• Document prepared in response to CLIVAR Global Synthesis and Observations Panel (GSOP) First Session Recommendation 28 (ICPO, 2005).

Guidelines for Evaluation of Air-Sea Heat, Freshwater and Momentum Flux Datasets

Simon A. Josey[1] and Shawn R. Smith[2]

1) National Oceanography Centre, Southampton, U.K.
2) Center for Ocean Atmospheric Prediction Studies, Florida State University, U.S.A.

Summary

A set of guidelines for evaluation of air-sea heat, freshwater and momentum flux datasets is developed in response to a recommendation of the CLIVAR Global Synthesis and Observation Panel (GSOP). The panel recognised the need for such guidelines in order to facilitate consistent evaluation and intercomparison of the many new flux datasets currently being developed (particularly those from ocean reanalyses). Our approach is to develop guidelines based on the use of both research quality data from flux buoys and research vessels (local evaluation) and large scale constraints (regional, global evaluation). The different evaluation techniques are discussed with reference to recent examples from the research literature. The document is not intended to be an exhaustive review of all methods for flux evaluation but rather to outline the main techniques that should be employed in order to meet the CLIVAR GSOP recommendation.

CLIVAR GSOP Recommendation 28 (ICPO, 2005):

Ocean Reanalysis fluxes: All future ocean reanalysis fluxes should be evaluated against research quality data (flux buoys, research vessels), in addition to large scale constraints (e.g. heat transport), to determine whether reanalysis has led to improvements (individual reanalysis groups).

CLIVAR GSOP Action 29 (ICPO, 2005):

Develop guidelines/recommendations for evaluating ocean reanalysis fluxes (S. Josey, S. Smith).

1. Introduction

Accurate global fields of the air-sea fluxes of heat, freshwater and momentum are vital for further progress in understanding climate change (e.g. IPCC AR4). Historically, gridded monthly mean surface flux datasets, at individual basin (e.g. Jones et al. 1995) to global scales (e.g. Bunker 1976, da Silva et al. 1994, Josey et al. 1999), were produced from in situ observations consisting primarily of voluntary observing ship reports and buoy measurements. More recently, atmospheric model reanalyses (Kistler et al. 2001) and satellite observations (Zhang et al. 2004) have provided

an alternative source of flux estimates which are now being combined using various techniques (e.g. Large and Yeager, 2004; Yu et al. 2004a). The advent of ocean reanalyses (i.e. data assimilation into ocean GCMs) has given rise to a fourth class of direct flux estimates (e.g. ECCO, Stammer et al. 2004). Finally, indirect estimates of the net air-sea heat flux have also been obtained using residual techniques that employ top-of-the atmosphere radiative flux measurements from satellites and estimates of the atmospheric flux divergence from reanalyses (e.g. Trenberth et al. 2001).

Given the diverse range of gridded flux datasets available, a common method of evaluation is needed to provide a means by which their accuracy can be compared and potential biases identified. Previous studies have been limited by the number of available high quality observations that can be used for evaluation. These observations comprise both local measurements of the fluxes from research buoys / vessels and large scale constraints, principally estimates of heat and freshwater transports across hydrographic sections, from which regionally averaged fluxes can be inferred. However, there has been a significant increase in the number of reference observations in recent years and further expansion is anticipated in the future. The surface flux community is now in a position where more extensive evaluations of new gridded flux datasets and the data input to these datasets are possible and indeed desirable.

For clarification of terminology we note that there are two main classes of flux dataset discussed here. The first consists of the large scale 'gridded flux datasets' (typically at spatial resolutions of order one degree and timescales from 6 hourly to monthly) produced from in situ, model or remote sensing sources, or some combination thereof. The second class of datasets are those described in the GSOP Recommendation as 'research quality data', most of which are in-situ point measurements (for example radiative fluxes and meteorological variables from research buoys / vessels) at high temporal resolution (typically available as averages on timescales of order minutes). We refer to such data as 'research quality fluxes' which may be used for the evaluation of the 'gridded flux datasets'.

Note that research quality turbulent (i.e. sensible and latent) heat fluxes and the wind stress are typically derived from the buoy / vessel measurements of the key meteorological variables (wind speed, air temperature and humidity, sea surface temperature) using bulk formulae. Turbulent flux measurements may also be obtained using direct techniques (eddy correlation, inertial dissipation) but these tend to have too much noise for useful comparisons at the level of 6 hour means. The general strategy should therefore be to use collections of direct turbulent flux measurements to more accurately define the bulk formula transfer coefficients and thus improve the accuracy of the bulk formula derived fluxes (see WGASF, 2000 for further details). In addition to the point measurements, area averaged fluxes derived from pairs of hydrographic ocean heat or freshwater transport estimates form a further class of evaluation data.

The need for flux evaluation guidelines has been recognised by the CLIVAR Global Synthesis and Observation Panel who have suggested that such evaluations should become a standard component of the production of new ocean reanalysis surface flux datasets (ICPO, 2005). This paper has been written in response to a recommendation from the panel that guidelines for the evaluation of fluxes should be drawn up. We note that a comprehensive review of flux estimation and evaluation techniques is included as part of a general review of air-sea flux research in the report of the WMO/SCOR Working Group on Air - Sea Fluxes ((WGASF 2000)); the current paper is intended to provide a short, updated summary of evaluation methods specifically for use in assessment of fluxes from ocean reanalyses although it may also be applied to other classes of flux dataset. Our approach is to consider a progression in spatial scale in terms of evaluation techniques that can be applied locally (Section 2), regionally (Section 3) and globally (also Section 3). We summarise the main points and recommendations in Section 4.

