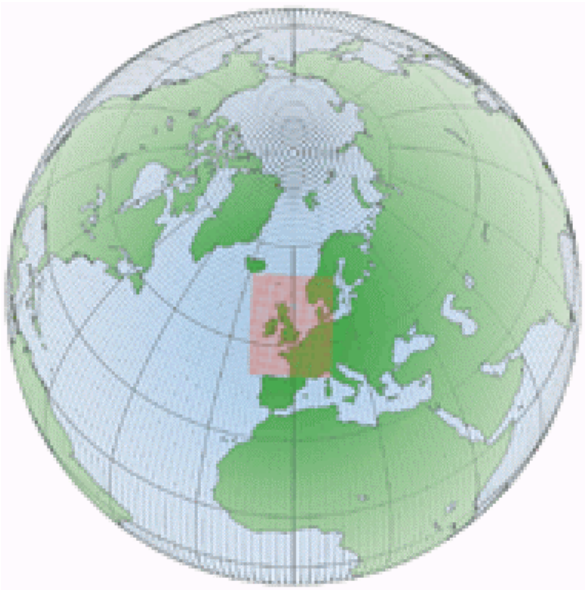
Numerical Weather Prediction

An error budget for bulk sea-surface temperatures derived from AATSR data



Forecasting Research Technical Report No. 411

Lisa A. Horrocks, Roger W. Saunders and Anne O'Carroll

email: nwp_publications@metoffice.com

AN ERROR BUDGET FOR BULK SEA-SURFACE TEMPERATURES DERIVED FROM AATSR DATA

Lisa A. Horrocks, Roger W. Saunders, Anne O'Carroll

Met Office, London Road, Bracknell, RG12 2SZ, UK, Email: lisa.horrocks@metoffice.com

Document History

Date	Version	Action/Comments	Approval
14/04/2003	1.0	Report provided as deliverable 4a/3/02 from Technical Annex 4a of CPP.	R. Saunders
03/06/2003	2.0	Draft of NWP Technical Report incorporating comments on Version 1.0 from J. Eyre.	
12/06/2003	2.1	Final version, with minor corrections to Version 2.0, following comments from J. Eyre.	J. Eyre

Executive summary

Sea-surface temperature measurements to high accuracy are required in many areas of climate research, not least for climate change detection. The Along-Track Scanning Radiometers have been designed to provide SSTs accurate to better than 0.3 K. By the end of the ENVISAT mission (2007) a 15-year record of data from these instruments will be available.

Estimates of the bias and random error associated with bulk SSTs derived from (A)ATSR are needed in order to assimilate these data into climate analyses or to interpret intercomparisons with alternative datasets.

A model for the error budget associated with (A)ATSR bulk SSTs has been described, and estimates of the expected bias and random error have been determined. An additional quality control has been suggested in order to screen out contaminated or anomalous data.

Optimal bulk SSTs from (A)ATSR are expected to be unbiased to ±0.1 K level of accuracy. Random error is dependent on the choice of retrieval algorithm and the atmospheric conditions pertaining to the observation, but should be below 0.25 K in all cases.

Work is underway to define two simple corrections to apply to retrieved skin SSTs in order to attain the expected level of bias. Further work to provide a quantitative validation of these estimates is recommended.

This work should enable widespread use of (A)ATSR SSTs in climate research in general and should contribute to improvements in the Hadley Centre datasets in particular.

1. INTRODUCTION

The primary purpose of the Advanced Along-Track Scanning Radiometer (AATSR), launched on ESA's ENVISAT platform in March 2002, is to continue the record of climate accuracy global SST established by ATSR-1 and -2. The ATSR instruments have been designed to deliver SST to better than 0.3 K accuracy, as required for climate datasets. We derive a bulk SST product from (A)ATSR data suitable for comparison with in situ SST datasets and analyses. For meaningful comparisons and, ultimately, blending of the (A)ATSR-derived SSTs with in situ datasets, it is essential that the errors associated with the (A)ATSR SSTs are accurately represented, both in terms of systematic errors (or biases) and random error.

Here we develop a logical model for the errors contributing to the (A)ATSR bulk SST product. We use published results or our own validation to establish the magnitudes of the different error components, attempting as far as possible to keep biases and random errors separate. A general estimate of error associated with the derivation of a bulk SST from (A)ATSR is determined, and we then consider the conditions which may increase the error associated with individual observations. We offer some validation of our error model through comparisons with HadISST1. Finally, we suggest a pragmatic quality check against a background SST field, which may be applied during operational data processing to ensure that spurious data are identified.

2. DATA PROCESSING AND ERROR MODEL

The (A)ATSR is sensitive to the ocean's infrared radiative skin temperature (skin SST), which is the temperature of only the top few 10^{-5} m. This temperature is different from the bulk SST sampled by in situ sensors at depths of 1 m or more because of the skin effect (e.g., Fairall et al. 1996) and any shallow temperature gradients which build up during the day. The skin-bulk temperature differences can be between ~0.2 and 0.6 K (skin cooler) at night or up to 1–3 K (skin warmer) during the day. Traditional SST datasets (e.g., Parker et al. 1995, Rayner et al. 2003) have been constructed from bulk SST measurements. (A)ATSR SSTs must be adjusted for skin-bulk differences to obtain comparability with these datasets because at the level of accuracy required for climate research (0.3 K, Allen et al. 1994) these differences are not negligible.

2.1 Processing scheme

(A)ATSR skin SSTs are retrieved from spatially-averaged infrared channel brightness temperatures using a linear retrieval defined from high resolution radiative transfer modelling. (A)ATSR has infrared channels at 3.7, 11 and 12 microns and it views the surface at nadir and forward along the orbital ground track. Skin SST can therefore be retrieved using a number of combinations of different channels and views. SSTs retrieved using the dual-view capability of the instrument are considerably better than those using only nadir view data, and so we discuss only the dual-view retrievals in this paper. We consider two algorithms. The first makes use of only the 11 and 12 micron channel data, and is referred to as D2. The second uses in addition data at 3.7 microns, and is referred to as D3. This algorithm cannot be used during the day when reflected solar radiation contributes to the radiance seen at 3.7 microns. In either case, skin SST (T_S) is retrieved by linear combination of the channel brightness temperatures (T_i), weighted by predetermined coefficients (a_i) (Eq. 1). The coefficients and offset (a_o) are defined by multiple regression of skin temperatures against brightness temperatures computed for a global suite of atmospheric profiles via a detailed radiative transfer model (Merchant et al. 1999).

$$T_S = a_o + \sum_{i=1,n} a_i T_i \qquad (1)$$

We then predict a skin-bulk temperature adjustment (ΔT) for each observation via a physical parameterisation for the skin effect (Fairall et al 1996). Under forced convection, this skin effect is dependent on both the net heat flux out of the ocean (Q) and the instantaneous wind speed through Eq 2

$$\Delta T = \frac{\lambda Q \nu}{k u_*}, \qquad (2)$$

in which u_* is the friction velocity in the water (in turn related to wind speed), k and ν are the thermal conductivity and kinematic viscosity, respectively, and λ is a dimensionless parameter. Fairall et al. made Eq 2 appropriate also to free convection, by redefining λ to be buoyancy-dependent: it attains a constant value at moderate to high wind speeds, and reduces as wind speeds reduce. Horrocks et al (2003a) provide a fuller description and show how this parameterisation is entirely adequate for application to ATSR-2 SSTs. Bulk SST (T_B) may then be derived by combination of skin SST and skin-bulk adjustment (Eq 3).

$$T_B = T_S + \Delta T \qquad (3)$$

A physical model to predict incidences of daytime shallow surface warming (e.g., Kantha and Clayson 1994) is used to identify potentially inflated SSTs.

2.2 Proposed error model

The bias and random error associated with the final estimate of bulk SST are the end result of errors incurred at each stage of the SST calculation. Figure 1 provides a schematic representation of each of these steps and the sources of error at each stage.

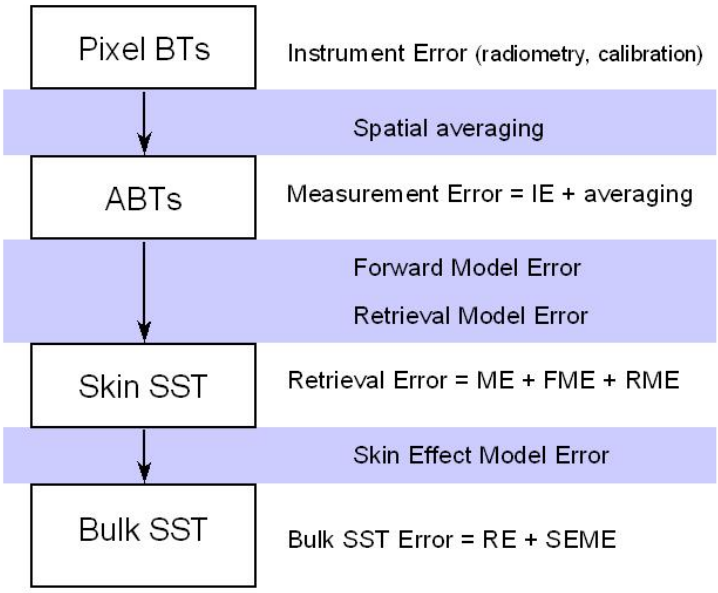


Figure 1. Bulk SST processing scheme and contributions to errors. Each named error can be split into systematic and random components. Not shown here are any additional errors arising from various sources of contamination or instrument/processing anomalies.

The retrieval error is the error associated with an estimate of skin SST from uncontaminated 10-arcminute averaged brightness temperature (ABT) data. It represents the combined effects of instrument, forward model and retrieval model errors. Retrieval error depends upon the combination of infrared channels used in the retrieval, and can vary with latitude. Retrieval error has been quantified for ATSR (e.g., Barton et al.1989, Merchant 1998), and there is no reason to expect AATSR to be much different.

The skin effect model error represents uncertainties associated with the prediction of skin-bulk adjustment. It combines errors in both the parameterisation itself and the flux data used to force the parameterisation. The total observation error associated with an estimate of bulk SST can then be defined as the combination of retrieval error and skin effect model error.

Additional unforeseen and non-gaussian errors can affect individual observations. Errors of this type result from various sources of data contamination (undetected cloud, undetected sea-ice, undetected daytime shallow warming, undetected land), unflagged instrument anomalies, or unexpected processing anomalies. Data affected by this second type of error should be identified and removed prior to further processing or assimilation. Auxiliary information (such as background SST fields, cloud climatologies, ice coverage analyses) are required for this quality control process.

In section 3 we assess contributions to the bulk SST observation error. In section 4, we consider other sources of data contamination. In section 5 we discuss our observation error estimates and suggest a quality check to identify data affected by contamination or other anomalies.

3. ESTIMATES OF BULK SST OBSERVATION ERROR

3.1 Skin SST retrieval error

Merchant (1998) studied the simulated retrieval error associated with skin SSTs determined from ATSR spatially averaged brightness temperatures (ABTs). Noise on the ABTs was assumed at the 0.01 K level. This is a conservative estimate computed by assuming ~0.05 K noise (typical NEΔT) on the 1-km pixel brightness temperatures and ~50% of the pixels contributing to one ABT. Simulated retrieval errors had mean (bias) and standard deviation (random error) of 0.01 K and 0.19 K for D2, and -0.002 K and 0.05 K for D3. An alternative approach from Barton et al (1989) who found similar values for the theoretical random errors on ATSR retrievals of 0.23 K for D2 and 0.08 K for D3. Barton et al. used an earlier version of the Radiative Transfer Model (RTM), whereas Merchant updated the spectroscopic data and aerosol assumptions used in the RTM, and this may account for the small differences between the two studies. Errors for AATSR skin SST retrievals might be expected to be smaller than those for ATSR owing to NEΔT at levels of only 0.025–0.037 K on AATSR infrared channels (Smith et al 2001). However, in orbit the AATSR channel noise has increased due to contamination of radiometer optics from residual water vapour held in the ENVISAT satellite structure (C. Mutlow, pers. com.).

