their effect on Tair–Tsst.

[42] Table 5 shows the results, where the GUIDE algorithm is used to fit the models. We see that deletion of net solar radiation produces the largest increase in estimated prediction mean squared error of 0.39 (0.38) for ERA-40 (COADS). The variable with the second largest increase is wind speed (0.35) for ERA-40 and vapor mixing ratio (0.33) for COADS. From the confidence intervals reported in the table, the difference between solar radiation and wind speed is statistically significant for ERA-40 but the difference between solar radiation and vapor mixing ratio is not statistically significant. For both data sets, there is no statistical significance in the differences between the other two remaining variables. Thus, the result for the second most important variable is inconclusive. It is safe, however, to say that the most important variable is net solar radiation, a conclusion that matches our answer for question (a).

[43] We also investigate whether or not the most important predictors which affect Tair–Tsst vary greatly by month. The GUIDE models for Tair–Tsst are first constructed using a single predictor variable for each month. Estimated prediction mean squared errors obtained using both the ERA-40 and COADS data sets reveal that, compared with the other predictors, net solar radiation at the sea surface is the most important variable which drives Tair–Tsst for all months except March, April, September, and October (Table 6). Similarly, when three of the four predictors are used in the GUIDE model, the exclusion of net solar radiation produced the largest estimated prediction mean squared error for all months except January–March, September, and October for both the ERA-40 and COADS data sets (Table 7). No other predictor variable exhibits the same consistent behavior.

[44] Finally, an examination of important atmospheric variables which mainly regulate Tair–Tsst is performed for two regions involving the major oceanic currents systems (Kuroshio and Gulf Stream) where Tair–Tsst experiences a large seasonal cycle, as discussed before (see Figure 2). Clearly, these are regions where advection by the major current systems could have a substantial impact on Tair–Tsst [e.g., *Dong and Kelly*, 2004], but an impact neglected in this study which is focussed on the response to atmospheric forcing variables.

[45] Similar to values given in Table 5, we calculate prediction mean squared error for each atmospheric variable in predicting Tair–Tsst based on the GUIDE analysis in both regions. The Kuroshio region is bounded by 25°N–40°N and 120°E–160°E, and the Gulf Stream region is bounded by 30°N—45°N and 40°W—80°W.

[46] Confidence intervals for each atmospheric variable are given in Table 8. The most important variable in predicting Tair–Tsst is still net solar radiation at the sea surface. This is true in both regions based on both ERA-40 and COADS data sets. The importance of the net solar radiation is even more significant when the relationship between Tair–Tsst and other variables is examined for the relatively small Kuroshio region in comparison to the global ocean. This is because the estimated error without net solar radiation (1.96) is more than 2 times larger than the error (0.78) for the second most important variable, mixing ratio. This is neither for the case for the Gulf Stream region (Table 8) nor for the global ocean (Table 5). We also note that precipitation seems to be the second

Table 5. Cross–Validation Estimates of Prediction Mean Squared Error[a]

	Results for ERA-40			Results for COADS		
Deleted Variable	Estimated Error	Confidence Interval	Statistically Significant?	Estimated Error	Confidence Interval	Statistically Significant?
Solar radiation	0.39	(0.37, 0.41)	Yes	0.38	(0.36, 0.40)	Yes
Wind speed	0.35	(0.33, 0.37)	Yes	0.30	(0.28, 0.32)	No
Mixing ratio	0.33	(0.31, 0.35)	No	0.33	(0.31, 0.35)	No
Precipitation	0.30	(0.28, 0.32)	No	0.32	(0.30, 0.34)	No

[a]The estimated prediction mean squared errors are obtained in each case by fitting a GUIDE model to all except the indicated variable in column 1. A variable is considered statistically significant if its confidence interval does not overlap with that of the variable with the lowest estimated error.

Table 6. Estimates of Prediction Mean Squared Error by Month Based on a Single Predictor[a]

