Remote Sensing Of Environment
November 2017, Volume 201 Pages 196-218
http://dx.doi.org/10.1016/j.rse.2017.08.016
http://archimer.ifremer.fr/doc/00403/51403/

Review and assessment of latent and sensible heat flux accuracy over the global oceans

Bentamy Abderrahim [1, *], Piolle Jean-Francois [1], Grouazel Antoine [1], Danielson R. [5], Gulev S. [6], Paul Frederic [1], Azelmat Hamza [1], Mathieu P. P. [2], Von Schuckmann Karina [3], Sathyendranath S. [4], Evers-King H. [4], Esau I. [5], Johannessen J. A. [5], Clayson C. A. [7], Pinker R. T. [8], Grodsky S. A. [8], Bourassa M. [9], Smith S. R. [9], Haines K. [10, 11], Valdivieso M. [10, 11], Merchant C. J. [10, 11], Chapron Bertrand [1], Anderson A. [12], Hollmann R. [12], Josey S. A. [13]

[1] Inst Francais Rech & MER IFREMER L4exploitat, Issy Les Moulineaux, France.
[2] European Space Agcy ESA ESRIN, Frascati, Italy.
[3] Mediterranean Inst Oceanog, Marseille, France.
[4] Plymouth Marine Lab, Plymouth, Devon, England.
[5] Nansen Environm & Remote Sensing Ctr, Bergen, Norway.
[6] PP Shirshov Inst Oceanol IORAS, Moscow, Russia.
[7] Woods Hole Oceanog Inst, Woods Hole, MA 02543 USA.
[8] Univ Maryland, College Pk, MD 20742 USA.
[9] Florida State Univ, Tallahassee, FL 32306 USA.
1 [0] Univ Reading, Natl Ctr Earth Observat, Reading, Berks, England.
1 [1] Univ Reading, Meteorol Dept, Reading, Berks, England.
1 [2] European Org Exploitat Meteorol Satellites EUMETS, Darmstadt, Germany.
1 [3] Natl Oceanog Ctr, Southampton, Hants, England.

* Corresponding author : Abderrahim Bentamy, email address : Abderrahim.Bentamy@ifrerner.fr

Abstract :

For over a decade, several research groups have been developing air-sea heat flux information over the global ocean, including latent (LHF) and sensible (SHF) heat fluxes over the global ocean. This paper aims to provide new insight into the quality and error characteristics of turbulent heat flux estimates at various spatial and temporal scales (from daily upwards). The study is performed within the European Space Agency (ESA) Ocean Heat Flux (OHF) project. One of the main objectives of the OHF project is to meet the recommendations and requirements expressed by various international programs such as the World Research Climate Program (WCRP) and Climate and Ocean Variability, Predictability, and Change (CLIVAR), recognizing the need for better characterization of existing flux errors with respect to the input bulk variables (e.g. surface wind, air and sea surface temperatures, air and surface specific humidities), and to the atmospheric and oceanic conditions (e.g. wind conditions and sea state). The analysis is based on the use of daily averaged LHF and SHF and the associated bulk variables derived from major satellite-based and atmospheric reanalysis products. Inter-comparisons of heat flux products indicate that all of them exhibit similar space and time patterns. However, they also reveal significant differences in magnitude in some specific regions such as the western ocean boundaries during the

Northern Hemisphere winter season, and the high southern latitudes. The differences tend to be closely related to large differences in surface wind speed and/or specific air humidity (for LHF) and to air and sea temperature differences (for SHF). Further quality investigations are performed through comprehensive comparisons with daily-averaged LHF and SHF estimated from moorings. The resulting statistics are used to assess the error of each OHF product. Consideration of error correlation between products and observations (e.g., by their assimilation) is also given. This reveals generally high noise variance in all products and a weak signal in common with in situ observations, with some products only slightly better than others. The OHF LHF and SHF products, and their associated error characteristics, are used to compute daily OHF multiproduct-ensemble (OHF/MPE) estimates of LHF and SHF over the ice-free global ocean on a 0.25° × 0.25° grid. The accuracy of this heat multiproduct, determined from comparisons with mooring data, is greater than for any individual product. It is used as a reference for the anomaly characterization of each individual OHF product.

Highlights

► Establishing reference input dataset maximizing the use of remotely sensed data ► Performing a cross-comparison of different heat flux algorithms and approaches ► Generating an ensemble of turbulent fluxes, including multiple approaches ► Evaluating the quality and consistency of ensemble realizations ► Exploiting integral heat constraints at local, regional and global scales

Keywords : Ocean Heat Flux, Latent heat flux, Sensible heat flux, Ocean heat content, Scatterometer, Surface wind, Specfic air humidity, OceanSites, Remotely sensed data

1 Introduction

Accurate estimation of the ocean surface turbulent and radiative fluxes is of great interest for a variety of air-sea interaction and climate variability issues. Surface fluxes of heat, moisture, momentum, and gases play a key role in the coupling of the Earth's climate system and control many important feedbacks between the ocean and the atmosphere (Gulev *et al*. 2013). Furthermore, consistency studies of turbulent flux estimates and ocean heat storage estimates are also essential for constraining the Earth's energy budget in order to "track" the energy flows through the climate system, which in turn is critical for improving understanding of the relationships between climate forcings, the Earth system responses, climate variability and future climate change (Trenberth et al., 2009; von Schuckmann et al., 2016). The longest time series of surface fluxes going back to the mid 19th century can be derived from the Voluntary Observing Ship (VOS) data (Woodruff *et al*., 2011, Gulev *et al*., 2013). However, these data are characterized by insufficient and time-dependent sampling (Gulev *et al*. 2007), and by inaccuracies in state variables used for flux computation (e.g. Josey *et al*. 1999, 2014). In contrast, atmospheric re-analyses, as well as remotely sensed data, potentially provide much more homogeneous time series of atmospheric state variables for surface flux computation. However, remotely sensed data are limited in time to a few decades while reanalyes can be strongly influenced by variations in the type and amount of data assimilated, particularly across the transition to the satellite era in the early 1980s.

In addition, surface flux products from reanalyzes and remote sensing are also subject to biases and uncertainties and require further improvement for turbulent flux determination. These include; improvements in spatial and temporal resolution, the accuracy, and the characterization of the spatial and temporal distribution of errors of each flux component. It is

one of the priorities of the World Climate Research Program (WCRP) to improve the accuracy of surface fluxes for climate studies to within "a few W/m²" and 10 W/m² for individual flux components and the large scale net heat fluxes, respectively (e.g. WGASF, 2000, Bradley and Fairall, 2007).. The Southern Ocean Observing System (SOOS) group recommends a better flux observation density for improving heat flux accuracies at regional scales (Gilles *et al*, 2016). These requirements impose challenges including the development of new parameterizations, achievement of global and regional heat budget closure, reducing sampling uncertainties, and better scaling parameters for surface flux estimates.