2. Local Evaluation

2.1. Data Source

Two main classes of reference observation for local evaluation of fluxes are available:

a.) Research Buoys

To date these have consisted primarily of WHOI moored buoys with more recent deployments, for example the Kuroshio Extension Observatory, by other groups. The research buoys are capable of providing accurate measurements (e.g. Weller et al. 1998) of basic meteorological variables and the downwelling radiative (longwave and shortwave) fluxes, and more recently precipitation. Estimates of the turbulent (latent and sensible) heat fluxes and the wind stress are then obtained from the meteorological variables using bulk formulae (e.g., Smith 1988; Fairall et al. 2003). Bulk formulae are constantly undergoing improvements and the COARE algorithm of Fairall et al. (2003) represents a significant advance relative to previous algorithms. However, it should be noted that the formulae are rarely tested under high wind speed conditions due to a lack of suitable validation data. Additional physical mechanisms may be important under high wind speed conditions and should be considered (e.g. Powell et al 2003, Andreas 2004, Bourassa 2006). Each mooring provides fixed point, relatively long (several months – several years) time series of the meteorological variables (air and sea surface temperature, wind direction and speed, atmospheric humidity and sea level pressure) and the radiative fluxes. Such data have been widely used for flux product evaluations (e.g. Weller et al. 1998; Josey, 2001; Yu et al. 2004b; Cronin et al. 2006). The number of moorings is now increasing via the OceanSites initiative which includes the development of a surface flux reference network as a major goal (http://www.oceansites.org/index.html). Currently, individual mooring programs (e.g. at WHOI) maintain archives for these data. However, an international global archive for surface flux reference site observations is being established as part of OceanSites. Note also that a summary list of links to websites holding research buoy data is included as part of the Florida State University in situ validation data website : http://www.coaps.fsu.edu/RVSMDC/FSUFluxes/validation.php

b.) Research Vessels

Research vessels typically provide shorter individual time series (1-2 months) of the same variables discussed above, with in some cases additional high quality direct turbulent flux measurements (see for example, the NOAA/ESRL/PSD Marine and Air-Sea Interaction group: http://www.etl.noaa.gov/et6/air-sea/). Historically research vessel data have been underutilized (Gould and Smith, 2006) for flux evaluations partly as a result of fragmented data stewardship practices but also because of the relatively short length of the time series. However, they can still be used to provide valuable insights into sources of bias, particularly for the turbulent fluxes (e.g. Renfrew et al. 2002). A good archive exists for the WOCE period (1988-1998; WOCE Data Products Committee, 2002), but otherwise observations tend to be held by vessel operators, individual chief scientists, or national data centres. The Shipboard Automated Meteorological and Oceanographic System (SAMOS) initiative (http://samos.coaps.fsu.edu/html/) is working to improve access and archival. In partnership with SAMOS, the WCRP Working Group on Surface Fluxes is providing guidelines to making climate quality meteorological and flux measurements on research vessels (Bradley and Fairall, 2006).

A primary consideration when using research quality data from moorings or vessels is to ensure that the observations have not been assimilated into the flux product being evaluated. In the case of many routine marine measurements (for example, those from voluntary observing ships, drifters, and some moorings, for example, TAO), the observed values are to some extent assimilated into most operational and many experimental flux products. Thus, withholding research mooring

and research vessel data is essential to ensure these observations are independent from the products under evaluation.

2.2. Method

Local evaluation of gridded flux datasets is carried out by first selecting the gridded values at the same spatial location and temporal period as the research quality data (a process which may involve averaging and/or interpolation of the gridded values) and then comparing the resulting flux and meteorological variable fields. If the spatial location of the research quality data does not correspond directly to a particular grid cell, a common practice is to average / interpolate values from several nearest neighbour grid points surrounding the location of the reference station (mooring or ship). A better approach takes into account the distance between the nearest grid cell and the mooring or ship. The second method allows assessment of the spatial dependencies of the comparison (especially important for moving platforms like ships). In addition, variations in the height of both the product and the in-situ observations need to be taken into account for comparison of air temperature and humidity, and wind speed. As gridded flux datasets tend to be produced at standard heights (typically 10 m for winds, 2 m for air temperature and humidity) while mooring and vessel measurements can be at widely varying heights, the general practice is to adjust the reference observations to the product height using a bulk flux algorithm (e.g. Smith, 1988; Fairall et al. 2003; Bourassa, 2004). However, care has to be taken as the height adjustment itself depends on the value of the surface flux. Thus, if the values of the research quality and gridded fluxes are markedly different, a different result will be obtained by adjusting the gridded product to the observation height, rather than vice versa.

Typically, research quality data are sampled at a temporal interval that is more frequent than the model time step used for gridded flux products. For example, numerical weather prediction (NWP) flux datasets are typically produced every 6 hours. In addition, values from gridded flux datasets can be either instantaneous or integrated quantities. In contrast, most research moorings or vessels can sample at intervals of seconds to minutes. This is necessary, for example, to properly calculate true wind velocity despite the many rapid course changes typical of a surveying research ship. It also allows for extensive quality control and assurance, for example detection of bad radiation data due to RF interference. The high temporal sampling of the reference observations should thus be employed to produce a high quality value that best matches the representative time interval of the gridded flux dataset.

Once gridded flux and research quality data pairs have been created, statistical evaluations can be carried out. The temporal scale of these comparisons depends on the objective of the analysis. Some approaches compare monthly averaged fluxes to determine longer time scale biases in different flux components (e.g. Josey, 2001). However, additional comparisons may also be made at sub-monthly time scales to provide a better understanding of the variability in the flux products on shorter temporal scales (e.g. Smith et al. 2001).