The random part of the retrieval error (i.e. standard deviation) for ABTs has two sources: fitting error and noise. The fitting error is seen when linear retrieval coefficients are reapplied to the brightness temperature dataset from which they were derived, and has a global value of ~0.10 K for D2. The noise depends upon the actual NEΔT on each channel measurement, and also the number of pixels averaged to make up the ABT, as summarised in Eq. (4):

$$\sigma_{\text{noise}} = \sqrt{\sum \left(a_i \frac{\text{NE}\Delta\text{T}}{\sqrt{n_{\text{pixels}}}} \right)^2} \qquad (4)$$

where σ_{noise} is the overall noise component, a_i are the retrieval coefficients, n_{pixels} is the number of pixels averaged to form the ABT, and NEΔT is ~0.05 K. Strictly, the noise, and hence retrieval error, therefore varies with the number of pixels included in each ABT. Given an overall random error of 0.19 K for D2, and fitting error of ~0.10 K, the noise component is expected to be ~0.16 K, on average. We assume an average value everywhere, since the majority of the AATSR observations passed through processing have more than 50 % of the pixels contributing to each ABT. At high latitudes where the number of available pixels is significantly reduced, we introduce an inflated noise component to compensate (see below).

The higher random error for D2, compared to D3, results mainly from the larger fitting error when only two channels are used in the retrieval instead of three. It is illustrated by comparison of scatter plots of retrieved *vs* true SST for D2 and D3 retrievals. For example, Merchant and Harris (1999) provide scatter plots of retrieved SST against buoy SST for a suite of ATSR–buoy coincidences in the tropics (their Figure 4), and scatter in the D2 case is wider, regardless of choice of retrieval coefficients. A <0.2 K random error on D2 indicates the limit of accuracy achievable by this linear algorithm on (A)ATSR spatially averaged data.

The 0.01 K global mean systematic retrieval error for D2 masks a small zonally-varying component reaching -0.1 K in the tropics and +0.1 K in mid-latitudes. This variable component results from non-linearities in the relationship between SST and brightness temperature differences. Brightness temperature differences are strongly controlled by total column atmospheric water vapour, which in turn is correlated with latitude; so we see an apparently latitude-dependent variability in systematic retrieval error in simulations. It is also evident in global comparisons of D2 and D3 retrievals of real night time AATSR data (Figure 2) because the equivalent non-linearities for D3 are minimal. The zonally-varying component of D2 bias could be removed

by application of a systematic adjustment to D2 retrievals, based on latitude, or air-mass. Merchant (1998) noted that D2 random error was little reduced by using latitude-dependent retrieval coefficients, although these achieved significant reduction in bias. We therefore assume that the application of an adjustment based on latitude, while effective in removing zonal SST biases, could produce only a marginal reduction (<0.01 K) in global random error. A latitude-dependent correction was developed for a CD-ROM of ATSR-2 data (Murray et al. 2000): a similar correction could be defined for both ATSR and AATSR. Work is well underway towards formulating a correction for systematic errors in AATSR retrievals, and since we intend to apply this correction, we hereafter ignore the zonal variation in D2 bias and assume values of 0.01 K and 0.19 K for D2 retrieval error mean and SD. In the absence of this correction, a zonally-varying contribution should strictly be included in estimates of D2 bias.

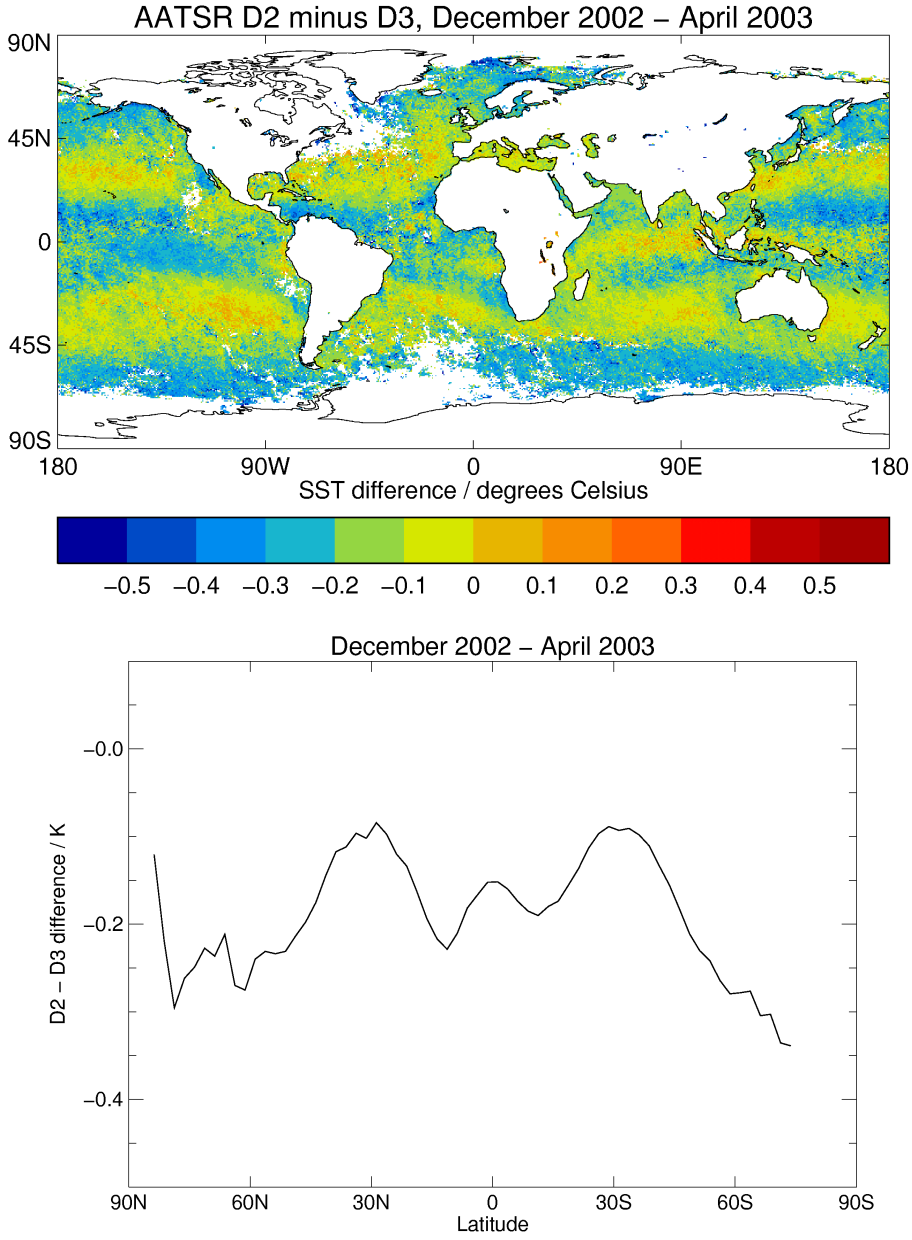


Figure 2. Mean differences between D2 and D3 for 4 months of night time AATSR data, displayed on a map at half-degree resolution (upper) and as zonal averages in 1-degree bands (lower).

The absolute bias on skin SST retrievals cannot be defined without a validation exercise. Merchant and Harris (1999) did this for ATSR. Horrocks et al (2003b) provided a similar validation of early AATSR data. In

both cases the absolute bias associated with D2 was negligible, whereas D3 was biased warm by ~0.2 K. A bias of this magnitude can be expected theoretically given the uncertainty associated with brightness temperatures specified by existing radiative transfer models (Merchant and Le Borgne, 2003). From Eq. (1), for unbiased retrievals, the offset, a_o, is defined as:

$$a_o = T_S - \sum_{i=1,n} a_i T_i \qquad (5)$$

Biases, e, in simulations of brightness temperatures cause a bias, e_{ao}, in the derived value of a_o weighted by the a_i coefficients:

$$e_{a_o} \simeq e\sqrt{\sum a_i^2} \qquad (6)$$

The a_i coefficients generally have larger absolute values for D2 retrievals (between 1 and 10) than D3 (between 0.1 and 4). For a_o to be specified with an accuracy of ~0.05 K for D2 retrievals, absolute errors in simulations of brightness temperatures would need to be <0.01 K. Current radiative transfer models may be accurate to ~0.05 K at best, which means that errors in the value of a_o could be up to ~0.2 K for D3 and ~0.4 K for D2. (Note that because a_i coefficients depend on differences in simulated brightness temperatures, retrieval error standard deviations of less than 0.2 K can be obtained from simulations accurate to ~0.05 K.)

Since the absolute accuracy of retrievals of skin SST is limited by the accuracy of radiative transfer modelling, the biases seen in (A)ATSR validation exercises are unsurprising. The obvious way to remove these biases is to adjust the a_o term by some empirically-determined amount. We intend to define a_o adjustments for AATSR on the basis of quality-controlled AATSR–buoy matchups, and to apply it routinely; we therefore assume hereafter that bias on both D2 and D3 skin SSTs can be maintained below ~0.05 K.

In summary, assuming correction of zonally-varying bias in D2 and appropriate empirically-determined offset adjustments for both D2 and D3, absolute bias on skin SSTs should be smaller than ±0.05 K with random error of ~0.19 and ~0.05 K for D2 and D3, respectively. In the absence of these corrections, absolute biases could be between ±0.1 K on D2, varying with latitude, and ~0.2 K on D3.

SST retrieval in high-latitude zones is liable to larger errors than retrievals in lower latitudes when a global set of coefficients is employed. This is due to the extreme atmospheric conditions that prevail in these regions. Matthiesen and Merchant (2003) set out a detailed study of SST retrieval from ATSR-2 image resolution data in marginal ice zones. They note that a standard SST retrieval can result in biases of up to 0.6 K, but that a retrieval tailored to ice zones can reduce this bias to ~0.1 K. The standard deviation of the retrieval error from use of the standard D2 retrieval coefficients on single pixel data in these regions was 0.15 K (Matthiesen and Merchant, Table 6). This number provides a measure of the fitting error part of the random error for retrievals of ABTs (since for retrievals of image resolution data no extra noise was included); this is larger than the mean global fitting error of ~0.10 K for D2.

Since the fitting error is larger for retrievals in high latitudes, the overall random error can be recalculated to include the expected increase. However, the noise component should also be increased because at high latitudes only about half the number of 1 km^2 pixels are available to contribute to each 10 arcminute cell compared to mid-latitudes/tropics (due to the curvature of the Earth). From Eq. (4), the overall noise should thus be increased by a factor of ~√2, from the average of ~0.16 K suggested earlier to ~0.23 K. (The assumption of similar NEΔT, even at the highest latitudes is justified, since calibration of the infrared channels on AATSR showed no significant increase in error until brightness temperatures dropped below ~250 K, Smith et al., 2001.) The overall random error on D2 ABT retrievals in high latitudes then becomes ~0.27 K, compared to ~0.19 K in the general case.