To meet these community requirements, the European Space Agency (ESA) launched a project called Ocean Heat Flux (OHF (http://www.oceanheatflux.org/) aiming at development, validation, and evaluation of satellite-based estimates of surface turbulent fluxes and their documentation, particularly those derived from ESA satellite/mission earth observation (EO) data, as well as all bulk parameters needed for turbulent flux calculations over the global ocean. OHF involves a number of objectives and studies. The main OHF objectives include (but are not limited to); establishing a reference surface flux dataset (to maximize the use of remotely sensed data including ESA products), development and accuracy assessment of an ensemble of ocean heat turbulent flux products available over decadal or longer timescales (in order to foster the use and validation of ESA mission data).

For these purposes, OHF uses in-situ, satellite-based, blended or synthetic, and reanalysis-derived surface fluxes over the global ocean, with synoptic and sub-synoptic spatial resolution for the period 1999 – 2009. The project makes use of the most modern global satellite surface flux data sets such as those from IFREMER (Institut Français pour la Recherche et l'Exploitation de la MER; France), HOAPS (the Hamburg Ocean Atmosphere Parameters and Fluxes from Satellite; Germany), SEAFLUX (Woods Hole Oceanographic Institution, Woods Hole (WHOI); USA), and J-OFURO (Japanese Ocean Flux Data sets with

Use of Remote Sensing Observations; Japan). These are used along with surface turbulent fluxes from three modern reanalyzes: ERA-Interim (Dee *et al*., 2011), NCEP-CFSR (Saha *et al*., 2010) and NASA MERRA (Rienecker *et al.,* 2011), as well as the synthetic OAFLUX product (Yu and Weller 2008) and the VOS based NOCS2 surface flux climatology (Berry and Kent, 2009). Because these flux products were derived using different approaches and data sources, they all have their strengths and weaknesses. Wide use of these products for different climate applications such as (among others) forcing ocean models (e.g., Ayina *et al.*, 2006), analyzing ENSO dynamics (Mestas-Nuñez *et al.*, 2006, 2013), and/or evaluating the intra-seasonal variability (Grodsky *et al*, 2009) requires a detailed quantitative assessment of each product's limitation and of and inter-product differences.

This study presents pilot results from the OHF project that describe uncertainties of the different flux products. Such intercomparison is supplemented by the validation of individual surface flux components against estimates based on in-situ buoy and ship data, especially buoy data included in the Flux reference OceanSites network (http://www.oceansites.org/). Consideration is also given to a new approach to using observations that are themselves incorporated into the flux products that are being validated.

The datasets used in this study are described in Section 2, while OHF products, all available at the same space and time resolution, are described in Section 3. Section 4 demonstrates the impact of recalibration on each OHF product. Regional product inter-comparisons are introduced in Section 5. The accuracy and quality of each OHF flux product, and the ensemble mean flux product, is discussed in Sections 6 and 7.

2 Flux products

2.1 IFREMER

In this study, we use the new IFREMER turbulent fluxes (version 4) available daily over the global ocean on a 0.25° regular grid. It is an updated version of (Bentamy *et al*, 2013). The bulk variables such as surface wind speed (U_{10}) and specific air humidity (q_a) at 10 m height are estimated from remotely sensed observations. U_{10} is mainly obtained from scatterometers onboard ERS-1 (1992 – 1996), ERS-2 (1996 – 2001), and QuikSCAT (1999 – 2009) satellites. More specifically, the main change with respect to IFREMER version 3 described in Bentamy *et al*. (2013) is the use of new ERS-1 and ERS-2 wind retrievals (Bentamy *et al*, 2013 and 2016). To enhance the sampling of surface winds, version 7 of wind speed from Special Sensor Microwave Imager (SSM/I) onboard Defense Meteorological Satellite Program (DMSP) F10, F11, F13, F14, and F15 satellites (Wentz, 2013)is used as ancillary data.

Specific air humidity is derived, over special sensor microwave imager (SSM/I) radiometer swaths, based on the use of the model relaying brightness temperature measurements (Tb) and q_a (Bentamy *et al*, 2013). For this study, a new reprocessing of q_a is performed with respect to the use of the recently reprocessed fundamental climate data record (FCDR) brightness temperatures (Sapiano *et al*, 2012).

2.2 HOAPS

Data used in this project are from HOAPS-3, which utilizes passive microwave data from SSM/I to retrieve bulk variables. HOAPS-3 latent heat flux is based on the bulk COARE3 algorithm (Fairall *et al.*, 2003) This algorithm requires atmospheric specific humidity (implemented after Bentamy *et al.*, 2003), sea surface saturation specific humidity

(q_s), as well as near surface wind speed (U_{10}). Sea surface temperature (*SST*) for q_s estimation is taken from the NODC/RSMAS Pathfinder SST (Casey *et al*, 2010), which uses AVHRR observations adjusted to drifting buoy data. This is, therefore, a 'bulk' *SST*, whereas ideally q_s should be estimated from a skin temperature, which can differ by a few tenths of degree Kelvin. HOAPS 3 near surface wind speed is retrieved from SSM/I measurements by a neural network approach. The HOAPS 3 LHF and SHF fluxes, as well as the related bulk variables, are derived from the newly daily analyses of HOAPS fluxes. They are calculated from swath retrievals based on the use of space and time interpolation method over a grid map of 0.50°×0.50°.

2.3 SEAFLUX

The SEAFLUX data are available over the global ice free ocean at high space (0.25°×0.25°) and time (3-hourly) resolution. Data are available from January 1998 through December 2007 (Clayson *et al*, 2013, see also http://SEAFLUX.org). Latent and sensible heat fluxes are estimated using the COARE3.0 algorithm. Wind speed at z=10m is obtained from the Cross-Calibrated Multi-Platform (CCMP) Ocean Surface Wind Components data (Atlas *et al.*, 2011). CCMP wind is calculated from cross-calibration and assimilation of wind retrievals from SSM/I, TMI, AMSR-E, QuikSCAT, and SeaWinds onboard ADEOS-2. These satellite wind retrievals are combined with atmospheric ECMWF ERA-40 and ECMWF operational analysis (January 1999 through June 2009). CMMP data are available at synoptic times (00h:00, 06h:00, 12h:00, 18h:00 UTC) on a 0.25°×0.25° grid. The specific air humidity at 10m and air temperature (*Ta*) are both retrieved from microwave brightness temperature (*Tb*) using the neural network method described in Roberts *et al.*, (2010). The method requires *SST* that is taken from NOAA *SST* (Reynolds et al., 2007).