When evaluating gridded flux data using reference quality data, knowledge of the accuracy of both the gridded and reference data, and allowance for it, is important if the results are to be properly interpreted. If the error characteristics of the gridded and reference data are different (for example differing random errors), spurious biases can arise during the comparison. This is well illustrated, in a different context (i.e. non-gridded data), by a comparison of ship and scatterometer winds carried out by Kent et al. (1998). These authors found biases at the high and low wind speed end of the comparison, and an incorrect slope for the regression, that were caused by differing magnitudes of the ship and scatterometer random errors. They concluded that data are best compared using a regression technique (Graybill, 1961) which accounts for the relative sizes of the random errors in the variables considered. Differences in spatial co-location can dominate the estimate of random errors unless tight spatial constraints are applied to the comparison data (Kent et al. 1998, Bourassa et al. 2003). Stoffelen (1998) provides another example of a thorough statistical

technique using triple co-location (winds from scatterometer, buoys, and NWP) to determine uncertainties and relative biases for all data sets. Careful choice of the correct statistical method for the comparison is thus vitally important (see also, Kent and Taylor, 1999).

In addition to statistical considerations, it is important to note that local evaluations should be carried out for both the heat flux components and the underlying meteorological variables (i.e. wind speed, sea surface temperature, air temperature and humidity, and the sea-air temperature and humidity differences). Evaluation of the meteorological variables is key to understanding the causes of any bias that may be present in the flux fields. Note also that, where average quantities are being compared, biases may occur if the correlations between the bulk formula variables are different for the research quality fluxes and the gridded flux dataset.

2.3. Examples

a) Research Buoy

There are many studies in the literature that have used research quality buoy data to evaluate gridded flux datasets (e.g. Moyer and Weller 1997; Weller et al. 1998; Josey et al. 1999; Josey, 2001; Yu et al. 2004; Cronin et al. 2006). The earlier analyses typically focussed on individual or small arrays of buoys. However, the more recent analyses of Yu et al. (2004) and Cronin et al. (2006) make use of larger datasets and are briefly reviewed here.

Yu et al. (2004) employ a combined research buoy and research vessel dataset to evaluate latent and sensible heat fluxes and the underlying meteorological variables from WHOI, NCEP2 and ECMWF in the Atlantic Ocean from 1988 to 1999. The comparisons were carried out using daily averaged research quality data and gridded fields. Standard linear statistics (for example, RMS error, mean bias, correlation) were employed to identify significant differences between the gridded products and the research data, see for example their Fig. 13, reproduced here as Fig. 1. This figure provides a simple bar chart representation of differences between the key fields which could be adopted as a standard for ocean reanalysis flux evaluations.

The analysis revealed that the mean and daily variability of the latent and sensible heat fluxes from the WHOI analysis are an improvement over the NWP fluxes in the sense that they are typically closer to the research quality data (see top row panels of Fig.1 which show biases of 20 to 40 Wm^{-2} in the reanalysis products). The improvement is due not only to the use of a better flux algorithm but also the improved estimates for the meteorological variables (see Yu et al. (2004) for full discussion of the results). The results discussed in their paper demonstrate that it is necessary to consider both the fluxes and meteorological variables in order to fully evaluate new gridded products.

The recent analysis of Cronin et al. (2006) provides a good example of what can be achieved using a widely spaced array of moorings. As part of their study, observations from the Eastern Pacific Investigation of Climate Studies (EPIC) mooring array were used to evaluate the surface radiative fluxes in the NCEP Reanalysis 2 (NCEP2) and 40-yr ECMWF Re-Analysis (ERA40) in the far eastern Pacific intertropical convergence zone and stratus cloud deck regions. The observations spanned a period of nearly four years and included downwelling surface solar and longwave radiation from 10 enhanced Tropical Atmosphere Ocean (TAO) moorings and a Woods Hole Improved Meteorology (IMET) mooring in the stratus region. The analysis revealed that both the NCEP2 and ERA40 incident shortwave is biased low on and north of the equator, and that both models have incident shortwave / longwave radiation biased high / low in the stratus deck region (see their Figs. 2 and 3). These results were subsequently analysed in terms of biases in the model surface cloud forcing (defined as the observed downwelling radiation at the surface minus the clear-sky value).

We note that as regards recent products, evaluations against research buoy measurements have been carried out for NOC 1.1 (Josey et al., 1999; Josey et al, 2002), NOC 1.1a (Grist and

Josey, 1999) and WHOI OAFlux (Yu and Weller, 2004b). However, the Large and Yeager (2004) product, which is based primarily on various adjustments to NCEP meteorological fields has not been evaluated in this way. Thus, it is not possible to say how much of an improvement it is on the original NCEP data and whether it is a more accurate product than others available.