Matthiesen and Merchant also showed that standard D2 retrievals at high latitudes had a mean bias of ~–0.16 K. We assume that this is just the high-latitude end of the zonal bias in D2 retrievals identified above,

and would be removed by application of a latitude-dependent correction. For D3 SST retrievals there was no significant increase in fitting error or bias in high latitude zones over the global case.

Matthiesen and Merchant calculated the range of SST that could be close to sea-ice (their figure 6): this delimits the category of "extreme conditions" for which standard retrievals may be deficient. At 0% ice coverage, SST ranged from -0.5 to 8 °C, based on a standard linear dependence of SST on ice fraction (Rayner et al. 2003) combined with random noise. We choose ~8 °C (281 K) to be the threshold of SST below which the random error associated with skin retrieval would be increased to 0.22 K for D2. The processor is already configured to flag immediately any retrieved skin SSTs below 273.15 K, with bulk SSTs derived only for unflagged observations. Retrieval errors are summarised in Table 1.

Table 1. Estimated retrieval errors for (A)ATSR data, in K.

	Dual-view, 2 channels		Dual-view, 3 channels
	Generally	SST < 281 K	Globally
Retrieval Maximum Bias	±0.05	±0.05	±0.05
Retrieval Random Error	0.19	0.27	0.05

Note: Errors assume that adjustments described in text have been applied.

3.2 Skin Effect Model Error

As outlined in section 2, we estimate the cool skin effect by means of a physical parameterisation in order to obtain comparability between (A)ATSR-derived SST and in situ bulk SST. Merchant and Harris (1999) also used the cool skin model of Fairall et al (1996) to compare ATSR SSTs with buoy SSTs in a tropical validation exercise.

Following arguments presented by Merchant and Harris, systematic errors in estimates of heat flux and friction velocity may be ~15 % from typical flux parameterisation schemes; from Eq. (2) these propagate to a systematic error in predicted skin effect of ~20 % (i.e., ~0.04 K for a typical skin effect of 0.2 K).

In operational (A)ATSR processing, we use flux data from the Met Office NWP model analysis. Errors in NWP surface flux fields are not well-constrained, but a mean global systematic error of ~30 W m^{-2} in net flux is often quoted (M. McCulloch, pers. com., following Isemer et al., 1989). This value is of order 15 %, but is likely to have strong geographical dependencies. NWP model wind fields may be better constrained than the surface fluxes. The ~20 % systematic error in skin effect suggested by Merchant and Harris (1999) should be an appropriate, and perhaps conservative, estimate of the skin effect model bias. Systematic errors in the flux data could be in either direction and variability in these errors is not accurately known. We assume that ±0.04 K delimits the usual bias in skin effect prediction caused by systematic errors in flux data.

Random errors in the flux data used by Merchant and Harris were estimated to be around 25 %, propagating to a ~35 % random error in predicted skin effect (i.e., ~0.07 K for a typical skin effect of 0.2 K). Again we assume that random errors in the NWP surface flux fields used in (A)ATSR processing will be similar to or less than this estimate

To confirm these estimates, we reconsider the skin effect observations and Fairall model predictions presented by Horrocks et al (2003a). Observations were from the Nauru99 cruise in the west Pacific, between latitudes 0–30°. For 1210 night time observations of skin effect, the Fairall model was used to predict the skin effect, first from flux data measured or derived in situ, and then with NWP model analysis flux fields; results are shown in Figure 3. When the in situ flux data were used, the mean difference between predicted and observed bulk–skin difference was -0.004 K with standard deviation of 0.034 K. Since the mean measured skin effect was 0.216 K, these values represent a typical bias and random error of 2 % and 16 %. When the NWP flux data were used, the mean difference between predicted and observed skin effects was 0.035 K with standard deviation of 0.050 K. Thus an average difference of 0.035 – (-0.004) ≈ 0.04 K in skin effect prediction was introduced when NWP flux data were used in place of in situ derivations. This discrepancy may arise from errors in both the NWP fields and the in situ derivations, but it illustrates that uncertainty in flux data can cause systematic errors in skin effect prediction at the level suggested by Merchant and Harris.

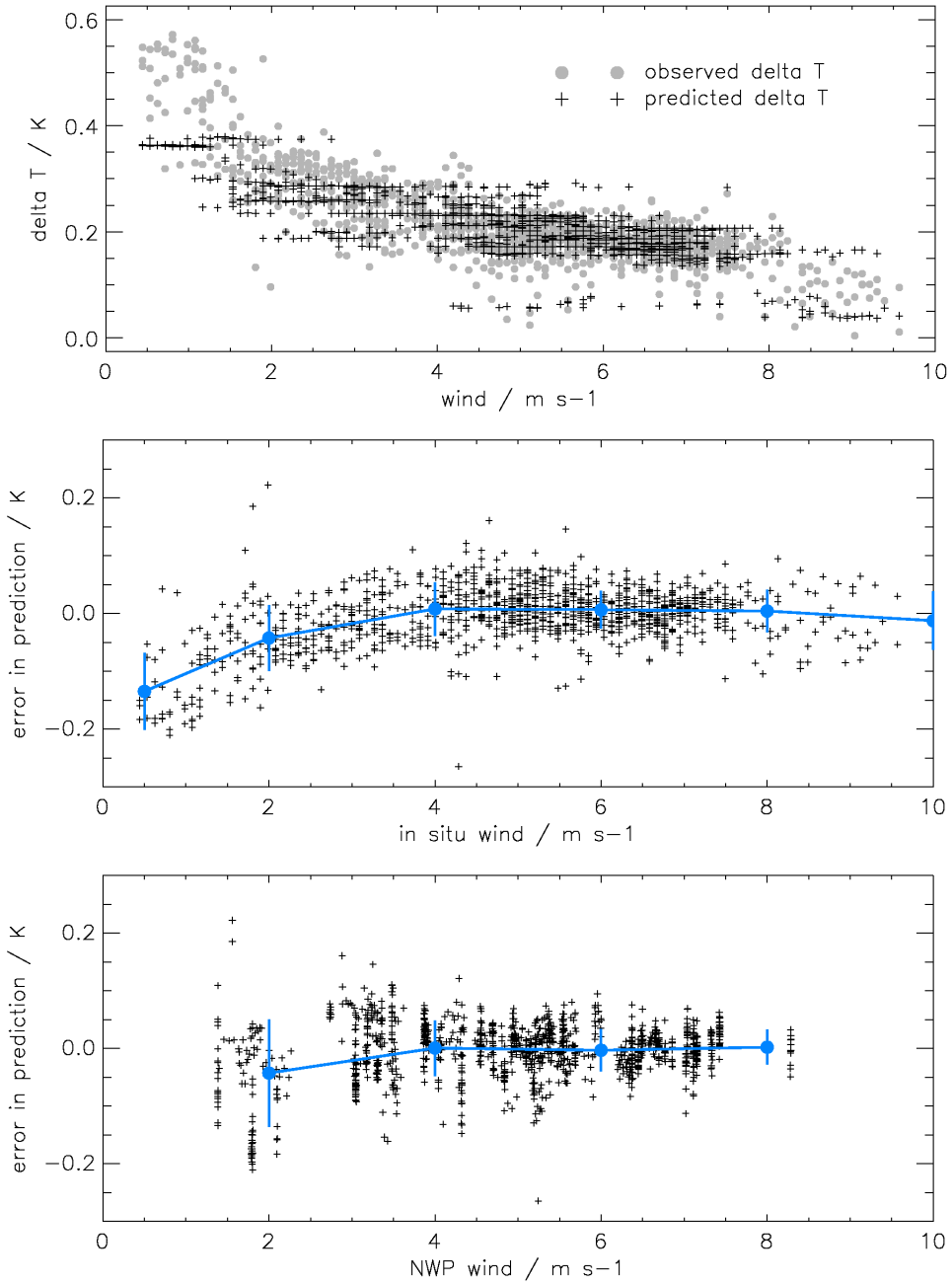


Figure 3. Observations of cool skin effect from the Nauru99 cruise, and predictions using the Fairall et al (1996) model forced by NWP analysis flux data. Upper: Bulk–skin differences against in situ wind speed. Middle: Predicted–observed skin effect error against in situ wind speed, blue line shows mean and standard deviations of data binned according to windspeed. Lower: As middle panel, but against NWP wind speed.

The bias with the in situ fluxes is low, but this is because the value of λ in Eq. 2 was set to 4.6 in order to minimise the mean bias in this dataset. In order to reduce the mean bias when NWP fluxes were used, Horrocks et al applied a systematic correction of -0.04 K to skin effect predictions from the NWP flux data. The mean bias with NWP fluxes then becomes -0.005 K, but this masks a significant trend with true wind speed (Figure 3, middle panel). For wind speeds above ~3.5 m s^{-1} (i.e., where the skin effect is roughly constant at 0.2 K), the mean bias is negligible, but at lower wind speeds, the predicted skin effect is biased low by ~0.15 K from a true skin effect of ~0.6 K. From these data, it would be possible to define a wind-dependent component to the skin effect model bias. However, for (A)ATSR processing, "true" wind speeds are not available. When the predicted–observed skin effect error is plotted against NWP wind speed

estimates, the trend is far less convincing (Figure 3, lower panel). We therefore propose that the best estimate of systematic error is still the value suggested by Merchant and Harris, but we recommend that where NWP wind speed is below ~4 m s^{-1}, the estimate of random error should be inflated. Clearly, systematic errors could be much larger if the value of λ were different. It is essential that exercises such as that by Horrocks et al. are repeated to ensure this value is tuned for minimum biases across a global dataset.

The standard deviation of the error when NWP flux data were used was 0.05 K, or ~23 %. This is less than the 0.07 K random error suggested by Merchant and Harris, but the data used in this study are temporally and spatially limited, and do not represent very high wind speeds. To ensure that our error estimate is appropriate for data under a wider range of conditions, we prefer the more conservative random error value of 0.07 K. Where the NWP wind speed is below 4 m s^{-1}, this value should be increased to 0.1 K (from the standard deviation of the lowest wind speed bin in the lower panel of Figure 3) to account for the deficiencies in skin effect prediction at low wind speed discussed above.

Gleckler and Weare (1996), investigated biases and random errors in surface flux fields derived from in situ data. For shortwave heat fluxes, they found that in the tropics, random errors of ~25 W m^{-2} were much more significant than biases. Further north, a bias of 10 W m^{-2} was dominant. For latent heat fluxes, they found that biases dominated random errors everywhere, with values of ~30 W m^{-2} in mid-latitudes and 10 W m^{-2} near the equator. It is clear that biases and random errors associated with in situ flux fields are complex and variable. Errors in model analysis fields are likely to be similarly complex. For estimates of error on operational (A)ATSR products, this level of complexity cannot be adequately represented. Instead, values of ±0.04 K for the range of systematic error, and 0.07 K (increasing to 0.1 K if NWP wind speed is less than 4 m s^{-1}) for random error associated with skin effect model predictions are conservative estimates which should be appropriate in all but the most extreme situations. These estimates are summarised in Table 2.

Table 2. Estimated skin effect model errors, in K.

	Wind > 4 m s^{-1}	Wind < 4 m s^{-1}
Skin Effect Model Maximum Bias	±0.04	±0.04
Skin Effect Model Random Error	0.07	0.10

3.3 Total bulk SST observation error

We estimate the maximum bias and random error on derived bulk SSTs from the errors associated with the skin retrieval and the skin effect model outlined above. Bulk SST is computed very simply, as Eq. (3). Therefore the maximum bias associated with bulk SST is the sum of the maximum biases associated with retrieved skin SST and predicted skin effect, and the random error is approximately the result of summing random errors in quadrature. Estimated values are summarised in Table 3.