The SEAFLUX product is three-hourly (averaged from 0000-0300Z, 0300-0600Z, 0600-0900Z, etc.). All variables are currently available from January 1, 1998, through December 31, 2007.

2.4 *J-OFURO*

Japanese Ocean Flux Data Sets with Use of Remote Sensing Observations (J-OFURO) provides global ocean latent and sensible heat fluxes and related bulk variables. In this project, J-OFURO version 2 referred as HF004 (http://dtsv.scc.u-tokai.ac.jp/j-ofuro/) is used. The fluxes are obtained based on the COARE3.0 bulk algorithm. The input parameter q_{a10} is derived from F10, F11, F13, and F14 SSM/I Tb using the empirical model (Schlussel *et al*, 1995). Wind speed at 10m is estimated using all available satellite data (Tomita and Kubota, 2006) including radiometers such as SSM/I, Aqua/AMSR-E, and TRMM/MI, and scatterometers onboard ERS-1, ERS-2, and QuikSCAT. Air temperature is derived from NCEP-2 re-analysis. Finally, J-OFURO2 *SST* is obtained from the newly merged Japan Meteorological Agency (JMA) multi-satellite and in situ product (MGDSST, Kurihara *et al.*, 2006, see also http://dtsv.scc.u-tokai.ac.jp/j-ofuro/index.html).

2.5 *OAFLUX*

The OAFLUX data are available for the 1985-2014 at daily resolution on a 1°× 1° grid (Yu *et al.*, 2008). For the flux computations OAFLUX uses the NOAA daily 0.25° *SST* (Reynolds *et al*, 2007). In addition to the NOAA *SST* dataset, OAFLUX also utilizes *SST* values from ERA-40 and NCEP-1 reanalyses. The *SST* data from the re-analyses are re-gridded by WHOI to 1° resolution for ease of synthesis with the Reynolds *SST* data through the objective analysis based on the Gauss-Markov approach and used for all surface meteorological variables. For estimation of specific air humidity at 2m (q_{a2},) OAFLUX

applies the Chou *et al*. (2003 and 2004) algorithm. Further blending of humidity fields employs also specific humidity from the NCEP and ECMWF re-analyses as inputs to objective analysis. For wind speed, OAFLUX uses QuikSCAT and version 6 of SSM/I data. The algorithm used to derive the SSM/I data is described in Wentz (1997). Wind data used in OAFLUX are 12-hourly averages at a swath resolution of 25 km. In addition, OAFLUX also utilizes AMSR-E data as well as data from NCEP and ECMWF re-analyses. A variational method applied in OAFlux is subjective due to the determination of weights. For flux computations, the analyzed winds are adjusted to the 10 m height and to the neutral stability. Air temperatures in OA-Flux are from NCEP and ECMWF re-analyses. Starting from 2002 OA-Flux air temperature is based on ERA-interim only. Bulk variables are converted to turbulent fluxes using the COARE-3 algorithm. Further details of OAFLUX development procedures are available in Yu *et al*. (2008) and at http://oaflux.whoi.edu/data.html.

2.6 *ERA-Interim*

Era-Interim (Simmons *et al*., 2006) refers to one of the reanalyses of atmospheric parameters produced by the ECMWF. It uses 4D-variational analysis on a spectral grid. This reanalysis covers the period from 1989 to the present day. The ERA-Interim data used in this study are on a 0.75° regular grid. The main parameters used are dew temperature and air temperature at 2m height available at synoptic times (00h:00, 06h:00, 12h:00, 18h:00 UTC), which are converted to q_{a10} and to T_{a10} utilizing the COARE3.0 algorithm. The quality of q_{a10} and of T_{a10} is checked through comparisons with moored buoy estimates. The main finding of interest is that ERA-I based Ta_{10} is underestimated for buoy Ta_{10} exceeding 20°C. A bias correction is determined from linear regression between ERA Interim and buoy Ta_{10} estimates.

The dedicated web site for ERA Interim data and documentation is (http://apps.ecmwf.int/datasets/data/interim_full_daily/)

2.7 *CFSR*

NCEP Climate Forecast System Reanalysis (CFSR) (http://rda.ucar.edu/pub/cfsr.html), developed by the US NOAA NCEP. The data used for this study are from the NOAA's National Operational Model Archive and Distribution System (NOMADS), which is maintained by the NOAA's National Climatic Data Center (NCDC) (Saha *et al*, 2010). The coupled model consists of a spectral atmospheric model at a resolution of T382 (38km) with 64 hybrid vertical levels and the GFDL Modular Ocean Model. The atmosphere and ocean models are coupled with no flux adjustment. The NCEP-CFSR uses the gridded statistical interpolation (GSI) data assimilation system for the atmosphere. Flow dependence for the background error covariances is included as well as first order time interpolation to the observation. Variational quality control of observations (Andersson and Järvinen, 1999) is also included. An ocean analysis for *SST* is also performed using Optimal Interpolation (OI). A full range of observations is used as in the other re-analyses which are quality controlled and bias corrected, including satellite radiances. Observations of ocean temperature and salinity are also used.

Details of CFSR data are available in (http://cfs.ncep.noaa.gov/cfsr/)

2.8 *MERRA*

The Modern-Era Retrospective Analysis for Research and Applications (MERRA; Bosilovich, 2008) is a reanalysis from NASA extending from 1979 to the present. It is routinely used to analyze NASA Earth Observing System (EOS) satellite data as well as conventional observations and operational satellite data in support of NASA science and field

missions. Rienecker *et al* (2011) provide an overview of MERRA. Surface winds are assimilated over the ocean using data from Special Sensor Microwave Imager (SSM/I) and scatterometer retrievals. Sea surface temperature and sea ice are prescribed from the Reynolds dataset (Reynolds *et al*, 2002). The prognostic variables atmospheric temperature and moisture at the lowest model level are used for computing the vertical gradients in moisture and temperature needed for calculation of the latent and sensible heat fluxes. The planetary boundary layer (PBL) scheme parameterization uses the Lock *et al*. (2000) and the Louis *et al*. (1982) schemes for unstable and stable conditions, respectively. Neutral transfer coefficients are computed based on standard similarity relationships using a momentum roughness length based on (Charnock, 1955), a roughness length for heat based on (Beljaars, 1995), and a roughness length for moisture that is a factor of 1.5 larger than the roughness length for heat. Air humidity and temperature at 10m height are estimated as diagnostic outputs based on the computed fluxes and transfer coefficients. MERRA data are available hourly on a 0.625°×0.5° longitude × latitude grid.