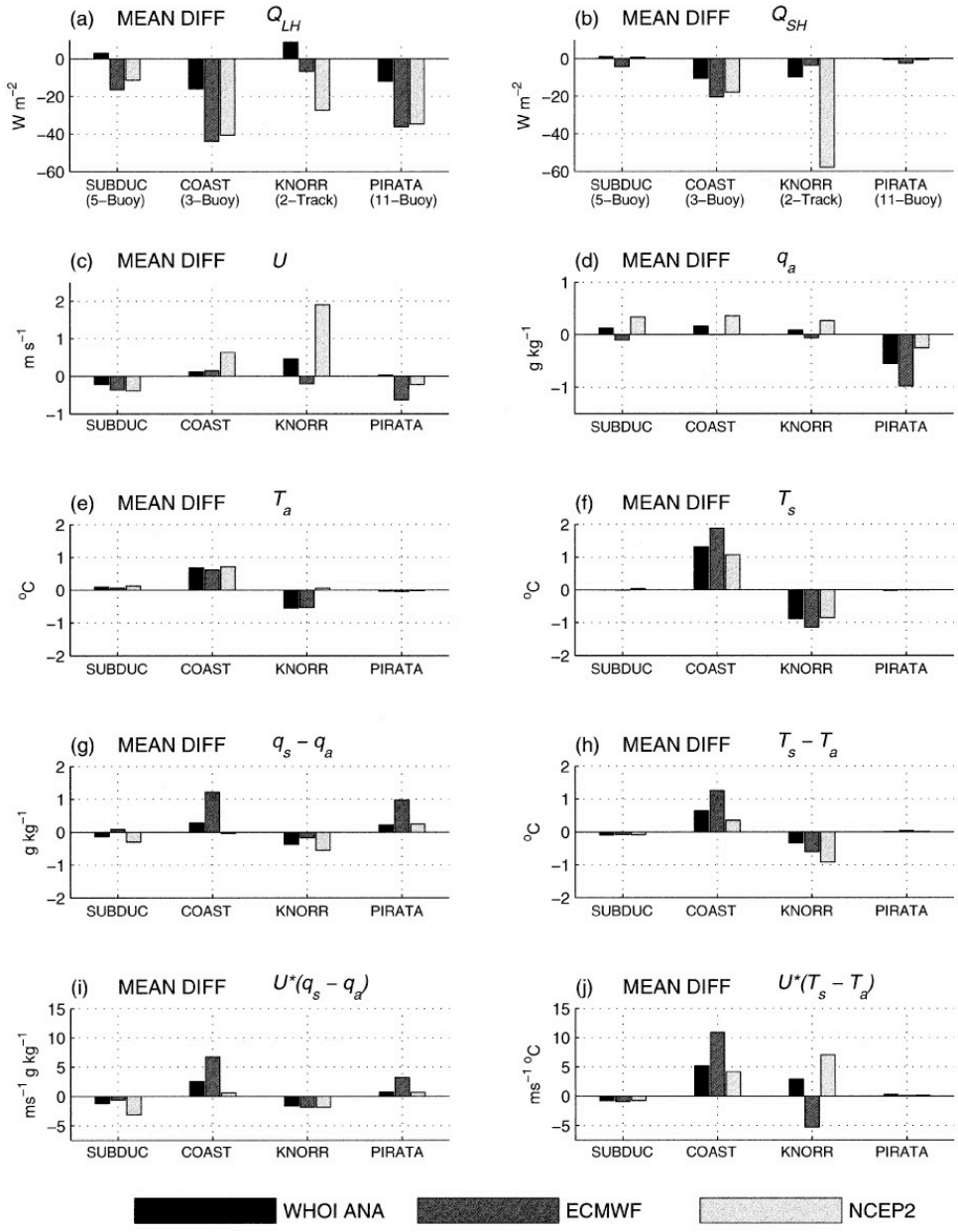

Figure 1. Example figure from Yu et al. (2004, Fig.13) showing the results of a comparison of various flux products with WHOI research buoy meteorological variable measurements and flux estimates. The vertical bars show mean daily differences (buoy – flux product) between the research buoys and gridded flux products from WHOI (WHOI ANA), ECMWF and NCEP (NCEP2), for four different buoy deployments (SUBDUC, COAST, KNORR, PIRATA). The variables are latent (Q_{LH}) and sensible (Q_{SH}) heat flux, wind speed (U), sea surface and air temperature (T_S and T_a) and sea surface and air specific humidity (q_S and q_a).

b) Research Vessel

Two examples of flux product evaluations using research vessel observations are provided by Renfrew et al. (2002) and Smith et al. (2001). In both studies, the observations were used to evaluate NWP flux products. Renfrew et al. (2002) compared observations from a single cruise of the R/V *Knorr* in the Labrador Sea to 6-hourly fluxes from the first NCEP reanalysis and the operational analysis of ECMWF. Smith et al. (2001) compared data from eight research vessels with cruises covering most of the global oceans to the first NCEP reanalysis.

In each case, bulk flux algorithms were employed to adjust the vessel data down to the flux product heights from the models. Subsequently bilinear interpolation was used to determine the model flux value at the vessel location. Finally, six hour averages of the research vessel data were calculated to match the six hour integrations in the NWP model fluxes. In this way, a total of 144 / 4773 matches was obtained for Renfrew et al. (2002) / Smith et al. (2001).