Table 3. Estimated (A)ATSR bulk SST observation errors, in K.

	Wind > 4 m s^{-1}			Wind < 4 m s^{-1}		
	Dual-view, 2 channels		Dual-view, 3 channels	Dual-view, 2 channels		Dual-view, 3 channels
	SST ≥ 281 K	SST < 281 K	Globally	SST ≥ 281 K	SST < 281 K	Globally
Retrieval Maximum Bias	±0.05	±0.05	±0.05	±0.05	±0.05	±0.05
Skin Effect Model Maximum Bias	±0.04	±0.04	±0.04	±0.04	±0.04	±0.04
Bulk SST Maximum Bias	±0.09	±0.09	±0.09	±0.09	±0.09	±0.09
Retrieval Random Error	0.19	0.27	0.05	0.19	0.27	0.05
Skin Effect Model Random Error	0.07	0.07	0.07	0.10	0.10	0.10
Bulk SST Random Error	0.20	0.28	0.09	0.21	0.29	0.11

The bias expected in bulk SST estimates should thus fall within the bounds of ±0.09 K: we expect estimates of bulk SST to be unbiased at the 0.1 K accuracy level. This assumes that absolute biases in skin SST retrievals have been minimised by application of the adjustments described in section 3.1. It further assumes that data are uncontaminated and not affected by instrument or processing anomalies of any kind. The random error determined here for bulk SSTs is in all cases within the error specified in the design of the (A)ATSR instruments for retrieval of skin SST (0.3 K), and provides confidence that data from this series of satellite instruments should meet requirements for climate research.

4. SOURCES OF POTENTIAL CONTAMINATION

The estimates of bulk SST observation error presented in section 3 assumed that the (A)ATSR data were uncontaminated. There are three sources of potential contamination of derived bulk SSTs. Residual undetected cloud at the pixel level results in some invalid brightness temperatures contributing to the average brightness temperature values for each 10-arcminute cell. Skin SSTs retrieved from cloud-contaminated brightness temperatures will be in error, and this will subsequently affect the estimate of bulk SST. The presence of undetected sea-ice in (A)ATSR cells can also change brightness temperatures from the values they would have had were the cell completely clear sea, and retrieved skin SSTs will consequently be in error. Finally, under conditions of strong insolation and low wind, thermal stratification of the surface ocean layers invalidates the assumption of an isothermal profile between "subskin" and "bulk" measurements. In this situation, the skin layer can become significantly warmer than water at depths of 1 m or more, and the value derived for T_B from Eq. (3) could be even a few K warmer than the true bulk SST at 1-m. These scenarios are considered below, and the errors that they may introduce into the bulk SST are estimated.

4.1 Cloud contamination

Residual undetected cloud in (A)ATSR data is probably the most significant source of error in SST fields determined from this sensor. Cloudy pixels are detected during the higher level (L1b) (A)ATSR processing, via a series of tests on the infrared channel brightness temperatures, and on accompanying visible channels when they are available (during the day). The cloud detection scheme was described by Závody et al (2000). Jones et al (1996) investigated the impact of undetected cloud on night time spatially-averaged SSTs, but this study was prior to the introduction of further tests by Závody et al, able to detect much of the previously unflagged homogeneous cloud. Nevertheless, Jones et al provide a useful upper limit to the error that cloud contamination may now introduce into retrieved SSTs. They noted that the impact of cloud on half-degree spatially averaged SSTs was generally seen as a cool bias of > 1 K, and standard deviations increased by > 1 K. Similar values arose from comparisons of 10-arcminute spatially-averaged ATSR SSTs with buoys by Merchant (1998), who attributed outliers at the ~1 K level to residual cloud contamination. Undetected cloud in (A)ATSR cells may not always result in cool SSTs. Whether cloud results in depression or elevation of retrieved SST depends upon the cloud's spectral signature coupled with the sign of the retrieval coefficients for the different channels and views.

The main challenge is how to identify those (A)ATSR cells which were passed as cloud-free, but which may still contain cloud. One possibility is to consider the number of 1 km^2 pixels which contribute to the 10 arcminute cell mean. Pixels which fail the cloud detection tests in L1b processing are discarded. If the number remaining is a only a small percentage of the maximum that could contribute to the cell, then there is a high chance that these few remaining pixels may also contain cloud, as yet unidentified. If this is the case, then we expect to find the largest (A)ATSR SST anomalies (from comparisons to independent measures of SST) where the percentage of contributing pixels is small. The maximum number of 1 km^2 pixels, N_{max}, in a 10 arcminute cell at latitude, ϕ, varies from ~345 at the equator as:

$$N_{max}^{\phi} = 345 \times \cos\phi \qquad (7)$$

In Figure 4 we show the monthly mean difference between ATSR-2 bulk SST and HadISST1, a monthly climate analysis of SST based on in situ data and bias-corrected AVHRR (Rayner et al. 2003). Largest differences are located either close to strong gradients in SST (e.g., the Gulf Stream) where the ATSR data are more likely than the interpolated analysis to capture the true location of high SSTs, or close to the data gaps where no cloud-free ATSR observations were available throughout the month. Note that the large gaps in the Atlantic and East Pacific are due to poor confidence flags against the received ATSR-2 products along these orbits: the reason why these data have been flagged is unclear and is under investigation with the data provider (RAL). The lower panel in Fig. 4 shows the number of 1 km^2 pixels used in the nadir view 10 arcminute average, as the percentage of the maximum number available (calculated from Eq. 7), and averaged across the month. A minimum 15% coverage in cells is required for the data to be considered

valid, and so cells with more than 85 % cloud cover are omitted from this plot and indeed any further SST processing.

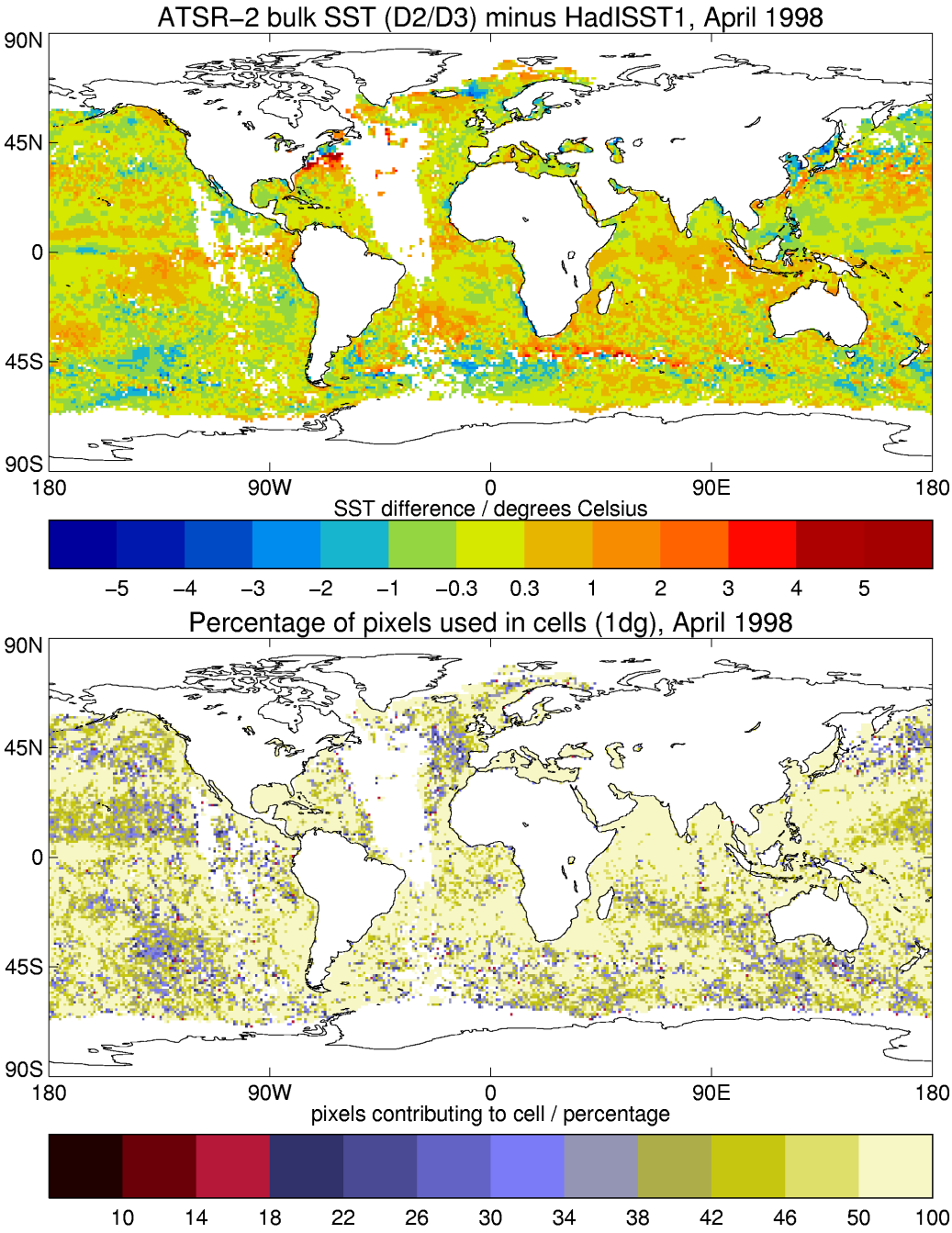


Figure 4. (upper) Differences between monthly mean ATSR-2 bulk SST and HadISST1 for April 1998. (lower) Number of 1 km^2 pixels in the nadir view 10 arcminute cell, as the percentage of the maximum number available, and averaged across the month.

Anomalies potentially related to residual cloud are in the north-east Pacific where gaps in coverage indicate high probability of cloud, and in the Southern Oceans, south of Africa and stretching eastwards, where large discrepancies are associated with a thin band of discarded data. The correspondence between these anomalies and pixel percentages is not clear: pixel percentages in parts of the north-east Pacific are low, but these do not match exactly the location of the largest SST differences. In the Southern Oceans, the pattern of low pixel percentages is less convincing. Pixel percentages are relatively low over the western North

Pacific, but SST anomalies are also low here. Areas often associated with persistent cloud cover (e.g. western seaboard of South America and Namibia) have relatively high pixel percentages, showing that ATSR cells were identified as either completely clear or completely cloudy (and thus discarded).

A similar comparison can be made between (A)ATSR bulk SSTs and collocated buoy SSTs to investigate whether largest SST anomalies correspond to low fractions of available 1 km^2 pixels. This comparison is free from the effects of monthly averaging, and so trends should be more obvious. Figure 5 shows the ATSR-2 – buoy SST difference for 3062 matchups from a global database for January–August 1998, plotted against the percentage of contributing pixels. The clear cutoff at 15 % pixels reflects the minimum coverage required for SST processing. The outliers for which the ATSR-2 SST is cooler than the corresponding buoy SST could be the result of residual cloud contamination. In fact, the warm outliers could also be the consequence of cloud, or they may represent shallow diurnal warming. Outliers in Figure 5 are distributed evenly throughout the range of pixel percentages. If the number of pixels used in the forward view is considered, the distribution is similar. Some authors suggest comparing the number of pixels available in the two views: where there are large discrepancies, this may be related to cloud contamination. A plot of SST difference against the nadir–forward pixel percentage difference revealed no significant trend, with the largest SST differences occurring where nadir and forward views had similar pixel coverage.