ERA-40	Wind Speed	Solar Radiation	Mixing Ratio	Precipitation
Jan	1.62 ± 0.10	0.99 ± 0.06	1.26 ± 0.06	1.63 ± 0.10
Feb	1.16 ± 0.06	0.79 ± 0.04	0.93 ± 0.05	1.11 ± 0.06
Mar	0.46 ± 0.02	0.45 ± 0.02	0.45 ± 0.03	0.42 ± 0.02
Apr	0.25 ± 0.01	0.24 ± 0.01	0.20 ± 0.01	0.27 ± 0.01
May	0.42 ± 0.01	0.25 ± 0.01	0.39 ± 0.01	0.43 ± 0.01
Jun	0.54 ± 0.02	0.28 ± 0.01	0.57 ± 0.02	0.57 ± 0.02
Jul	0.35 ± 0.01	0.25 ± 0.01	0.38 ± 0.01	0.34 ± 0.01
Aug	0.35 ± 0.01	0.26 ± 0.01	0.33 ± 0.01	0.32 ± 0.01
Sep	0.28 ± 0.01	0.25 ± 0.01	0.25 ± 0.01	0.26 ± 0.01
Oct	0.31 ± 0.01	0.30 ± 0.01	0.29 ± 0.01	0.29 ± 0.01
Nov	0.60 ± 0.03	0.40 ± 0.02	0.58 ± 0.03	0.58 ± 0.03
Dec	1.07 ± 0.06	0.66 ± 0.04	0.93 ± 0.05	1.06 ± 0.06
COADS	Wind Speed	Solar Radiation	Mixing Ratio	Precipitation
Jan	0.90 ± 0.05	0.47 ± 0.02	0.90 ± 0.04	0.92 ± 0.04
Feb	0.69 ± 0.03	0.50 ± 0.02	0.70 ± 0.03	0.72 ± 0.03
Mar	0.33 ± 0.01	0.30 ± 0.01	0.32 ± 0.01	0.30 ± 0.01
Apr	0.24 ± 0.01	0.21 ± 0.01	0.20 ± 0.01	0.23 ± 0.01
May	0.44 ± 0.02	0.34 ± 0.03	0.39 ± 0.02	0.41 ± 0.02
Jun	0.49 ± 0.01	0.32 ± 0.01	0.49 ± 0.01	0.48 ± 0.01
Jul	0.45 ± 0.01	0.30 ± 0.01	0.46 ± 0.01	0.42 ± 0.01
Aug	0.44 ± 0.01	0.35 ± 0.01	0.42 ± 0.01	0.41 ± 0.01
Sep	0.31 ± 0.01	0.28 ± 0.01	0.30 ± 0.01	0.28 ± 0.01
Oct	0.38 ± 0.02	0.34 ± 0.01	0.34 ± 0.01	0.36 ± 0.02
Nov	0.61 ± 0.02	0.35 ± 0.01	0.60 ± 0.02	0.56 ± 0.02
Dec	0.88 ± 0.04	0.43 ± 0.02	0.89 ± 0.04	0.82 ± 0.03

[a]Cross–validation estimates of prediction means squared error, when a single predictor variable is used in the GUIDE analysis. The values are given for ERA-40 and COADS data sets separately. Standard errors (±) are also provided. Each 95% confidence interval is obtained by taking two times the standard error around the estimate.

most significant variable in controlling Tair–Tsst in the Gulf Stream, a region where heat loss through evaporation is significant.

6. Summary and Conclusions

[47] The relationship between the pairs of Tair versus Tsst and Tair–Tsst versus four independent atmospheric forcing variables (net solar radiation at the sea surface, wind speed, vapor mixing ratio at 10 m above the sea surface, and precipitation at the sea surface) is investigated. Our analysis is based on global monthly mean climatologies of these variables from two global data sets: ERA-40 and COADS.

[48] The results clearly reveal that while there is a strong correlation between Tair and Tsst over the global ocean, the relationship between the two is not as simple as can be described by a linear least squares approach. The reason is that skill is very low between Tair and Tsst due the large unconditional bias (i.e., the bias due to the large differences between mean Tair and Tsst) in some regions (e.g., equatorial regions).

[49] The atmospheric response to Tair–Tsst is generally a function of more than one atmospheric forcing variable in many regions over the global ocean on climatological timescales. Therefore, a tree–based statistical methodology that allows for nonlinear relationships between Tair–Tsst and other atmospheric variables is used to determine the most important of the variables considered in influencing Tair–Tsst. The method fits piecewise linear models to Tair–Tsst. Results using combined data (i.e., all 12 months) from ERA-40 and COADS shows that net solar radiation at the sea surface is the most important predictor for Tair–Tsst over the seasonal cycle. The same variable is also picked when a similar analysis is performed using data for each month separately. In particular, net solar radiation at the sea surface is found to be a crucial parameter in predicting Tair–Tsst for May through August, November and December. Both data sets (ERA-40 and COADS) yield almost identical results, reinforcing the importance of net solar radiation as a predictor for Tair–Tsst. They also indicate the robustness of the relationship between Tair–Tsst and other atmospheric variables. The results, as revealed by the regression tree models, point to the importance of the large–scale environment in influencing Tair–Tsst. In addition, the methodology presented in this paper shows that regression trees can be applied to data with highly non–symmetric distributions because the models do not require strong distributional assumptions.