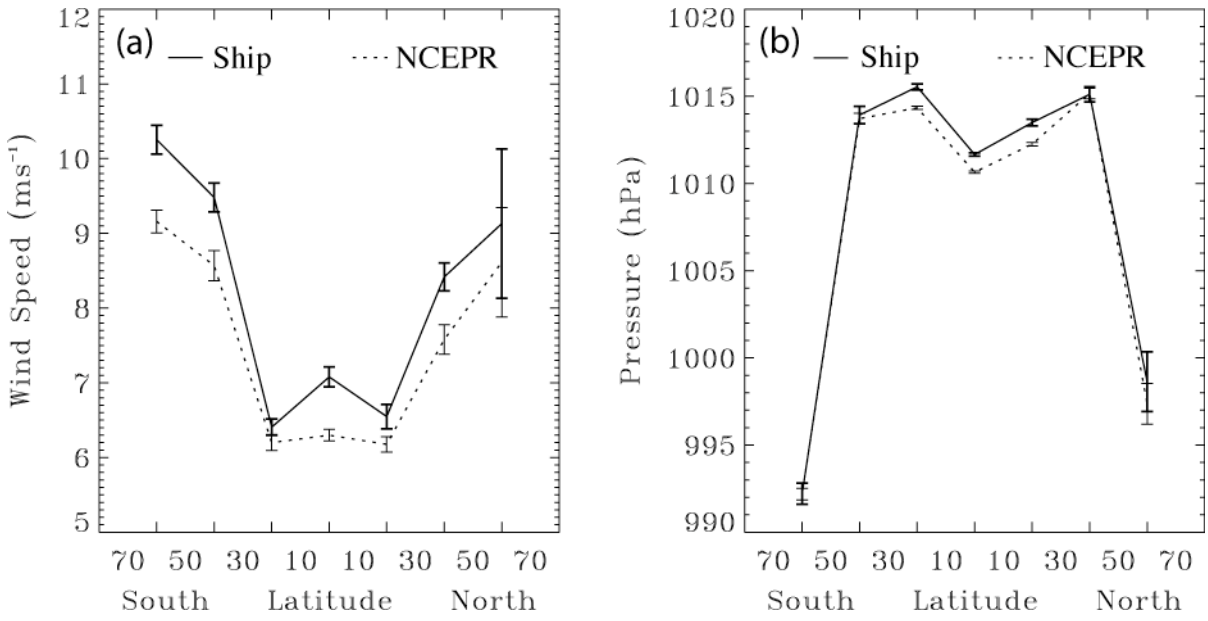

Figure 2. NCEP reanalysis (NCEPR) and WOCE research vessel (ship) measurements of (a) wind speed and (b) sea-level pressure averaged in 20º latitude intervals. Standard error bars are plotted on the mean for each latitude interval. Redrawn version of Smith et al. 1999, Fig. 6.

Smith et al. (2001) found that the NCEP wind speeds were significantly lower than the research quality values over a range of latitude bands (see Fig. 2) and that this was potentially linked to an underestimation of the strengths of the main atmospheric high and low pressure centres in the reanalysis. Renfrew et al. (2002) found that in the Labrador Sea, NCEP overestimates the sensible and latent heat fluxes by 51% and 27%, respectively. They ascribed these biases to an inappropriate choice for the roughness length formula in the NCEP reanalysis under large air–sea temperature difference and high wind speed conditions. Thus, they were able to extend conclusions drawn from an analysis in a specific region to provide an indication of biases that are likely to arise in other regions experiencing similar conditions (e.g. the Gulf Stream and Kuroshio in winter).

3. Regional and Global Evaluation

Regional and global methods of flux dataset evaluation are grouped together in this section as they essentially make use of the same class of constraint : closure of the climatological mean heat and freshwater budgets at either regional or global scales.

3.1. Data Source

Reference evaluation data is provided by hydrographic estimates of the heat and freshwater transport across typically zonal sections within a basin. For reviews of the methods used to obtain these estimates see Bryden and Imawaki (2001) and Wijffels (2001). It is important to note that these estimates may also contain significant errors which should be taken into account using the relevant published error estimates during the evaluation, see example below.

Note also that the level of uncertainty in hydrographic freshwater transport estimates tends to be larger than that for the heat transport. In addition, the effects of runoff need to be taken into account when using freshwater transport estimates as constraints. Consequently, evaluations of flux products using heat transport observations are more widespread than those using freshwater estimates. Thus, the focus here is on the heat flux but the same approach can also be applied to freshwater given knowledge of the runoff term which has improved in recent years (Dai and Trenberth, 2002).

3.2. Method

3.2.1. Regional.

In the regional case, hydrographic section estimates of the heat and freshwater transport are used to identify biases in the gridded flux dataset climatological net air-sea heat flux and net evaporation either through:

a.) Comparison of the climatologically implied property transport (determined by integrating the climatological net flux relative to a given reference section) with the hydrographic estimates of the property transport, or

b.) Comparison of regionally averaged flux estimates from the climatology with corresponding values determined from hydrographic section pairs.

Many examples exist (e.g. da Silva et al. 1994) of evaluations of gridded heat flux datasets by comparison of the implied ocean heat transport with hydrographic estimates i.e. Method (a.). The disadvantage with this approach is that errors accumulate in the transport calculation with the potential to generate differences between the climatological and hydrographic heat transport estimates in regions where the surface fields are reliable. In contrast, regional imbalances in the climatological fluxes linked to processes that are not well represented in the flux calculations may be identified by Method (b.).

3.2.2. Global.

In the global case, the evaluation criterion is simply that the globally averaged net heat flux must average to zero on sufficiently long timescales that heat storage terms are negligible. This criterion is complicated slightly by the effects of global warming with recent estimates suggesting a warming of the ocean by on average 0.2 Wm^{-2} over the last 40 years (Levitus et al. 2005). Thus, closure to within a limit close to this value should be expected on these timescales. At timescales of order a decade or less a larger value is more appropriate, Grist and Josey (2003) adopted a value of 2 Wm^{-2} for their analysis of a climatological dataset spanning the 14 year period 1980-1993. Note that the global average should be made over the entire ocean including the ice covered oceans which is feasible for ocean reanalysis datasets but provides an additional source of uncertainty when evaluating flux products that do not include values over ice. However, as a result of the

relatively small area covered by ice, this is a fairly minor source of error (Josey et al., 1999).

For the freshwater flux, the globally integrated value of net evaporation must equal the total river runoff into the ocean (R) at sufficiently long timescales. Dai and Trenberth (2002) have carried out a recent analysis of river runoff and discuss the various problems inherent in estimates of this variable. They obtain a global total value for R of 37,288 ± 662 $km^3 yr^{-1}$which may be used as an evaluation measure for the total sea-air freshwater exchange in gridded flux datasets.

3.3. Example

Evaluations using the implied ocean heat transport are well established and there is no need to provide a detailed example of this technique here. However, relatively few studies have utilised the regional comparison approach (Method (b.) above). We illustrate this approach with an example drawn from the evaluation of the original SOC climatology (Josey et al. 1999)[1].

Josey et al. (1999) calculated implied surface heat fluxes for a number of boxes bound by hydrographic sections in the Atlantic and compared them with the corresponding NOC1.1 climatological mean values (note the climatology covers the period 1980-1993). The variation of the two sets of estimates with latitude is shown in Fig. 3, in which the box mean values are plotted against the mid-latitude for each box. Error values for the climatological means are an estimate of the upper limit for the random error while the hydrographic errors are obtained by combining the error estimates for each section in a pair. Agreement within the error range was obtained in the majority of regions considered, the main exception being the shaded area in the mid-latitude North Atlantic for which the climatological heat loss is some 70 Wm^{-2} weaker than the hydrographic value. Separate local evaluations using buoys reported in Josey et al. (1999) indicated that this bias is due to an underestimate of the heat lost in the western part of the basin.