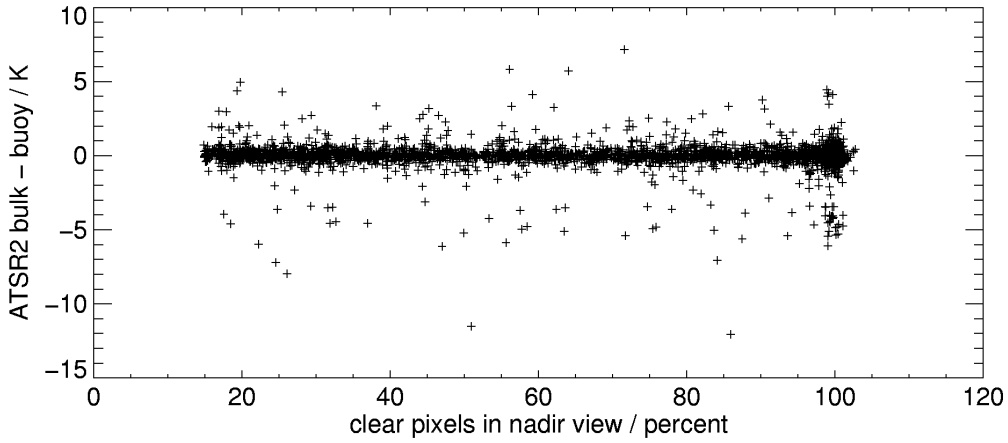


Figure 5. ATSR-2 bulk SST minus buoy SST as a function of the percentage of pixels used in the nadir view.

In summary, for pixel coverages above 15 %, we find no significant relationship between the number of pixels contributing to the 10 arcminute cells and cloud-related SST biases. Within the (A)ATSR processing scheme, an initial check comparing nadir-only and dual-view SSTs is made; this screens out observations where residual cloud present in only one view causes a large dual–nadir SST difference. However, we still suspect some observations to be affected by undetected cloud, causing the outliers in the buoy comparison, and the large SST differences near cloudy areas in the HadISST comparison. Similarly, in previous comparisons of ATSR-2 to Hadley Centre in situ datasets we attributed increased bias and standard deviation in the N Pacific during northern hemisphere summer to undetected cloud (Horrocks et al. 2001).

The challenge of how to identify the relevant observations remains. Since the percentage of pixels used in each cell cannot be related confidently to observed anomalies, there seems to be no information within the (A)ATSR data alone with which to identify cloud-contamination. Instead, it will be necessary to use comparisons with independent datasets. A check against a 30-year SST climatology is not precise enough because the magnitude of discrepancies caused by cloud is similar to discrepancies expected as a result of real climate variability (e.g., El Niño). For near-real time purposes, SST analyses from NWP could be used to identify spurious (A)ATSR data, although the accuracy of NWP SST fields is probably worse than (A)ATSR. In the future, with atmospheric models able to resolve cloud, it may be possible to use cloud analyses to assist in locating observations likely to be beneath cloud.

A more effective comparison may be against SST retrieved from microwave instruments. Since clouds are transparent to some microwave frequencies, it is possible to retrieve SST beneath cloud. The disadvantage of microwave observations is that they have a coarser resolution than infrared (minimum pixel size of ~50 km). A relatively coarse resolution global field from microwave data, perhaps updated over a rolling 5-day interval, should be adequate to identify the (A)ATSR SST outliers at the ~2 K level resulting from residual cloud. We have compared ATSR-2 data with SSTs from the Tropical Rainfall Monitoring Mission (TRMM) microwave imager (TMI) as a first step in this direction (Saunders et al. 2003a). TRMM has a low earth orbit, and data are therefore only available for latitudes between ± 40 °. Future microwave instruments (e.g. AMSR) will provide global SST coverage, with data available in near real time. However for the earlier data, from ATSR-1/-2, progress in improving pixel level cloud detection is still required and we have reported on how this may be achieved (Saunders et al. 2003b).

4.2 Sea-ice contamination

If undetected sea-ice is present in an (A)ATSR cell, it affects the average brightness temperatures. Ideally, sea-ice should be identified at the 1 km pixel level in a similar way to cloud, and only those pixels deemed ice-free should contribute to the mean brightness temperatures and SST in the cell. Ice coverage over more than 50% of the cell area should perhaps result in rejection of the cell for SST purposes. Unfortunately, the pixel level (A)ATSR processing currently includes no method of sea-ice detection: unless ice-covered pixels trigger some of the cloud detection tests (which they are quite likely to do), they will be included in the spatially-averaged products (A. Birks, pers. com.). Clearly this is an area where further work could be important.

There is no information within the (A)ATSR products received at the Met Office that allows an easy assessment of residual sea-ice in each 10-arcminute cell. However, we can use estimates of ice fraction from the NWP model analysis to identify observations corresponding to high model ice fractions. The NWP ice fraction field is built up from passive microwave data. Wherever the NWP model ice fraction is non-zero, surface flux estimates will be strongly controlled by the presence of ice. Our skin effect model has been developed for open water situations, as are the bulk formulae employed in the NWP model for air-sea fluxes. Both of these parameterisations may not hold in areas close to sea-ice where alternative processes can dominate. For that reason, NWP model flux data are only considered valid in the (A)ATSR processing system for observations where the NWP ice fraction is zero. One consequence of this is that bulk SSTs are only derived where the NWP ice fraction is zero. Observations with the highest likelihood of residual ice contamination are therefore automatically screened out of the bulk SST product.

4.3 Impact of shallow diurnal warming

A conversion from skin SST to nominal bulk SST at 1-m relies upon an accurate characterisation of the temperature gradient in this top metre of the ocean. At night, we assume an isothermal profile from beneath the skin layer down to several metres: heat is invariably being lost from the ocean into the atmosphere, encouraging convective mixing even in the absence of forced mixing due to stronger winds. Thus the only temperature gradient is that across the skin layer, which can be well-quantified by a physical parameterisation (e.g., Horrocks et al. 2003a). Profile 1 in Figure 6 illustrates this situation. During the day, under cloud-free conditions, strong insolation is absorbed at the ocean's surface. In the absence of sufficient mixing energy from moderate winds, thermal stratification develops, temperature gradients become stable, and the temperature differential over a few metres can increase through the day. Profile 2 in Figure 6 shows the impact of this diurnal warming.

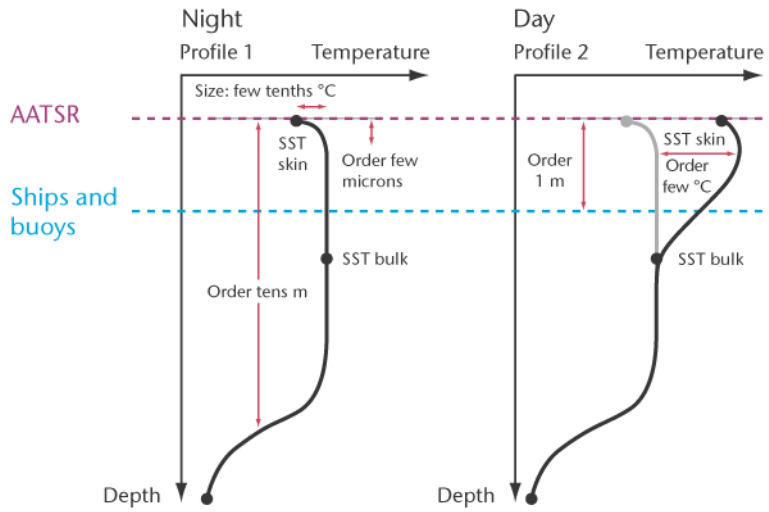


Figure 6. Schematic temperature profiles at the ocean surface for night, or cloudy, windy conditions (Profile 1) and calm clear daytime (Profile 2).

If a strong temperature gradient between skin and bulk SST exists at the time of (A)ATSR observation, then our derived value of bulk SST is warmer than the true 1-m SST by an amount determined by that gradient. Differences of up to 2–3 K between skin and 1-m SST are not uncommon in this situation. Undetected diurnal thermoclines can be a significant source of error in daytime calculations of bulk SST from (A)ATSR.

The development of the SST profile during the day depends upon net heat fluxes, including insolation, and surface wind speed, and the history of these variables from local dawn to the time of observation. Several one-dimensional models appropriate to the problem of the development of a diurnal thermocline exist (e.g., Kraus and Turner 1967, Price et al. 1986, Kantha and Clayson 1994). We have integrated two of these into the (A)ATSR processing system, with the aim of identifying (A)ATSR observations which may be affected by the presence of a diurnal thermocline. In both simulations and comparisons with observed diurnal warming in AVHRR SSTs the model suggested by Kantha and Clayson (1994) provided more realistic estimates of diurnal SST differences. A report on the application of models for diurnal warming to AATSR processing is currently being prepared.

As with the skin effect model, surface heat fluxes and wind speeds from NWP are used to force the diurnal thermocline model, which is integrated from dawn through to the (A)ATSR overpass time for each observation. A prediction of the surface to 1-m temperature difference (neglecting skin effect) at the observation time is obtained; if this value is high enough to offset a typical skin effect (i.e., > 0.2 K) it is flagged. This procedure could now be revised so that it is observations with predicted thermocline SST differences greater than the possible bulk SST bias (±0.09 K) which are flagged.

To assess whether the diurnal thermocline model is robust in the mean, we can compare a monthly mean of predicted incidences of diurnal warming with the difference between monthly mean day time and monthly mean night time AATSR SSTs. This comparison is shown in Figure 7. The day–night SST difference must be evaluated in the mean sense because there is insufficient overlap of day and night coverage from AATSR on a daily basis. Since the diurnal thermocline prediction is the monthly mean of daily computations, interpretation of Figure 7 is difficult. The day–night difference also includes the effect of different cloud detection at day and night. Cloud detection is more conservative at night because there are no visible data; areas where the daytime SST is significantly cooler than at night may be due to remnant cloud in the daytime data. The day–night difference is unaffected by interalgorithm biases because the D2 retrieval was used throughout. Coverage for the thermocline predictions is different because these were tied to daytime data only.