[50] The approach using the statistically–based GUIDE algorithm should be more effective in combination with other dynamical ocean models, such as the Princeton Ocean Model (POM) and HYbrid Coordinate Ocean Model (HYCOM). In addition, all of the analyses in this paper are based on the assumption that Tair–Tsst is mainly driven by local near–surface atmospheric variables. An examination of the effects of dynamical processes, such as oceanic upwelling and advection in the atmosphere and the ocean in driving the seasonal cycle of Tair–Tsst deserves a future study. Such processes do not assume one–dimensional oceanic response to the local atmospheric forcing.

Appendix A: GUIDE Algorithm

[51] A brief description of how the GUIDE algorithm proceeds is provided here. While there are other tree–based prediction algorithms [e.g., *Breiman et al.*, 1984; *Alexander*

Table 7. Estimated Increase in Prediction Mean Squared Error From a Multipredictor Model[a]

ERA-40	Wind Speed	Solar Radiation	Mixing Ratio	Precipitation
Jan	0.46 ± 0.03	0.62 ± 0.03	0.73 ± 0.05	0.52 ± 0.03
Feb	0.39 ± 0.02	0.44 ± 0.03	0.52 ± 0.03	0.48 ± 0.04
Mar	0.24 ± 0.02	0.20 ± 0.01	0.22 ± 0.01	0.28 ± 0.01
Apr	0.13 ± 0.01	0.15 ± 0.01	0.15 ± 0.01	0.13 ± 0.01
May	0.14 ± 0.01	0.24 ± 0.01	0.16 ± 0.01	0.13 ± 0.01
Jun	0.18 ± 0.01	0.33 ± 0.02	0.18 ± 0.01	0.13 ± 0.00
Jul	0.14 ± 0.00	0.20 ± 0.01	0.15 ± 0.00	0.12 ± 0.00
Aug	0.12 ± 0.01	0.17 ± 0.01	0.16 ± 0.00	0.11 ± 0.00
Sep	0.17 ± 0.00	0.18 ± 0.00	0.19 ± 0.00	0.19 ± 0.00
Oct	0.18 ± 0.01	0.19 ± 0.01	0.20 ± 0.01	0.21 ± 0.01
Nov	0.23 ± 0.01	0.33 ± 0.02	0.26 ± 0.01	0.23 ± 0.01
Dec	0.36 ± 0.02	0.48 ± 0.03	0.48 ± 0.02	0.40 ± 0.02
COADS	Wind Speed	Solar Radiation	Mixing Ratio	Precipitation
Jan	0.33 ± 0.02	0.34 ± 0.02	0.26 ± 0.01	0.38 ± 0.02
Feb	0.30 ± 0.02	0.30 ± 0.01	0.25 ± 0.01	0.34 ± 0.02
Mar	0.16 ± 0.01	0.16 ± 0.01	0.17 ± 0.01	0.22 ± 0.01
Apr	0.14 ± 0.01	0.15 ± 0.01	0.15 ± 0.01	0.14 ± 0.01
May	0.18 ± 0.02	0.29 ± 0.02	0.27 ± 0.02	0.18 ± 0.01
Jun	0.17 ± 0.01	0.33 ± 0.01	0.21 ± 0.01	0.18 ± 0.01
Jul	0.15 ± 0.00	0.29 ± 0.01	0.20 ± 0.01	0.15 ± 0.00
Aug	0.22 ± 0.01	0.33 ± 0.01	0.27 ± 0.01	0.23 ± 0.01
Sep	0.22 ± 0.01	0.24 ± 0.01	0.23 ± 0.01	0.22 ± 0.01
Oct	0.24 ± 0.01	0.26 ± 0.01	0.29 ± 0.02	0.24 ± 0.01
Nov	0.23 ± 0.01	0.41 ± 0.02	0.26 ± 0.01	0.23 ± 0.01
Dec	0.32 ± 0.02	0.46 ± 0.02	0.29 ± 0.01	0.31 ± 0.01

[a]Cross-validation estimates of increase are shown by month, using GUIDE with all except one of the predictor variables.