2.9 Moorings

Data from about 200 moored buoys were collected and investigated prior to any use for flux product validation purposes.

Buoy measurements provide several oceanic and/or atmospheric variables required for turbulent flux estimation. Twelve moorings located off the French and England coasts are maintained by UK Met-Office and/or Météo-France (MFUK), 96 buoys located off and near U.S coasts are maintained by the U.S. National Data Buoy Center (NDBC), 66 buoys of the TAO array are located in the equatorial Pacific, 13 buoys of the PIRATA network are located in the equatorial Atlantic, and 6 RAMA moorings in the Indian Ocean. TAO, PIRATA, and RAMA buoys will be hereafter referred as Tropical buoys (Figure 1). Buoy data are hourly

available at heights varying between 3m and 10m. Buoy wind speeds, specific air humidity, air temperature are converted to values at the standard 10m height by the COARE3.0 algorithm. The latter is also used to estimate buoy turbulent fluxes from buoy bulk variables.

High quality bulk variable measurements are obtained from the OceanSites buoy network (http://www.oceansites.org). These moorings are an integral part of the Global Ocean Observing System (GOOS). Most of the OceanSites buoys are located in the tropical zones of the Atlantic, the Indian, and the Pacific oceans. Only Kuroshio Extension Observatory (KEO) buoys are extra-tropical moorings. The number of OceanSites buoys increased from 7 in 1999 to 37 in 2009 (Figure 1).

One should notice that most of the measurements from moorings used in this study such as 10m wind, 2m air temperature, and 2m relative humidity are assimilated by ERA Interim, CFSR, and MERRA.

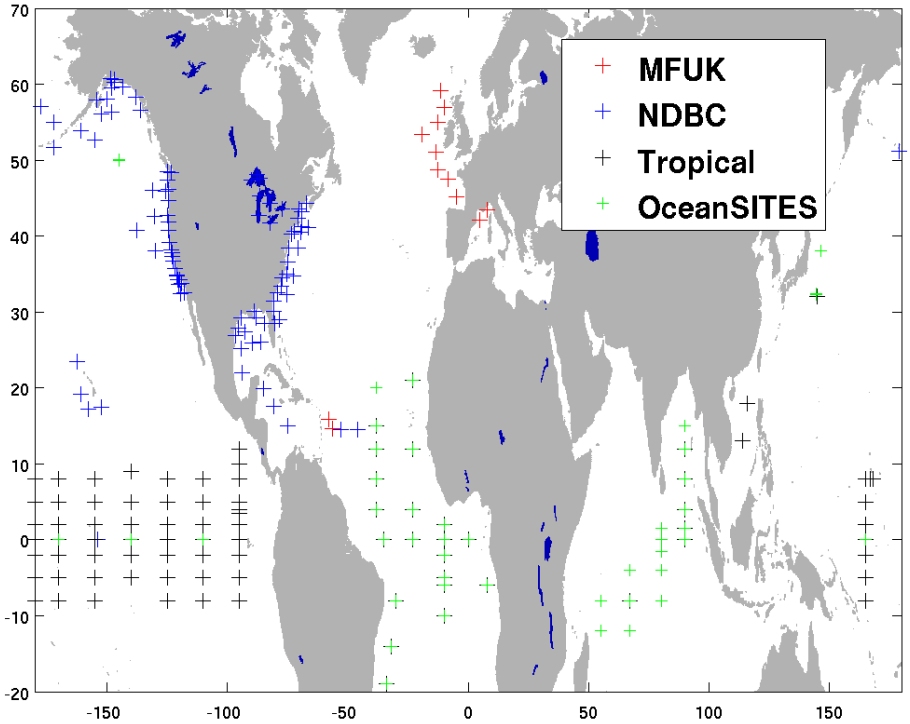

Figure 1: Moored buoy Locations

Turbulent fluxes are calculated from validated hourly buoy 10m wind speed, specific air humidity, and air temperature in combination with sea surface temperature. The adjustment to the 10m height of basic variables (U_{10}, q_{a10}, T_{a10}) as well the estimation of turbulent fluxes is performed using COARE3.0 algorithm. For each day, daily averaged buoy estimates of bulk variables and heat fluxes are calculated if at least 6 hourly measurements are available during day and night.

2.10 ICOADS ships and buoys

A compilation of in situ surface marine observations in the International Comprehensive Ocean–Atmosphere Data Set (ICOADS Version 3; Freeman *et al*. 2017**)** is also considered as a reference for analyses. ICOADS includes quality controlled ship and mooring data collected over years by various countries. ICOADS quality indicators are used to select observations between January 2000 and December 2007 that are within 2.8 standard

deviations of a smoothed monthly climatology. Turbulent heat fluxes are calculated using the COARE 3.0 algorithm for each ship and/or buoy observation with the required bulk variables (i.e., sea level pressure, *SST*, wind speed, air temperature, and dew point). Each ICOADS heat flux is then associated with the center of the nearest grid box of the standardized analysis (at ¼-degree resolution) and daily averaged. About 2.6 million collocations that are common to all OHF products are then retained. The resulting distribution of daily observations (Figure 2) provides relatively good spatial coverage in the Northern Hemisphere midlatitudes but mostly poor coverage elsewhere particularly in the Southern Ocean which is largely unsampled. Good temporal coverage occurs only at mooring locations (Figure 1) and along major ship routes. . It is convenient to divide this dataset by even and odd year and by common and extreme flux. The odd year subset permits an independent check on calculations. Below, only the even-year subset is discussed but all conclusions apply equally to the odd-year subset (and all heat flux collocations are available at Danielson 2017). Extreme fluxes (greater than a few hundred Wm^{-2}) are further ignored in the calculation of covariance, following Hubert et al. (2012). Because covariance is sensitive to outliers (McColl et al., 2014; Su et al., 2014), collocation groups are trimmed by about 10% before other calculations are performed.