It should be noted that the set of regionally averaged net heat flux estimates from hydrography quoted in Fig.3 are based on sections undertaken during the WOCE period. A similar set (which also includes other ocean basins) should be developed using more current post-WOCE (i.e. 1998-present) hydrographic sections. The problem for developing such a set is that at present there is no central archive of section based heat and freshwater transport estimates. Some effort therefore needs to be devoted to generating an archive of transport estimates (including their temporal ranges, latitude and longitude bounds and error estimates) for use as benchmarks for future gridded flux products. There also needs to be a mechanism to update these estimates using the CLIVAR/Carbon repeat hydrography lines.

In addition to evaluations using hydrography, gridded flux datasets are often compared to other gridded products in the literature and conclusions drawn from the level of agreement with the comparison dataset in different regions. This approach has to be employed with extreme caution as the comparison dataset is likely to contain significant biases. For example the widely used Hellerman and Rosenstein (1983) wind stress fields were obtained using an overly strong drag coefficient parameterisation (Harrison, 1989; Josey et al., 2002). Comparisons against gridded products that are based on observationally sound exchange coefficients can provide useful information on differences between gridded datasets but they do not yet provide an established means for validation of new datasets. It should also be noted that for such comparisons to be of use the base time periods for the products must be closely matched, and ideally exactly the same, to prevent differences arising from temporal variability.

---

[1] Note: The original SOC climatology has now been renamed the NOC1.1 climatology, while the adjusted SOC climatology produced by Grist and Josey (2003) has become the NOC1.1a climatology, see the following website http://www.noc.soton.ac.uk/JRD/MET/fluxclimatology.php for details.

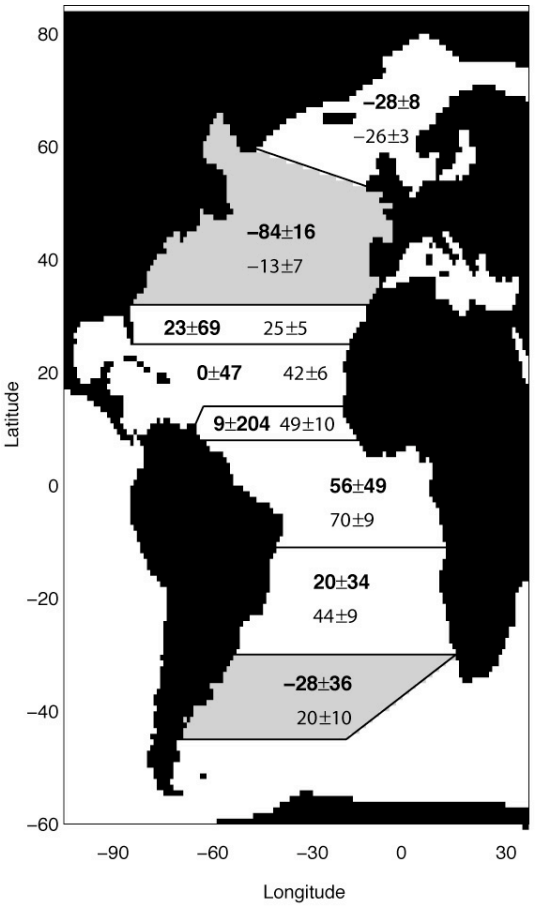

Figure 3. Comparison for the Atlantic Ocean of regionally averaged net surface heat flux estimates from hydrography (bold text, Wm^{-2}, negative values indicate ocean heat loss) with corresponding values from the NOC1.1 climatology (normal text). Grey shading indicates regions in which the values do not agree within the stated error limits. Redrawn version of Josey et al. (1999) Fig.8.

4. Summary and Discussion

An outline method for evaluating gridded flux datasets has been presented in response to a recommendation of the CLIVAR Global Synthesis and Observation Panel (GSOP). The suggested method incorporates a range of measurements from local to regional to global scales. All three scales should be employed where possible for the evaluation of gridded flux datasets. In summary, the key evaluation points are as follows:

a.) Local evaluation of time averaged fluxes and meteorological variables at specific grid locations with corresponding research quality data from surface flux reference moorings and vessels.
b.) Regional evaluation of either gridded flux product ocean transports or, preferably, area averaged fluxes with corresponding research quality data from hydrographic sections.
c.) Global evaluation of gridded flux product area weighted mean fluxes through closure of the appropriate property budget within observational constraints.

Various extensions to these methods are possible, as discussed in the cited literature. These may be considered in addition to the basic set advocated here, which should form the minimum evaluation criteria for gridded flux products. In addition, other types of research quality data may be useful in specific studies. For example, Josey and Marsh (2005) have employed coastal rain gauge observations to evaluate trends in the NCEP and ECMWF reanalysis precipitation datasets in the context of freshening of the North Atlantic sub-polar gyre.

Applications of the outlined method will require access to the research quality evaluation data. Local data from surface flux buoys and, to a lesser extent research vessels are available on the websites listed under Sec 2.1. However, these data are typically at high time resolution (hourly or less) while monthly or 6 hour averaged fields are likely to prove most useful for the community wanting to use them as evaluation data. Some consideration should therefore be given to the establishment of a surface flux reference data archive comprising fields from both research buoys and vessels at these timescales. It may be that this is best generated by effort from within the ocean reanalysis community i.e. as a 'standard' dataset for evaluation of new models.