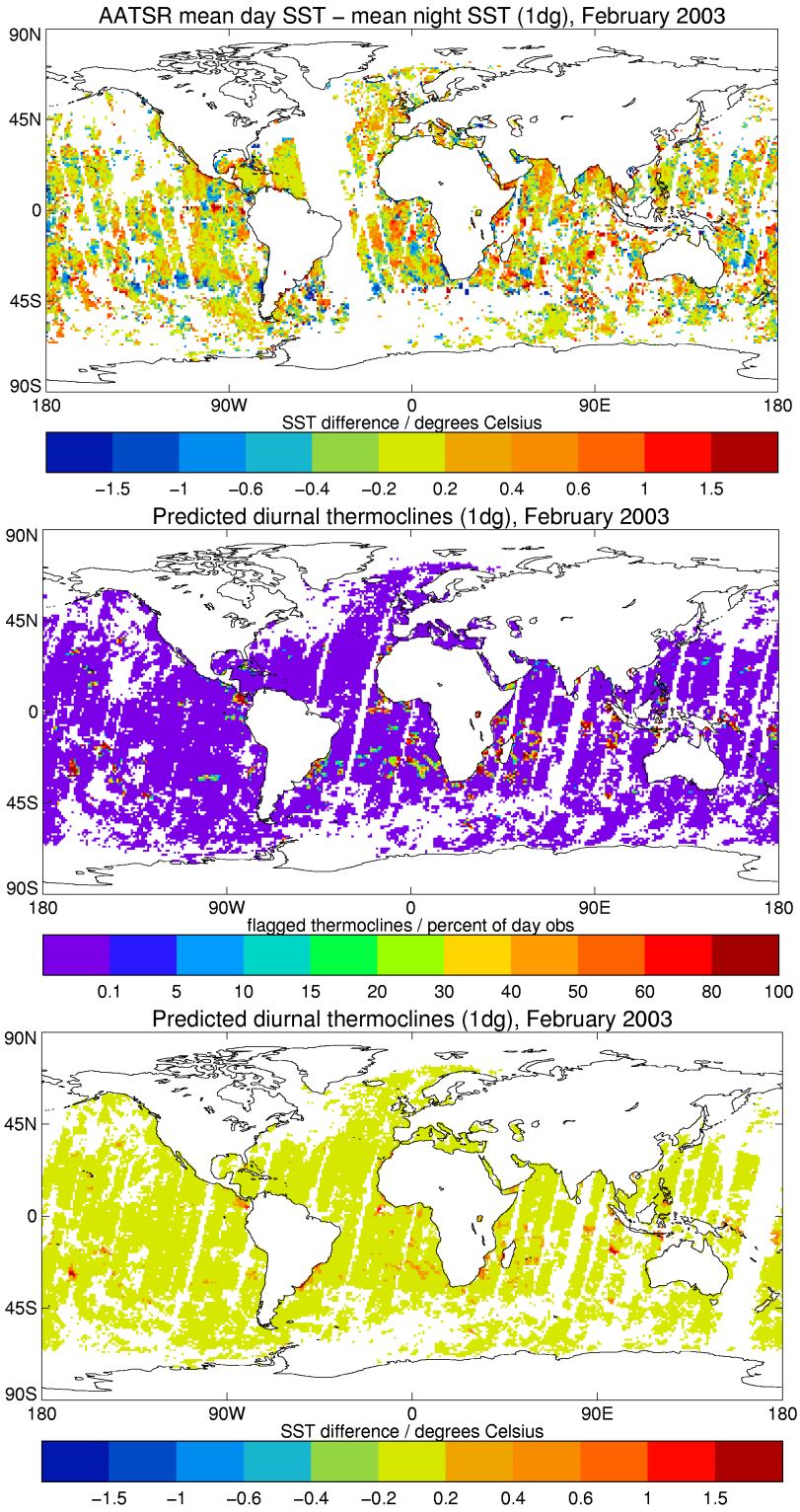


Figure 7. Comparison of diurnal thermocline predictions with observed differences between monthly mean daytime and night time skin SST fields. (Upper) Observed AATSR differences. (Middle) Predicted thermoclines (where skin–1-m difference more than 0.2 K) as a percentage of daytime observations. (Lower) Monthly mean predicted skin–1-m difference.

The predicted thermoclines are located in the tropics or mid-latitudes away from strong wind belts, as expected for the Southern hemisphere summer. The strongest thermoclines are predicted where the flux data reflect conditions of no cloud and very low wind. The observed day–night comparison is more difficult to interpret, with no areas of strong diurnal warming particularly obvious.

An alternative validity check can be made against a monthly mean of day–night SST differences obtained from AVHRR data. The wider swath of this instrument allows for direct comparison of consecutive day and night SSTs, providing a more reliable measure of diurnal warming. The comparison for February 2003 was not available; Figure 8 shows instead the global distribution of mean day–night difference for January 1989 (A. Stuart-Menteth, pers. com.). The daytime AVHRR overpass is mid-afternoon, corresponding to the peak of the diurnal cycle, compared to ~1000–1030h for (A)ATSR. Figure 8 shows the day–night temperature difference, while the lower panel in Figure 7 shows the mean skin to 1-m temperature difference: although related, these values need not be the same. Nevertheless, the predicted diurnal warming corresponds to some extent with the mean pattern of observed diurnal warming in Figure 8. Our diurnal thermocline model shows some skill in identifying potentially affected observations, at least in the mean.

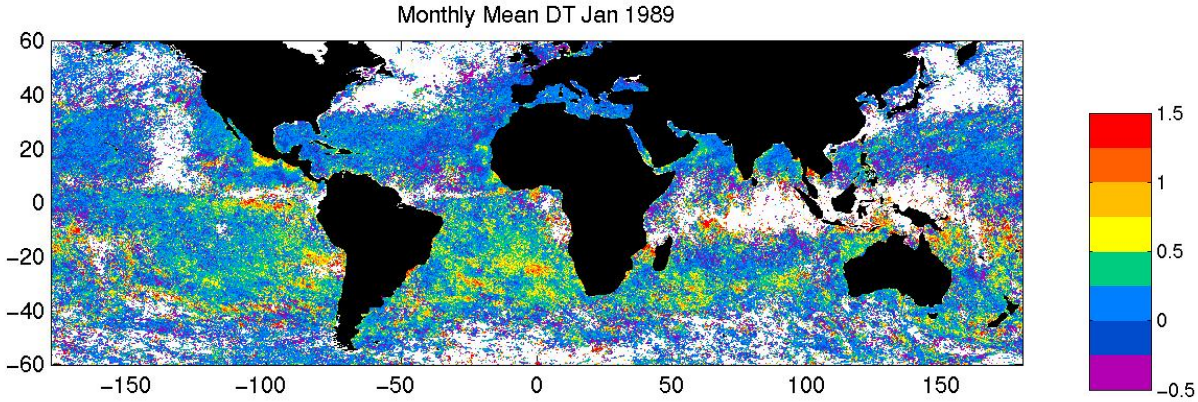


Figure 8. Observed day–night SST differences from AVHRR data during January 1989. AVHRR has a wider swath than (A)ATSR and differences between consecutive day and night SSTs can be computed. The afternoon overpass in mid-afternoon is close to the peak of the diurnal cycle. *(Figure courtesy of A. Stuart-Menteth, SOC)*.

Given that the diurnal thermocline model can identify potentially affected observations, there are two options available to the user. The more cautious approach is to reject any observations for which the model predicts a skin–1-m temperature difference of more than 0.2 K. Alternatively, affected data could be used, but with inflated error and bias to reflect the associated uncertainty. At present, we recommend the former approach, but for completeness in this report we describe a method for the latter.

The diurnal thermocline model predictions can be used to define inflated errors for affected observations. The absolute bias associated with modelled skin to 1-m temperature differences is not known, although the values commonly obtained from the model are generally of the right magnitude (up to ~3 K). Of greater significance in terms of uncertainty is the reliability of the NWP flux data over the period relevant to each observation: if the NWP model has failed to predict cloud in the correct place, then solar fluxes can be very different from the true value at a given location. In section 3.2, we estimated systematic and random errors in surface fluxes of the order of 15 % and 20 %, respectively; systematic error in the solar flux may occasionally be much larger than this. Since the performance of the model is dependent upon the quality of the flux data, which is highly variable, propagation of these values through to their impact on predicted ΔT is not sufficient. In fact, because we do not use the predicted ΔT in an additive way in the estimate of bulk SST, the exact error on the predicted value is not required. Instead, we can use the predicted ΔT as an estimate of bias on the bulk SST, while increasing the random error to account for the large variability in confidence in the flux data used to determine this estimate. What limit can we give to the random error?

Consider a population of diurnal thermocline events. If the model successfully predicts the event in 50 % of the cases, but fails in the remainder due to inaccurate flux data, half the time with an error equal to the maximum ΔT in the positive sense, and half the time in the negative sense, then the variance of the error is given by:

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}\left(x_i - \bar{x}\right)^2$$
$$\sigma^2 = \frac{1}{N}\left(\frac{N}{2}(0)^2 + \frac{N}{4}x_m^2 + \frac{N}{4}\left(-x_m\right)^2\right) \qquad (8)$$
$$\sigma^2 = \frac{x_m^2}{2}$$

where N is the number of events, x_i are the errors in the predictions, x is the mean error (in this case zero), and x_m is the maximum thermocline ΔT. For observations at mid-morning, excursions of more than ~1.5 K are rare. Taking this value, the variance in our example would be ~2.25 K, and the standard deviation little over 1 K.

It would thus be possible to use the skin–1m SST difference predicted from our implementation of the Kantha-Clayson model to estimate an additive component of the bias on affected bulk SSTs. An additional component to the random error of ~1 K should be included for these observations. This proposal would provide an increment to the error for potential diurnal warming, but only for those observations where the thermocline model predicts an effect. It cannot provide additional error for observations for which the model has failed to predict any diurnal warming. It should also be noted that if the model used to flag diurnal warming is changed, then this method of error estimation may no longer be appropriate.

5. DISCUSSION

5.1 Validation of error estimates

The new error model has been applied to AATSR data for February 2003. Code has been written in PV-WAVE to demonstrate the implementation of the scheme, and this can be easily translated to Fortran90 for incorporation into the Met Office (A)ATSR Processing Scheme (Horrocks et al. 2002). Biases and random errors have been computed separately for both D2 and D3 retrievals. The calculations are summarised in Eq. (9) and (10), in which ε represents bias and σ random error (or standard deviation) and the subscripts B, S, sem, and d, stand for bulk SST, skin SST retrieval, skin effect model and predicted diurnal thermocline, respectively.

$$\varepsilon_B = \varepsilon_S + \varepsilon_{sem} \quad \left(+\varepsilon_d\right) \qquad (9)$$

$$\sigma_B = \sqrt{\sigma_S^2 + \sigma_{sem}^2 \quad \left(+\sigma_d^2\right)} \qquad (10)$$

σ_S depends upon the type of retrieval and for D2 retrievals it is larger for SSTs below 8°C (section 3.1). σ_{sem} is larger at low wind speeds (section 3.2). The diurnal thermocline components are relevant only for daytime, low wind speed observations for which diurnal thermoclines have been predicted (section 4.3). Currently we advise that data with predicted diurnal thermoclines are discarded and in this case, these errors are not relevant. We include them here for demonstration purposes. Results from the error computation are shown in Figure 9 as a monthly mean at half-degree resolution. For comparison the difference between a bulk SST (composite of D2 retrievals during the day and D3 at night) and HadISST1 is also shown.

The first observation from Figure 9 is that the strongest anomalies in the AATSR–HadISST difference plot are not represented in the modelled error fields. Significant differences between AATSR and HadISST can arise from the different resolutions of the datasets, sampling inconsistencies in the AATSR data, and error in the analysis, as well as from contamination of AATSR data. These effects are not included in the error model.