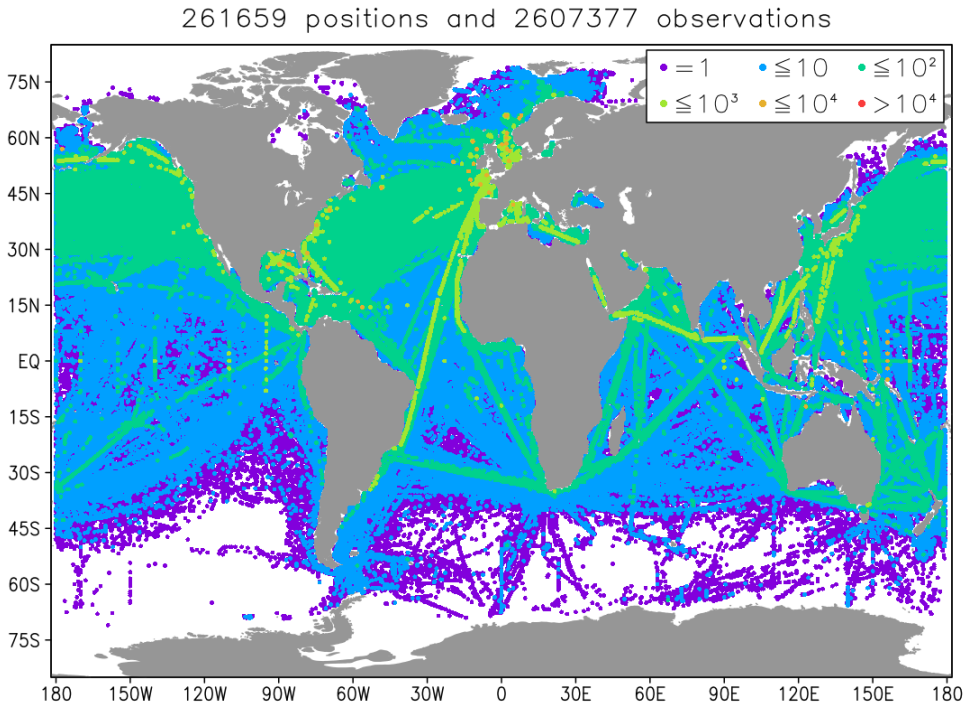

Figure 2: Number of selected ICOADS (Version 3) ship and buoy observations between January 2000 and December 2007 (order of magnitude in color). The two criteria for selection are that a) valid collocations exist with all OHF products and their ensemble and b) all ICOADS variables required to calculate a COARE flux estimate are within 2.8 estimated standard deviations of their respective smoothed monthly climatology. Shown are values at the ¼-degree resolution of the project's reference grid.

3 Determination of Ocean Heat Flux products

3.1 Standardized flux products

Table 1 provides the spatial and temporal resolution characteristics of flux products used in this study. The spatial resolution of products varies from 0.25 to 1 degree and the highest temporal resolution varies from 3 hourly to daily. For further intercomparisons, we interpolated all products onto a standard 0.25 degree grid and at daily time resolution.

Each flux product listed in Table 1 is interpolated onto the same regular 0.25° latitude/longitude grid, using two methods, namely spline interpolation and the modified method of local procedures (Akima 1970). The latter is based on a piecewise function with slopes at the junction points determined locally by a set of polynomials. Both methods are found to be suitable for the re-gridding of flux products and give very similar results. Linear regression slopes between the original and standardized daily LHF product, are 0.99 or higher with the intercepts being lower than 1 W/m². Similar results are found for SHF assessments. The interpolated daily flux products (including bulk atmospheric state variables and heat fluxes) are referred to hereafter as the standardized products.

Table1: Spatial and temporal characteristics of flux products available for OHF project.

	Spatial resolution	Temporal resolution	Period of availability
IFREMER	0.25°×0.25°	Daily 6-hourly	1992 – 2012
HOAPS	0.5°×0.5°	Daily Monthly	1987 - 2008
OAFLux	1°×1°	Daily	1985 - 2014
SEAFLUX	0.25°×0.25°	3-hourly	1998 - 2007
J-OFURO	1°×1° 0.25°×0.25°	Daily Monthly	1988 - 2008
ERA Interim	0.75°×0.75°	6-hourly	1992 - Present
CFSR	0.38°×0.38°	6-hourly	1992 - Present

MERRA	0.50°×0.66°	hourly	1992 - Present

To assess the impact of the space interpolation on the resulting fields, comparisons the standardized and original data are performed. To avoid any further errors related to space and temporal collocation of the original and interpolated data, the interpolation impact is only investigated based on the comparison of original and interpolated statistical distributions. Figure 3 illustrates the distribution comparisons based on the statistical quantiles estimated from original and interpolated LHF data, respectively. Comparisons are shown for IFREMER (Figure 3a), HOAPS (Figure 3b), OAFLUX (Figure 3c), SEAFLUX (Figure 3d), J-OFURO (Figure 3e), and ERA Interim (Figure 3f). The comparisons are performed for data occurring over the global oceans on 3rd of January 2000. These examples indicate that the two kinds of LHF distribution are comparable for most of the variable ranges. As expected, the best agreement is found for interpolated data estimated from products available with 0.25° spatial resolution (Table 1). Slight departures are found for extreme values. The interpolated values tend to be underestimates compared to the original data.

Further controls are performed to assess the quality of the interpolated data. For instance, the original and the standardized LHF and SHF distribution quantiles are calculated for every day in 2000 from global and regional data (high latitudes of the Atlantic ocean (55°N – 65°N), Gulf Stream, the Atlantic tropical zone (15°S-15°N), and the Mediterranean Sea). We found that the slope of linear regression between quantiles from the original and interpolated data varies between 0.96 and 0.99 for LHF and between 0.95 and 0.98 for SHF (not shown). The associated intercepts are lower than 1W/m² for both LHF and SHF.