A further problem noted in Sec 3.3. is the lack of a central archive of heat and freshwater transports determined from hydrographic sections. Previous evaluation studies (e.g. Grist and Josey, 2003) have tended to gather available transport estimates from the literature but as the number of such estimates increases, an archive of reliable estimates, possibly established by the CLIVAR Hydrographic Data office would be desirable. An issue which also needs to be addressed here is the method used to estimate the error on such transport estimates which can vary significantly from one analysis to another.

In some cases it is difficult to establish information regarding use of observations and flux algorithms for certain gridded products and this limits the extent of evaluations that can be carried out using research quality data. In particular, it is often hard to determine which flux algorithms and model settings were used to calculate fluxes in model reanalyses. In addition, it is difficult to quantify the influence of observations on the analyses, much of the voluntary observing ship data may essentially be independent as a result of elimination / downweighting during the NWP model assimilation and could therefore be potentially used for evaluation. In this case, an approach can be taken in which the model is compared to both research buoy and voluntary observing ship data for the same area. If similar results are obtained then the voluntary observing ship data can potentially be used to extend the comparisons to other climatically similar regions (Taylor et al., 2001, Fig.8). This approach may be expected to develop further with the advent of a voluntary observing ship subset for which high-quality meteorological data with extensive associated metadata are available as a result of the VOSClim project (http://www.ncdc.noaa.gov/oa/climate/vosclim/vosclim.html).

Finally, we note that it is also often unclear whether research quality data has been excluded from the analyses. Furthermore, the method of temporal integration tends to vary between products and this also complicates comparisons. These are issues which will also have an impact on evaluations of ocean reanalysis flux products. Some effort therefore needs to be directed towards providing the documentation for atmospheric and oceanic reanalysis products which will allow a full evaluation of their output fields at the air-sea interface.

Despite the various issues discussed above, progress can be made now towards developing a standard method for evaluating gridded flux datasets. Thus, we urge the ocean reanalysis, and surface flux, communities to adopt the method outlined here as a basis for future evaluations of surface meteorological and flux products.

Acknowledgements

We thank Mark Bourassa, Peter Taylor, Bob Weller and Lisan Yu for many useful comments in the preparation of this paper.

References

Andreas, E. L., 2004: Spray Stress Revisited, *J. Phys. Oceanogr.*, **34**, 1429-1440.

Bourassa, M. A., D. M. Legler, J. J. O,Brien, and S. R. Smith, 2003: SeaWinds Validation with Research Vessels, *J. Geophys. Res.*, **108**, 3019, DOI 10.1029/2001JC001081.

Bourassa, M. A., 2004: An Improved Seastate Dependency For Surface Stress Derived from In Situ and Remotely Sensed Winds. *Advances in Space Res.*, 33, 1136-1142.

Bourassa, M. A., 2006: Satellite-based observations of surface turbulent stress during severe weather, Atmosphere - Ocean Interactions, Vol. 2., ed., W. Perrie, Wessex Institute of Technology Press, p. 35 - 52.

Bradley, F., and C. Fairall, 2006: A guide to making climate quality meteorological and flux measurements at sea. NOAA [to go to press in Summer 2006].

Bryden, H. L. and S. Imawaki, 2001: Ocean Heat Transport. Ocean Circulation and Climate. G. Siedler, J. Church and J. Gould, Academic Press: 455-474.

Bunker, A. F., 1976: Computations of surface energy flux and annual air-sea interaction cycles of the North Atlantic Ocean. *Mon. Wea. Rev.*, **104**: 1122-1140.

Cronin, M. F., N. A. Bond, Fairall, C. W. and R. A. Weller, 2006: Surface cloud forcing in the East Pacific stratus deck/cold tongue/ITCZ complex. *J. Clim.* 19(3): 392-409.

Dai, A. G. and K. E. Trenberth, 2002: Estimates of freshwater discharge from continents: Latitudinal and seasonal variations. *J. Hydromet.* 3(6): 660-687.

da Silva, A. M., C. C. Young, and S. Levitus, 1994: Atlas of Surface Marine Data Vol. 1: Algorithms and Procedures. *NOAA Atlas series*, pp.74.

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**(4): 571-591.

Gould, W. J and S. R. Smith, 2006: Research vessels: Underutilized assets for climate observations. *EOS, Trans Amer. Geophys. Union*, 87, 214-215.

Graybill, F. A., 1961: *An Introduction to Linear Statistical Models, Volume 1*, New York, McGraw-Hill.

Grist, J. P. and S. A. Josey, 2003: Inverse Analysis of the SOC Air-Sea Flux Climatology Using Ocean Heat Transport Constraints, *J. Clim.*, **16**(20), 3274-3295.

Harrison, D. E., 1989: On climatological monthly mean wind stress and wind stress curl fields over the World Ocean. *J. Clim.*, **2,** 57 – 70.

Hellerman, S. and M. Rosenstein, 1983: Normal monthly wind stress over the World Ocean with error estimates. *J. Phys. Oceanogr.*, **13,** 1093 - 1104.

International CLIVAR Project Office, 2005: Report of the First Session of the CLIVAR Global Synthesis and Observations Panel (GSOP), 10-12 November 2004. April 2005. International CLIVAR Project Office, CLIVAR Publication Series No. 90.

Jones, C. S., D. M. Legler, and J. J. O'Brien, 1995: Variability of surface fluxes over the Indian Ocean; 1960-1989. *The Global Atmosphere and Ocean System*, 3, 249-272.

Josey, S. A., 2001: A comparison of ECMWF, NCEP/NCAR and SOC surface heat fluxes with moored buoy measurements in the subduction region of the North-East Atlantic. *J. Clim.*, **14**(8): 1780 -1789.

Josey, S. A., E. C. Kent and P. K. Taylor, 1999: New insights into the ocean heat budget closure problem from analysis of the SOC air-sea flux climatology. *J. Clim.*, **12**(9): 2856 - 2880.