Table 8. Cross–Validation Estimates of Prediction Mean Squared Error for Kuroshio and Gulf Stream Region[a]

Kuroshio Region	Results for ERA-40			Results for COADS		
Deleted Variable	Estimated Error	Confidence Interval	Statistically Significant?	Estimated Error	Confidence Interval	Statistically Significant?
Solar radiation	1.96	(1.84, 2.08)	Yes	0.91	(0.85, 0.96)	Yes
Mixing ratio	0.78	(0.73, 0.83)	Yes	0.32	(0.32, 0.33)	Yes
Wind speed	0.57	(0.53, 0.61)	Yes	0.24	(0.22, 0.25)	No
Precipitation	0.44	(0.41, 0.46)	No	0.21	(0.19, 0.22)	No
Gulf Stream Region	Results for ERA-40			Results for COADS		
Solar radiation	0.87	(0.83, 0.91)	Yes	0.60	(0.58, 0.63)	Yes
Precipitation	0.73	(0.70, 0.76)	Yes	0.49	(0.47, 0.52)	Yes
Mixing ratio	0.62	(0.59, 0.65)	No	0.50	(0.48, 0.52)	Yes
Wind speed	0.62	(0.59, 0.65)	No	0.44	(0.42, 0.46)	No

[a]The results are shown as in Table 5, i.e., estimated prediction mean squared errors for the deleted variable for each case. A variable is considered statistically significant if its confidence interval does not overlap with that of the variable with the lowest estimated error.

and Grimshaw, 1996; *Ripley*, 1996], GUIDE, as used in this paper has advantages over them because of three main reasons: (1) it has a negligible selection bias, (2) it includes categorical predictor variables, and (3) it is sensitive to pairwise interactions between regressor variables.

[52] Suppose that a random observation $(\mathbf{x}, y)$ is generated by the relation $y = f(\mathbf{x}) + \varepsilon$, where $f(\mathbf{x})$ is an unknown function, ε represents random variation with zero expectation, and $\mathbf{x}$ may be a vector of predictor variables $\mathbf{x} = (x_1, x_2, \ldots, x_k)$. In this context, f is called a regression function.

[53] Given a data set $\{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \ldots, (\mathbf{x}_n, y_n)\}$ of n observations, there are many methods of estimating f. Clearly, it is desirable to choose an estimate $\hat{f}$ such that the expected square prediction error $E\{y^* - \hat{f}(\mathbf{x}^*)\}^2$ is small, where $(\mathbf{x}^*, y^*)$ denotes a future observation that is not in the data used to construct $\hat{f}$.

[54] If f is known to be a linear function of $x_1, x_2, \ldots, x_k$, the least squares method often yields an estimate with excellent expected square prediction error. But the latter can be very large if f is not a linear function. In that case, a nonparametric method that adapts to the complexity of f is usually preferred. One such method is GUIDE which yields a piecewise linear estimate of f. Besides being completely automatic and adaptive, GUIDE has the unique feature that the data partitions defining the piecewise linear estimate can be displayed graphically as a binary decision tree.

[55] The algorithm aims to identify the predictor variable whose plot exhibits the highest degree of non–randomness, because this variable is most likely to have a nonlinear effect on the dependent variable. GUIDE employs the chi–square test [e.g., *Jaisingh and Rozakis*, 2000] to measure the degree of nonlinearity in each predictor variable. Specifically, it first groups the predictor values into four groups at the sample quartiles and then cross–tabulates the grouped values with the signs of the residuals. The predictor variable yielding the most significant chi–square statistic is selected to split the node with an inequality of the form $x \leq c$. Each value of c divides the data into two subsets and a multiple linear regression model is fitted to the data in each subset. The value of c that yields the smallest total sum of squared residuals in these two regression models is selected as the best split of the node. Complete details on the curvature test and the pruning algorithm are given in *Loh* [2002].

[56] **Acknowledgments.** Special thanks go to A. J. Wallcraft and E. J. Metzger of NRL for their valuable discussions and help in processing the atmospheric forcing data. Much appreciation is extended to the reviewers for their constructive comments. This work is funded by the Office of Naval Research (ONR) under the 6.1 project, Global Remote Littoral Forcing via Deep Water Pathways. The work of W.–Y. Loh is partially supported by the National Science Foundation and the U.S. Army Research Office. This paper is contribution NRL/JA/7304/05/6066 and has been approved for public release.

References

Alexander, W. P., and S. D. Grimshaw (1996), Treed regression, *J. Comp. Graph. Stat.*, *5*, 156–175.

Barron, C. N., A. B. Kara, H. E. Hurlburt, C. Rowley, and L. F. Smedstad (2004), Sea surface height predictions from the Global Navy Coastal Ocean Model (NCOM) during 1998–2001, *J. Atmos. Oceanic Technol.*, *21*, 1876–1894.

Breiman, L., J. H. Friedman, R. A. Olshen, and C. J. Stone (1984), *Classification and Regression Trees*, 358 pp., Wadsworth, Stamford, Conn.