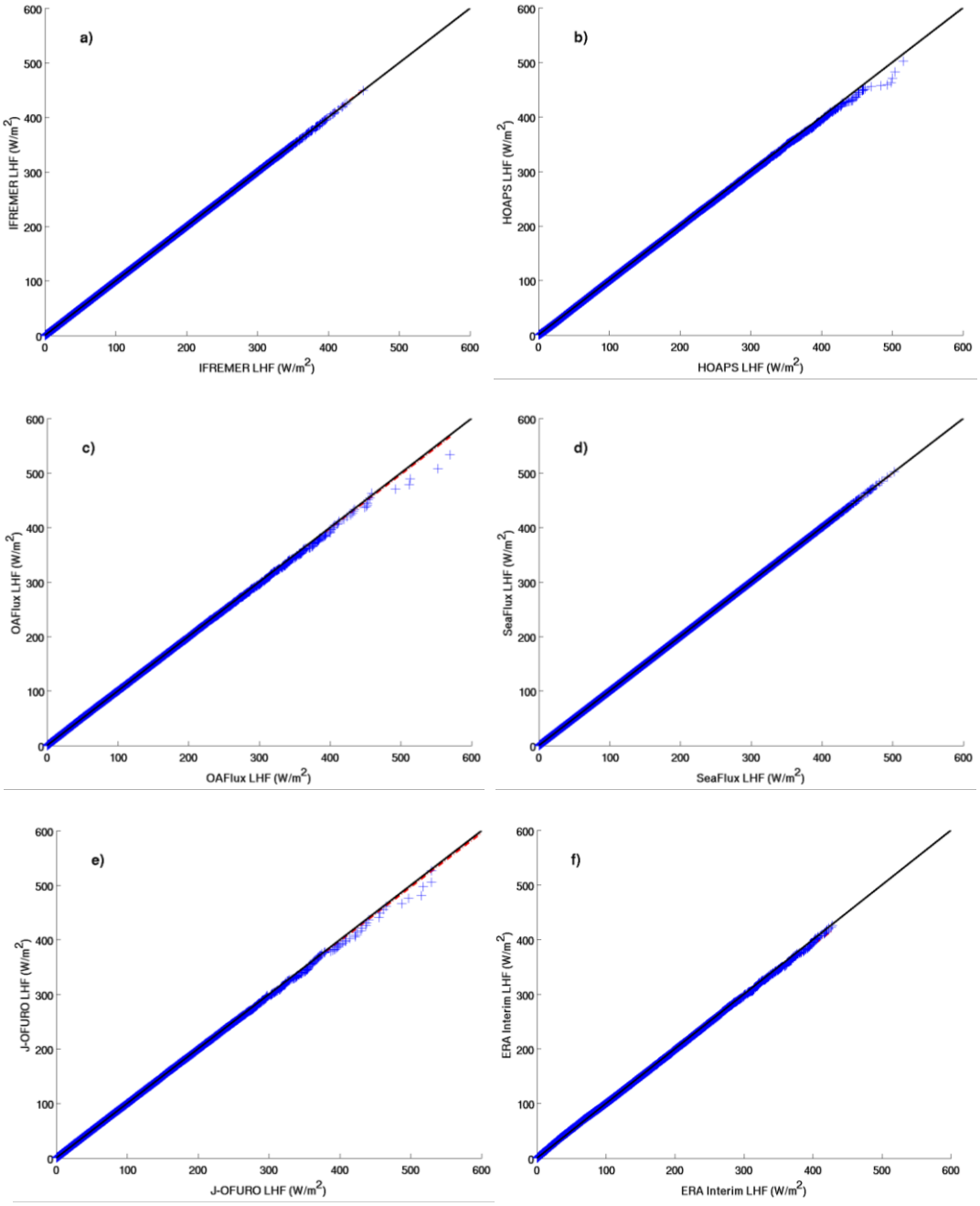

Figure 3 : Comparisons of Original and interpolated LHF from IFREMER (a), HOAPS (b), OAFlux (c), SEAFLUX (d), J-OFURO (e), and ERA Interim (f). They are estimated from global data occurring on 3 January 2000. x- and y- axis indicate the original and standardized data, respectively.

The main spatial and temporal features of the resulting standardized products are also assessed through comparisons with LHF and SHF patterns estimated from original data, as investigated in previous publications (e.g. Grodsky *et al*, 2009; Mestas *et al*, 2013; Smith *et al*, 2011). For instance, Figure 4 (similar to Figure 3 of Smith *et al*, 2011) shows the global spatial distributions of LHF and SHF estimated from the eight standardized daily products averaged over the period 2000 – 2007. All products exhibit similar large LHF spatial patterns (Figure 4, 1st column) mostly characterized by high values localized along western boundary currents (Gulf Stream and Kurishio currents), the southern African zone (Agulhas current), under the subtropical highs, and in the north Indian Ocean. The eight products show that the lowest LHF values are mostly located along the cold tongues in the Atlantic and Pacific equatorial zones, along the main upwelling zones, and at high latitudes. As in previous studies, the main differences between product LHF patterns are seen in the magnitude and are associated with specific air and/or surface humidity issues (e.g. Grodsky *et al*, 2009, Bentamy *et al*, 2013). For SHF (Figure 4, 3rd column), differences are of the same order as those found in (Smith *et al*, 2011). The spatial variability of LHF and SHF (Figure 4, 2nd and 4th columns) is also similar to results found previously from the original data. Although, all products exhibit quite similar LHF and SHF standard deviation patterns, significant magnitude differences are revealed.