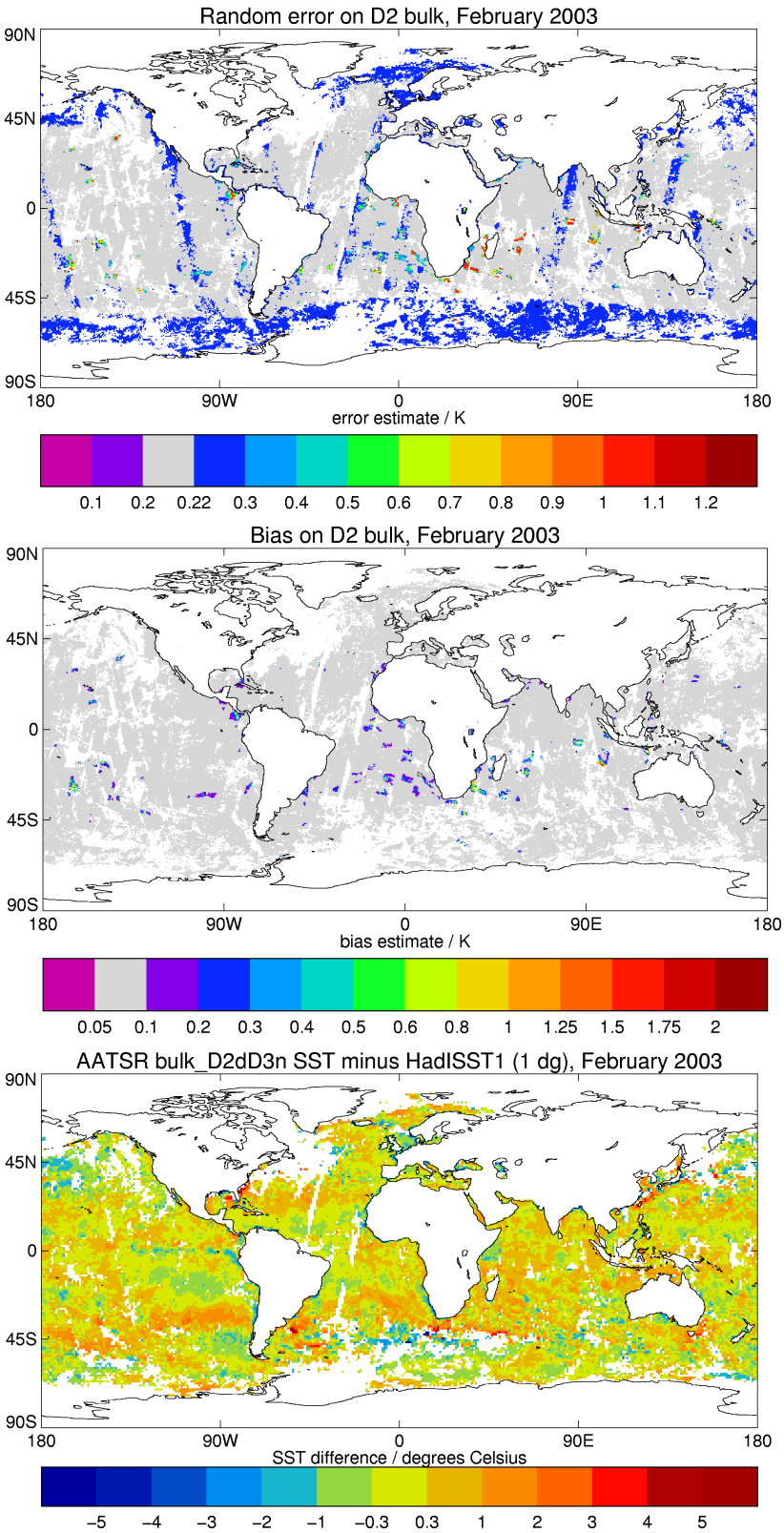


Figure 9. Error estimates in February 2003. (Upper) Random error associated with bulk SST from D2 retrieval (day and night) averaged to half-degree. (Middle) Estimated bias on D2 retrieval averaged to half-degree. (Lower) Difference between monthly mean AATSR bulk SST and HadISST1 at 1° resolution.

Random error associated with D2 bulk SST is between 0.2 and 0.29 K, as expected from the values in Table 3. The skin effect model contribution is larger for low wind speeds, or where there are no wind speed data available. Note that the skin effect is computed by reference to the wind-mixing energy field rather than surface wind speed. Higher random error is also seen where SST is low: large areas of the Southern Ocean and northern Pacific have been identified as below the 8 °C threshold. In some cases these areas correspond to slightly larger anomalies of AATSR compared to HadISST1 in the lower plot. Increased error is also seen where diurnal thermoclines are predicted. These do not match the HadISST anomalies exactly, but they do fall in the right general areas. The true extent of diurnal warming is underpredicted because the necessary flux data were not always available for near-real time processing. Bias associated with the D2 bulk SST is uniform at ~±0.09 K, except where diurnal thermoclines have been predicted. The direction of the bias has not been specified because the systematic errors on NWP fluxes are not known accurately enough (see section 3.2); however, real biases resulting from diurnal warming would be positive. The bias associated with D3 bulk SST (not shown) is a uniform ±0.09 K, while the associated random error is 0.09 K or 0.11 K at low windspeeds (Table 2).

In calculating the error field, we have assumed that the corrections mentioned in section 3.1 have been applied. However, the correction for zonal biases in D2 retrievals, and the global offset, have yet to be defined for AATSR: the absence of these corrections therefore contributes to the differences in the upper panel of Figure 9. Work to define these corrections for AATSR is in progress and should be achieved by end-May 2003. The mean and standard deviation of the AATSR–HadISST1 difference field were 0.11 K and 0.69 K respectively. In Table 3, we suggested that the maximum bias expected in (A)ATSR bulk SSTs was ±0.09 K: this should be easily achieved once the skin SST corrections are included. The 0.69 K standard deviation has a number of contributions. Natural variability in SST over the course of a month is implicit. AATSR data available during the month were not uniformly distributed in space or time, but were meaned for comparison with HadISST on a 1° grid. There are no estimates of error on HadISST, and so the contribution to the 0.69 K from AATSR alone cannot be determined. However, it's likely that the AATSR error indicated by this comparison is greater than the values estimated in Table 3 because potentially contaminated data have not been adequately screened out.

5.2 Additional quality control

For semi-operational applications, it is necessary to perform a quality control on retrieved bulk SSTs to detect contaminated data or unforeseen instrument or processing anomalies. In the existing scheme within the processor, a background SST field is derived from previous AATSR bulk SSTs themselves. A ten-arcminute grid is initialised as the 1961-1990 climatology and each gridpoint is subsequently updated daily with any valid bulk SSTs. D2 and D3 derivations are kept separate, and each cell is only considered valid if it has been updated within the previous 15-day period. A check is made requiring the processed bulk SST to be within 8 K of the background value. There are a number of advantages to this method: the background field is a comparable SST measurement, with D2 and D3 separate, and it may be more accurate than daily analyses developed from alternative data sources. However there are also disadvantages. Some parts of the globe can have few or no valid AATSR observations over the course of a month (witness data gaps on upper panel, Fig. 9): in these areas any valid data are likely to be compared to climatology values, which may be quite different as a result of climate variability rather than deficiencies in the AATSR data. Since comparisons may be made against climatology a high threshold for allowed deviations of 8 K is needed, not really sufficient to detect poor data. Some areas of the globe may have persistently cloud-contaminated AATSR data. Under this system, the background field may develop persistent deviations from "truth" in these areas, and subsequent poor data would not be flagged.

An alternative option is to use daily NWP model analyses of surface temperature as a background field. While the accuracy of an NWP SST analysis may be worse than (A)ATSR (quoted at ~1 K?), the field is updated regularly, is consistent with the analysed daily atmospheric conditions, free from cloud contamination, and globally complete. To safeguard against any minor anomalies and ensure relevance for observations at the beginning and end of a day, we suggest deriving a rolling 3-day mean field, centred on the AATSR observation date.

Figure 10 shows a 3-day mean surface temperature field from NWP, and the corresponding anomalies calculated from a one-day half-degree mean AATSR bulk SST field.

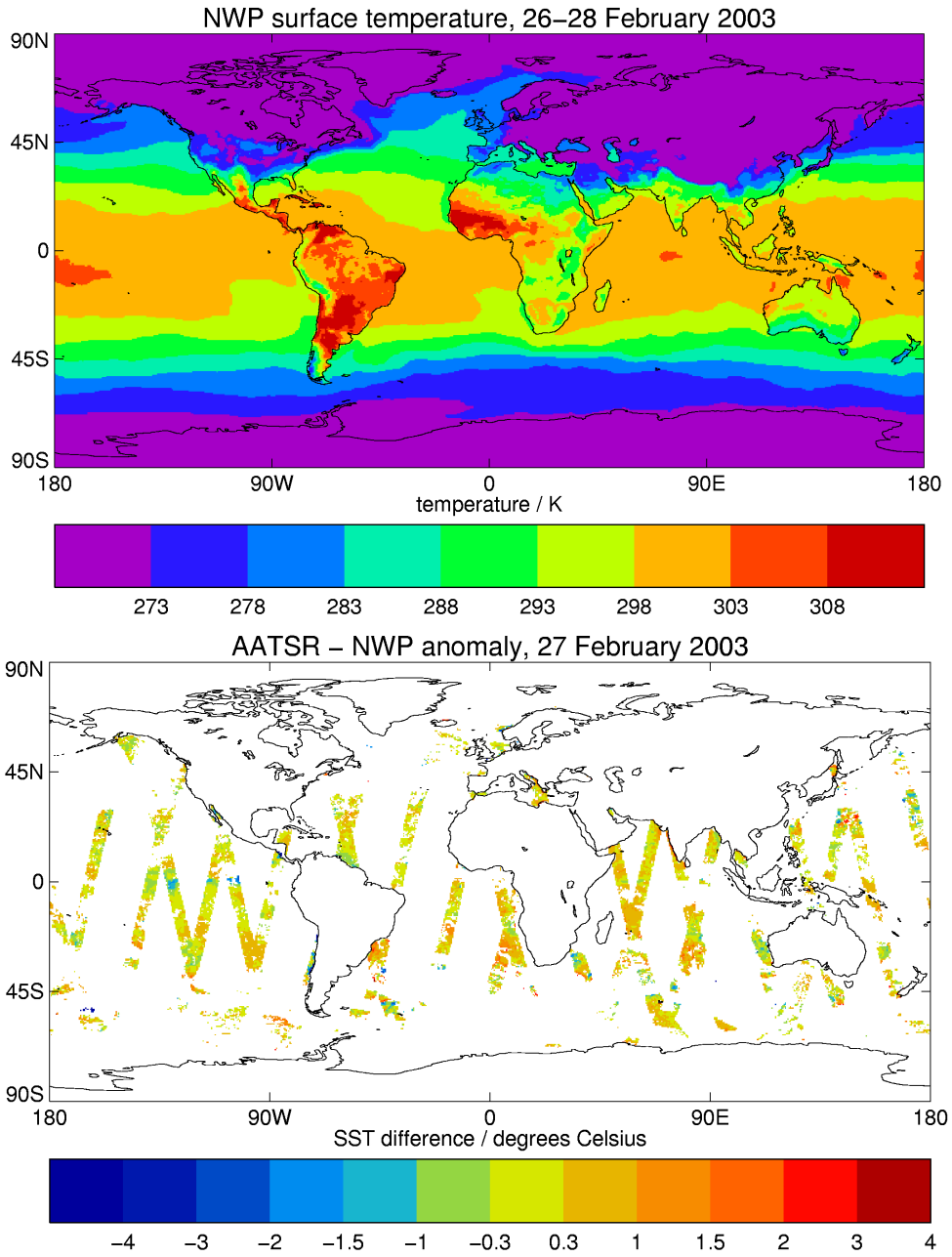


Figure 10. (Upper) A 3-day mean field from the NWP surface temperature analysis. (Lower) Difference between AATSR bulk SST meaned at half-degree resolution and the 3-day mean NWP surface temperature field regridded to the same resolution. Observations flagged with predicted diurnal thermoclines were omitted following current practice in the AATSR processor.