Josey, S. A., E. C. Kent, and P. K. Taylor, 2002: On the Wind Stress Forcing of the Ocean in the SOC Climatology : Comparisons with the NCEP/NCAR, ECMWF, UWM/COADS and Hellerman and Rosenstein Datasets. *J. Phys. Oceanog.*, **32**, 1993-2019.

Josey, S. A. and R. Marsh, 2005: Surface Freshwater Flux Variability and Recent Freshening of the North Atlantic in the Eastern Subpolar Gyre, *J. Geophys. Res.*, **110**, C05008, doi:10.1029/2004JC002521.

Kent, E. C., P. K. Taylor, and P. Challenor, 1998: A comparison of ship and scatterometer-derived wind speed data in open ocean and coastal areas, *Int. J. Remote Sensing*, **19**, 3361-3381.

Kent, E. C. and P. K. Taylor, 1999: Accounting for Random Errors in Linear Regression: A Practical Guide, *Q. J. e Royal Met. Soc.,* **125***,* 2789-2790.

Kistler, R., E. Kalnay, W. Collins, S. Saha, G. White, J. Woollen, M. Chelliah, W. Ebisuzaki, M. Kanamitsu, V. Kousky, H. van den Dool, R. Jenne, and M. Fiorino, 2001: The NCEP–NCAR 50-Year Reanalysis: Monthly Means CD-ROM and Documentation. *Bull. Amer. Met. Soc.*, **82**, 247-267.

Large, W. G., and S. G. Yeager, 2004. Diurnal to decadal global forcing for ocean and sea-ice models: the data sets and flux climatologies. NCAR Technical Note NCAR/TN-460+STR, 111 pp.

Levitus S., J. Antonov and T. Boyer, 2005: Warming of the world ocean, 1955–2003, *Geophys. Res. Lett.*, **32**, L02604, doi:10.1029/2004GL021592.

Moyer, K. A. and R. A. Weller, 1997: Observations of surface forcing from the subduction experiment : a comparison with global model products and climatological data sets. *J. Clim.* **10**(11): 2725 - 2742.

Powell, M. D., P. J. Vickery, and T. A. Reinhold, 2003: Reduced drag coefficient for high wind speeds in tropical cyclones. *Nature*, **422**, 279-283.

Renfrew, I. A., G. W. K. Moore, P. S. Guest and K. Bumke, 2002: A comparison of surface-layer and surface heat flux observations over the Labrador Sea with ECMWF and NCEP reanalyses. *J. Phys. Oceanogr.* **32**: 383-400.

Smith, S. D., 1988: Coefficients for sea surface wind stress, heat flux, and wind profiles as a function of wind speed and temperature. *J. Geophys. Res.*, **93**, 15, 467-15,472.

Smith, S. R., D. M. Legler, and K. V. Verzone, 2001: Quantifying uncertainties in NCEP reanalyses using high quality research vessel observations. *J. Clim.*, **14**, 4062-4072.

Stammer, D., K. Ueyoshi, A. Kohl, W. G. Large, S. A. Josey and C. Wunsch, 2004: Estimating Air-Sea Fluxes of Heat, Freshwater and Momentum Through Global Ocean Data Assimilation, *J. Geophys. Res.*, **109**, C05023, doi:10.1029/2003JC002082.

Stoffelen, A., 1998 : Toward the true near-surface wind speed: Error modeling and calibration using triple co-location, *J. Geophys. Res.*, **103**, C4, 7755 - 7766, 1998.

Taylor, P. K., E. F. Bradley, C. W. Fairall, L. Legler, J. Schulz, R. A. Weller and G. H. White, 2001: Surface Fluxes and Surface Reference Sites. in *Observing the Oceans in the 21st Century* (Koblinsky & Smith eds.), ISBN 0642 70618 2, Bureau of Meteorology, Melbourne, Australia, 604pp.

Trenberth, K. E., J. M. Caron and D. P. Stepaniak, 2001: The atmospheric energy budget and implications for surface fluxes and ocean heat transports. *Clim. Dyn.*, 17, 259-276.

Weller, R. A., M. F. Baumgartner, S. A. Josey, A. S. Fischer, and J. Kindle, 1998: Atmospheric forcing in the Arabian Sea during 1994-1995: observations and comparisons with climatology and models. *Deep Sea Res. II*, **45**(11): 1961 - 1999.

WGASF, 2000: Intercomparison And Validation Of Ocean-Atmosphere Energy Flux Fields, Final Report Of The Joint WCRP/SCOR Working Group On Air-Sea Fluxes SCOR Working Group 110, ed. P.K.Taylor.: 312.

Wijffels, S.. 2001, Ocean Freshwater Transport, in *Ocean Circulation and Climate*, G. Siedler, J. A Church and J. Gould, Eds., *Academic Press,* London, p.475-488.

WOCE Data Products Committee, 2002: WOCE Global Data, Version 3.0, Surface Meteorology. WOCE Report No. 180/02, World Ocean Circulation Experiment International Project Office, Southampton, UK, [cdrom]

Yu, L. S., R. A. Weller and B. M. Sun, 2004a: Improving latent and sensible heat flux estimates for the Atlantic Ocean (1988-99) by a synthesis approach. *J. Clim.*, **17**(2): 373-393.

Yu, L. S., R. A. Weller and B. M. Sun, 2004b: Mean and variability of the WHOI daily latent and sensible heat fluxes at in situ flux measurement sites in the Atlantic Ocean. *J. Clim.*, **17**(11): 2096-2118.

Zhang, Y. C., W. B. Rossow, A. A. Lacis, V. Oinas, and M. I. Mishchenko, 2004: Calculation of radiative fluxes from the surface to top of atmosphere based on ISCCP and other global data sets: Refinements of the radiative transfer model and the input data. *J. Geophys. Res.*, **109**, D19, ISI:000224429700004.