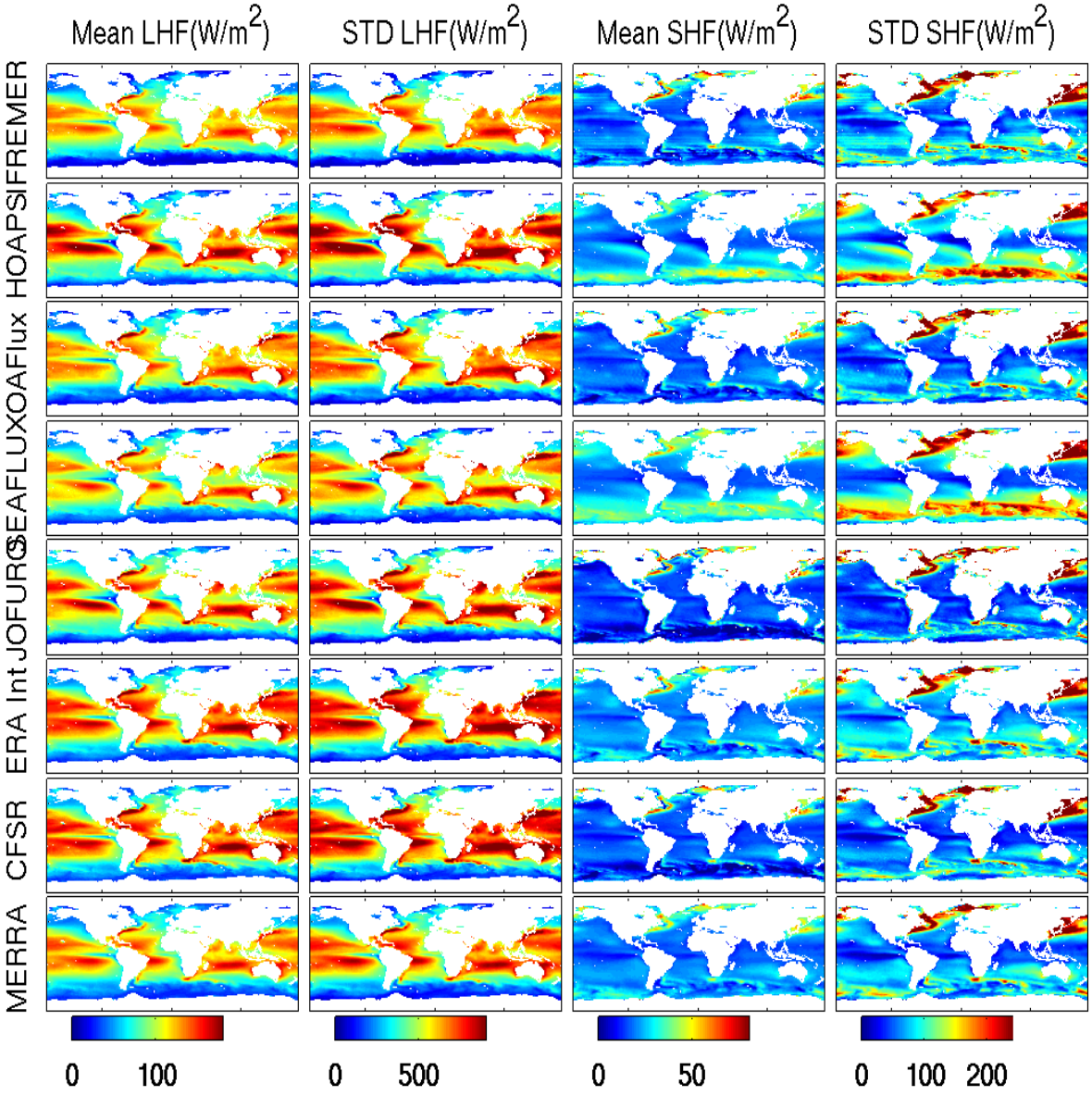

Figure 4 : Mean and associated standard deviation (STD) of LHF (1st and 2nd columns) and SHF (3rd and 4th columns) estimated from the standardized OHF products for the period 2000 – 2007. Panels shown in 1st trough 8 rows are related to IFREMER, HOAPS, OAFLUX, SEAFLUX, J-OFURO, ERA Interim, CFSR, and MERRA, respectively.

3.2 Ensemble flux product

The OHF multiproduct ensemble (OHF/MPE) is estimated based on the use of the standardized IFREMER, HOAPS, OAFLUX, SEAFLUX, J-OFURO, ERA Interim, and

CFSR daily fluxes. It is calculated on a daily basis over the standardized OHF product grid map (0.25°×0.25°) over the ice free global ocean. MERRA-2 is not included in the OHF/MPE ensemble and is kept for further inter-comparison issues.

More specifically, for each day and at each grid point the median and the associated standard deviation of each bulk variable 10m wind speed, specific air humidity, air temperature and surface specific humidity and temperature, are calculated from available and valid U_{10}, q_{a10}, T_{a10}, q_s, and *SST*, derived from the standardized product mentioned above. In order to minimize the impact of outliers, median values are considered as OHF/MPE bulk variable estimates. The latter are used to estimate latent and sensible heat fluxes for each day of 2000 – 2007 period and at each grid 0.25°×0.25° over global oceans. COARE3.0 parameterization is used for OHF/MPE LHF and SHF calculation.

4 Inter-Comparisons

The nine LHF products exhibit quite similar latitudinal variations in the zonal means, averaged over the Atlantic, Indian, and Pacific Oceans (Figure 5). Notably, all standardized LHF products, including MERRA, are within one STD from OHF/MPE. For the three basins, a local minimum in LHF is present near the equator due to the combination of low wind speed and relatively small range of surface humidity departures from saturation. This equatorial minimum is apparent in the Atlantic and Pacific where the cold tongue *SST* is responsible for lower $q_a - q_s$. In general, stronger LHF occurs over warmer *SST* due to the temperature dependence of the saturated humidity. Local maxima of LHF correspond to the combination of relatively warm *SST* and rather strong winds. Such a combination is present in the trade wind belts in all basins. In the Atlantic, all products indicate that the highest LHF are centered around 15°N and 12°S, due to trade winds, and near 38°N due to the combination of high wind speed and large differences between q_s and q_a in the western boundary *SST* frontal

region (e.g. Bentamy *et al*, 2013). Similarly good agreement of the highest LHF is found in the Pacific around 34°N (Kuroshio zone) as well as around 15°N and 18°S (trade wind related maxima). In the Indian Ocean, all nine highest LHF values are observed in the vicinity of 17°S (southeasterly trade wind zone). The largest spread of LHF products is found at latitudes corresponding to the western boundary currents. In particular, the spread reaches 40 W/m² at 38°N in the Atlantic and at 34°N in the Pacific. LHF from IFREMER and ERA Interim tend to be consistently lower or higher than the other products, except in the equatorial area for IFREMER.

SHF zonal means compare well north of 20°S (Figure 5), whereas their spread is larger at more southerly latitudes. This agrees with results found by Smith *et al*. (2011), except for the IFREMER product. In the new version 4 of IFREMER, SHF has been significantly improved in comparison with the previous version 3 and the SHF is now close to the ensemble mean. Some other products (notably J-OFURO, SEAFLUX, and HOAPS) show large SHF variations south of 40°S that are not present in other products. These large variations are indicative of spurious differences in the air-sea temperature difference.