Cayan, D. R. (1992), Latent and sensible heat flux anomalies over the northern oceans: Driving the sea surface temperature, *J. Phys. Oceanogr.*, *22*, 859–881.

da Silva, A. M., C. C. Young, and S. Levitus (1994), *Atlas of Surface Marine Data*, vol. 1, *Algorithms and Procedures*, *NOAA Atlas NESDIS 6*, 83 pp., Natl. Oceanic and Atmos. Admin., Silver Spring, Md.

Dong, S., and K. A. Kelly (2004), Heat budget in the Gulf Stream region: The importance of heat advection and storage, *J. Phys. Oceanogr.*, *34*, 1214–1231.

Fairall, C. W., E. F. Bradley, J. E. Hare, A. A. Grachev, and J. B. Edson (2003), Bulk parameterization of air-sea fluxes: Updates and verification for the COARE algorithm, *J. Clim.*, *16*, 571–591.

Frankignoul, C. (1985), Sea surface anomalies, planetary waves, and air-sea feedback in the middle latitudes, *Rev. Geophys.*, *23*, 357–390.

Haidvogel, D. B., and F. O. Bryan (1992), Ocean general circulation modeling, in *Climate System Modeling*, edited by K. E. Trenberth, pp. 371–412, Cambridge Univ. Press, New York.

Jaisingh, L. R., and L. Rozakis (2000), *Statistics for the Utterly Confused*, 318 pp., McGraw-Hill, New York.

Kållberg, P., A. Simmons, S. Uppala, and M. Fuentes (2004), *ERA-40 Proj. Rep. Ser.*, 17, 31 pp.

Kara, A. B., H. E. Hurlburt, P. A. Rochford, and J. J. O'Brien (2004), The impact of water turbidity on the interannual sea surface temperature simulations in a layered global ocean model, *J. Phys. Oceanogr.*, *34*, 345–359.

Kara, A. B., H. E. Hurlburt, and A. J. Wallcraft (2005), Stability-dependent exchange coefficients for air-sea fluxes, *J. Atmos. Oceanic Technol.*, *22*, 1077–1091.

Kraus, E. B., and J. A. Businger (1994), *Atmosphere-Ocean Interaction*, 2nd ed., 362 pp., Oxford Univ. Press, New York.

Loh, W.-Y. (2002), Regression trees with unbiased variable selection and interaction detection, *Stat. Sinica*, *12*, 361–386.

Murtugudde, R., J. Beauchamp, C. R. McClain, M. R. Lewis, and A. Busalacchi (2002), Effects of penetrative radiation on the upper tropical ocean circulation, *J. Clim.*, *15*, 470–486.

Qu, T., Y. Y. Kim, M. Yaremchuk, T. Tozuka, A. Ishida, and T. Yamagata (2004), Can Luzon Strait transport play a role in conveying the impact of ENSO to the South China Sea?, *J. Clim.*, *17*, 3644–3657.

Reynolds, R. W., N. A. Rayner, T. M. Smith, and D. C. Stokes (2002), An improved in-situ and satellite SST analysis for climate, *J. Clim.*, *15*, 1609–1625.

Ripley, B. D. (1996), *Pattern Recognition and Neural Networks*, 415 pp., Cambridge Univ. Press, New York.

Send, U., R. C. Beardsley, and C. D. Winant (1987), Relaxation from upwelling in the Coastal Ocean Dynamics Experiment, *J. Geophys. Res.*, *92*, 1683–1698.

Soloviev, A. V., and R. Lukas (1997), Large diurnal warming events in the near-surface layer of the western equatorial Pacific warm pool, *Deep Sea Res.*, *44*, 1055–1076.

Soloviev, A. V., R. Lukas, and P. Hacker (2001), An approach to parameterization of the oceanic turbulent boundary layer in the western Pacific warm pool, *J. Geophys. Res.*, *106*, 4421–4435.

Wilks, D. S. (1995), *Statistical Methods in the Atmospheric Sciences*, 467 pp., Elsevier, New York.

Yasuda, I., T. Tozuka, M. Noto, and S. Kouketsu (2000), Heat balance and regime shifts of the mixed layer in the Kuroshio Extension, *Prog. Oceanogr.*, *47*, 257–278.

Yu, L., R. A. Weller, and B. Sun (2004), Improving latent and sensible heat flux estimates for the Atlantic Ocean (1988–1999) by a synthesis approach, *J. Clim.*, *17*, 373–393.

H. E. Hurlburt and A. B. Kara, Oceanography Division, Naval Research Laboratory, Code 7320, Bldg. 1009, Stennis Space Center, MS 39529, USA. (birol.kara@nrlssc.navy.mil)

W.-Y. Loh, Department of Statistics, University of Wisconsin, Madison, WI 53706, USA.