Cool AATSR data in the equatorial Pacific, and southern Pacific and Atlantic may indicate cloud contamination. Warm data in the South Atlantic occur on daylight orbits and may indicate unpredicted diurnal warming (data flagged with predicted diurnal thermoclines were omitted). Figure 11 is a histogram of the anomalies plotted in Figure 10. The differences are characterised by a mean of 0.20 K and standard deviation of 0.76 K. Again, the mean difference is not zero because we have not yet applied an adjustment to the skin SST retrievals for consistency with in situ data. Early AATSR validation indicated that D3 skin retrievals were ~0.2 K warmer than expected compared to buoys (Horrocks et al. 2003b). The standard

deviation is a combination of the errors associated with the two datasets; assuming, for ease of calculation, that all of the AATSR bulk SSTs have random error of 0.20 K (general value for D2 retrievals), then the random error from the background field is ~0.73 K, which is within the accepted error for NWP surface temperature fields. If we consider that points outside of a 3-σ range of the mean represent contaminated data which should be discarded, then any differences outside the range -2.0 to +2.4 K indicate data to be removed. From Figure 10, the few data implicated by this threshold could well result from undetected cloud or diurnal warming.

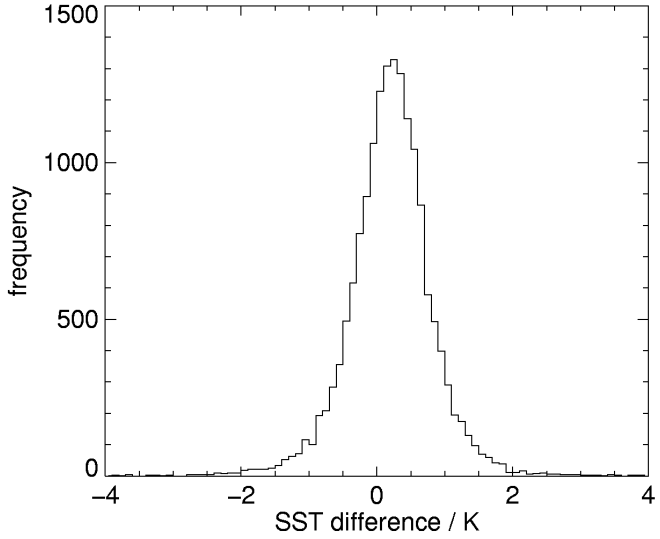


Figure 11. Histogram of differences between AATSR bulk SST for 27 February 2003 and a mean surface temperature field for 26–28 February from NWP analysis.

6. CONCLUSIONS

We have presented a model for the error budget associated with derivations of bulk SST from (A)ATSR spatially-averaged data. For uncontaminated data, we expect estimates of bulk SST to be unbiased at the ±0.1 K accuracy level, provided that: (a) dual-view two-channel retrievals of skin SST are corrected for known latitudinal biases related to non-linear response to total column water vapour; and (b) retrieved skin SSTs are adjusted to offset constant bias arising from limited accuracy in radiative transfer modelling. These corrections are expected from theoretical arguments. The first can be determined from radiative transfer simulations, but the second requires an empirical validation. Work to develop the corrections for AATSR is underway in collaboration with the University of Edinburgh. The random error associated with bulk SSTs is expected to range from 0.09 K to 0.24 K depending on the retrieval algorithm and atmospheric conditions (extreme cold, and wind speed). These are within the error specified in the design of the (A)ATSR instruments for retrieval of skin SST (0.3 K). Bulk SSTs from this series of satellite instruments should meet the stringent requirements of climate monitoring.

Comparisons of AATSR bulk SSTs with HadISST1 indicate that our estimates of error are plausible numbers. Quantitative validation of our estimates is difficult because the errors associated with the climate analysis are unknown. Once the skin SST corrections for AATSR have been developed, we suggest that further validation of our error estimates, using correlative buoy SSTs is desirable. A detailed study of a full year of AATSR–HadISST comparisons is also needed, since interpretation of the difference fields is not simple, but could highlight areas for improvement in both datasets.

We have suggested a quality control check which could be implemented in the (A)ATSR processor to eliminate potentially contaminated data, based on comparison of derived bulk SSTs and a 3-day mean surface temperature field from NWP analysis.

Contamination of (A)ATSR data by undetected cloud continues to present difficulties. While the worst affected data can be screened out through a quality check, it would be preferable to see improvements in the pixel-level cloud detection so that contamination was reduced further. One option for the future may be to integrate a stage for detection of cloud contamination in the spatially-averaged data into our processor. This could be based on comparison of retrieved SST with near-coincident coarser-resolution microwave SST which is unaffected by cloud. Some workers have gone further, to develop merged infrared-microwave SST products (e.g., Wick, 2003). Combination of these data types is probably best achieved through data assimilation into an analysis, although there may be disadvantages to this approach in the case of AATSR, firstly because it may degrade the SST accuracy possible from optimal uncontaminated AATSR data, and secondly because it would reduce the spatial resolution.

ACKNOWLEDGEMENTS

AATSR data are © European Space Agency and were provided within the framework of ENVISAT cal/val activities. ATSR-2 CABT data were provided by R.A.L. We thank N. Rayner of the Hadley Centre for provision of HadISST1 dataset, and A. Stuart-Menteth of the Southampton Oceanography Centre for provision of Figure 8. The clarity of the text was greatly improved following discussions with C. Merchant and J. Eyre.

7. REFERENCES

Allen M. R., Mutlow C. T., Blumberg G. M. C., Christy J. R., McNider R. T., and Llewellyn-Jones D. T. Global change detection. *Nature*, Vol. 370(6484), 24–25, 1994.

Barton I. J., Závody A. M., O'Brien D. M., Cutten D. R., Saunders R. W., and Llewellyn-Jones D. T. Theoretical algorithms for satellite-derived sea surface temperatures. *J. Geophys. Res*., Vol. 94 (D3), 3365–3375, 1989.

Envisat Products Handbook. http://envisat.esa.int/dataproducts/

Fairall C.W., Bradley E.F., Godfrey J.S., Wick G.A., Edson J.B., and Young G.S. Cool-skin and warm-layer effects on sea surface temperature. *J. Geophys. Res*. Vol. 101 (C1), 1295–1308, 1996.

Gleckler P. J, and Weare B. C. Uncertainties in global ocean surface heat flux climatologies derived from ship observations. *J. Climate*, Vol. 10, 2765–2781, 1996.

Horrocks L., O'Carroll A., Candy B., and Saunders R. Progress with the production of an ATSR-1/-2 bulk SST record. *Deliverable (4a/2/00) from Technical Annex 4a of the 2000–01 Climate Prediction Programme*, Met Office, 2001.

Horrocks L. A., O'Carroll A. G., Saunders R. W., Candy B., and Harris, A. R. Near-real time processing of data from the Advanced Along-Track Scanning Radiometer (AATSR) to provide bulk sea surface temperatures. *AATSR Science Document No 1*, Met Office, 2002.

Horrocks L. A., Candy B., Nightingale T. J., Saunders R. W., O'Carroll A., and Harris A. R. Parameterisations of the ocean skin effect and implications for satellite-based measurement of sea-surface temperature. *J. Geophys. Res*., Vol. 108(C3), 3096, doi:10.1029/2002JC001503, 2003a.

Horrocks L.A., Watts J.G., Saunders R.W., and O'Carroll A. Validation of the AATSR Meteo product sea-surface temperature against in situ observations and analyses. *ESA Special Publication* Vol. 531, in press, 2003b.

Isemer H. J., Willebrand J., and Hasse L. Fine adjustment of large scale air-sea energy flux parameterizations by direct estimates of ocean heat transport. *J. Climate*, Vol. 2, 1172–1184, 1989.

Jones M. S., Saunders M. A., and Guymer T. H. Global remenant cloud contamination in the along-track scanning radiometer data: Source and removal. *J. Geophys. Res.* Vol. 101(C5), 12141–12147, 1996.

Kantha L.H., and Clayson C.A. An improved mixed layer model for geophysical applications. *J. Geophys. Res.* Vol. 99 (C12), 25235–25266, 1994.

Kraus E.B., and Turner J.S. A one-dimensional model of the seasonal thermocline II. The general theory and its consequences. *Tellus*, Vol. 19, 98–105, 1967.

Matthiesen S., and Merchant C. Sea surface temperature in marginal ice zones. *Final report on Pathfinder project, Hadley Centre contract number PB/B3529*, MetOffice, 2003.

Merchant C. J. Eliminating bias in satellite retrievals of sea surface temperature. *PhD thesis, University College, London*, 1998.

Merchant C. J., and Harris A. R. Toward the elimination of bias in satellite retrievals of sea surface temperature 2. Comparison with in situ measurements. *J. Geophys. Res.* Vol. 104 (C10), 23579–23590, 1999.

Merchant C. J., and Le Borgne P. Retrieval of sea surface temperature from space, based on modeling of infrared radiative transfer. Submitted, 2003.

Merchant C. J., Harris A. R., Murray M. J., and Závody A. M. Toward the elimination of bias in satellite retrievals of sea surface temperature 1. Theory, modeling and interalgorithm comparison. *J. Geophys. Res.*, Vol. 104 (C10), 23565–23578, 1999.

Murray M. J., Abolins J., Allen M. R., Birks A. R., Dancey K., Mutlow C. T., Merchant C. J., and Harris A. R. Sea-surface temperatures from the ATSR series. *Booklet accompanying the October 2000 release of ATSR2 SST data CD-ROM from Rutherford Appleton Laboratory*, 2000.

Parker D. E., Folland C. K., and Jackson M. Marine surface temperature: observed variations and data requirements. *Climatic Change*, Vol. 31, 559-600, 1995.

Price J.F, Weller R.A., and Pinkel R. Diurnal cycling: observations and models of the upper ocean response to diurnal heating, cooling, and wind mixing. *J. Geophys. Res.*, Vol. 91, 8411–8427, 1986.

Rayner N. A., Parker D. E., Horton E. B., Folland C. K., Alexander L. V., Rowell D. P., Kent E. C., and Kaplan A. Global analyses of SST, sea ice and night marine air temperature since the late nineteenth century. *J. Geophys. Res.*, in press, 2003.

Saunders R. W., Roberts T., O'Carroll A., and Horrocks L. A. Comparison of ATSR and TMI SSTs with the Hadley Centre SST analyses. *NWP Technical Report*, Met Office, in prep. but soon to be available from http://www.metoffice.com/research/nwp/publications/, 2003a.

Saunders R., Horrocks L., and Watts J. Status of (A)ATSR SST data record and feasibility of a reprocessed 15-year dataset for climate studies. *Deliverable (4a/4/02) from Technical Annex 4a of the 2002–03 Climate Prediction Programme*, Met Office, 2003b.

Smith D. L., Delderfield J., Drummond D., Edwards T., Mutlow C. T., Read P. D., and Toplis G. M. Calibration of the AATSR instrument. *Adv. Space Res.* Vol. 28, 31–39, 2001.

Wick G.A. Blended sea-surface temperature product. http://www.etl.noaa.gov/satres/blended_sst.html. 2003.

Závody A. M., Mutlow C. T., and Llewellyn-Jones D. T. Cloud clearing over the ocean in the processing of data from the Along-Track Scanning Radiometer (ATSR). *J. Atmos. Oceanic Tech.* Vol. 17, 595–615, 2000.