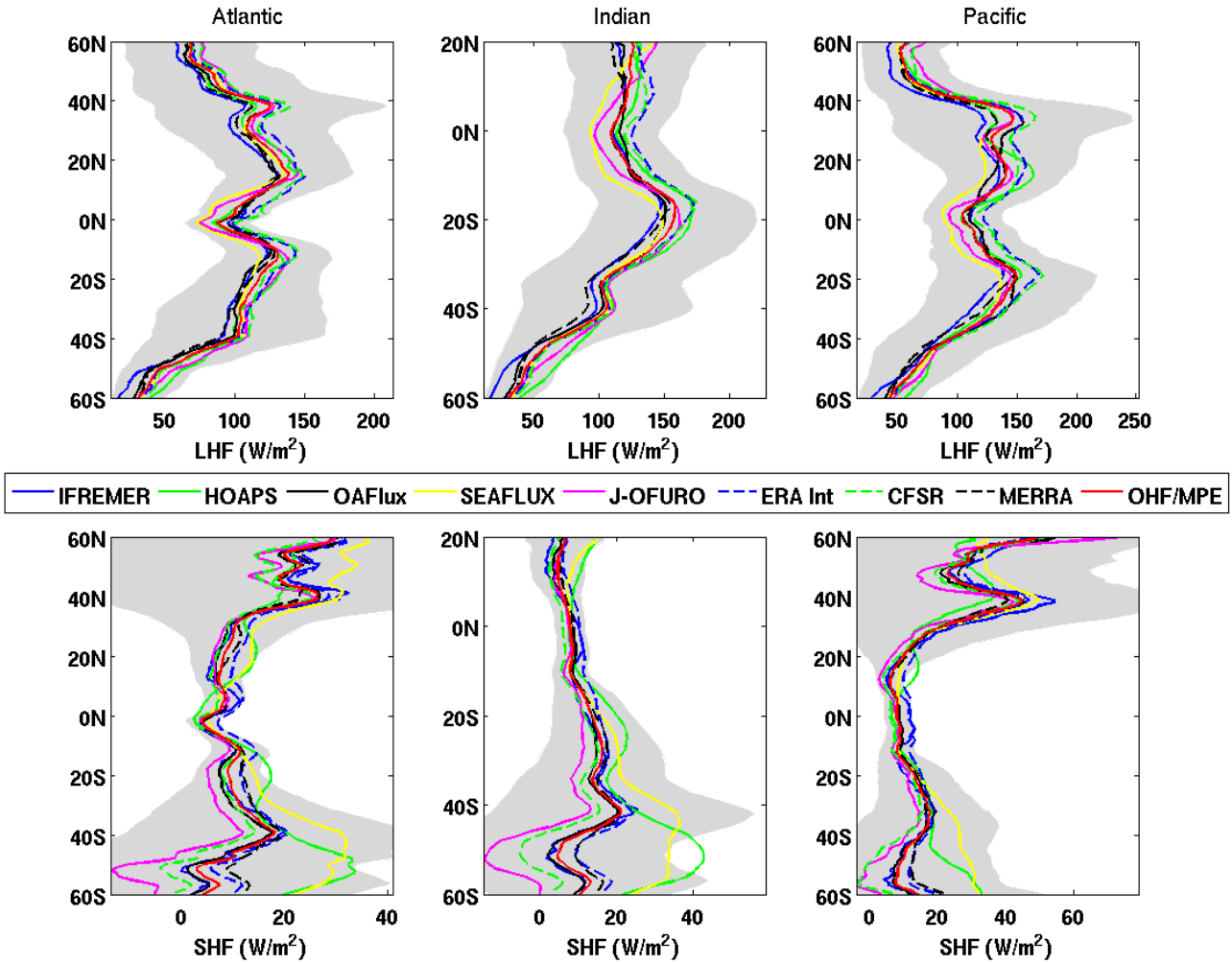

Figure 5 : Latitudinal behaviors of OHF LHF (top) and SHF (bottom) products estimated as averages of daily data occurring over the Atlantic (left), Indian (middle), and Pacific (right) Oceans during the period 2000 – 2007. Shaded area indicates one STD of OHF/MPE LHF and SHF data indicating their longitudinal variability.

5 Product calibration considering correlated errors

Spatial and temporal coverage of the intersection of two (or more) datasets can be orders of magnitude smaller than the coverage of just one gridded dataset. In this sense, collocations only allow one to infer the bias and performance of a full dataset and sometimes such inferences may be lacking (cf., Josey et al, 2014). However, it is common for gridded products to benefit from observations in assimilation windows that are typically as large or larger, than the grid interval on which a true or target variable is represented. Nonlocal, propagated, or shared signal and noise is the norm and inferences based on collocations can be useful. Although some frameworks for assessing bias (e.g., conventional regression and triple collocation) assume independent errors (Stoffelen 1998, McColl et al. 2014), a corresponding framework for slowly varying and well resolved (correlated) error is worth exploring (Su et al. 2014, Gruber et al. 2016). Accommodation of truth/signal (*t*), error/noise (ε), as well as error propagation (λ) is needed, via analyses that incorporate observational error at the time of observation, with decreasing, but roughly symmetric impact at times before and after. There appears to be a family of error models (of which the following is a member) that provide the simplest possible framework for further exploration:

$$
\begin{aligned}
\text{in situ} \quad I &= t + \epsilon_I \\
\text{nowcast} \quad N &= \alpha_N + \beta_N t + \lambda_N \epsilon_I + \epsilon_N \\
\text{forecast} \quad F &= \alpha_F + \beta_F t + \lambda_F(\lambda_N \epsilon_I + \epsilon_N) + \epsilon_F \\
\text{extended forecast} \quad E &= \alpha_E + \beta_E t + \lambda_E(\lambda_F(\lambda_N \epsilon_I + \epsilon_N) + \epsilon_F) + \epsilon_E \\
\text{revcast} \quad R &= \alpha_R + \beta_R t + \lambda_R(\lambda_N \epsilon_I + \epsilon_N) + \epsilon_R \\
\text{extended revcast} \quad S &= \alpha_S + \beta_S t + \lambda_S(\lambda_R(\lambda_N \epsilon_I + \epsilon_N) + \epsilon_R) + \epsilon_S
\end{aligned}
\tag{1}
$$

This error model consists of two heat flux datasets: an in situ estimate and an OHF estimate, where the OHF nowcast is collocated in space and time with the ICOADS in situ estimate and the forecast and revcast are simply samples taken at adjacent locations on the OHF grid (e.g., "persistence" over one or two days is our forecast/revcast method). With an ICOADS heat flux estimate as the calibration reference, each of the OHF samples (*NFERS*) has its own additive and multiplicative bias (α and β). Although product error variance changes under a recalibration to remove bias neither ICOADS error nor true variance (common to both ICOADS and product) changes. As all products are calibrated to the same ICOADS collocations (Figure 2), this permits a separate comparison of truth (and error) across products.

Heat flux analyses assimilate ICOADS observations of *SST*, wind speed, etc. Although they do not assimilate our estimate of heat flux, we accommodate an assimilation of ICOADS information above. Specifically, a parameterization of shared or propagated error into and across an analysis (*NFERS*) is quantified by a retrieval of the λ coefficients. Note that error in the INFERS model employs an AR-1 autoregressive form because this is arguably the simplest. It follows from our application of an AR-1 error model that the minimum number of equations (or samples of the gridded dataset) to match the total number of unknowns is four (*NFER* or *NFRS*). The symmetry of five samples (*NFERS*) simply facilitates a retrieval of the model parameters. Retrieval is done using the full covariance matrix, with the INFERS variance terms and all covariance terms involving *I* and *N* defining all parameters except true variance, σ_t^2, and nowcast multiplicative bias, β_N. As in Danielson et al. (2017), a familiar approximation of β_N is obtained by matching OHF variance to that of ICOADS. This corresponds to the assumption that signal to noise ratio (SNR) is the same for both OHF and ICOADS flux estimates (Su et al. 2014). Numerical estimation of true variance is obtained

from the remaining six terms of the covariance matrix, denoted the autocovariance equations as they involve only the OHF forecast and revcast samples (*FERS*). Given that all retrieved variance is expected to be positive, the locus of minima in the LHS minus the RHS of the autocovariance equations yields the σ_t^2 estimates.

Table 2 provides metrics of calibration (additive and multiplicative bias) and performance (common signal, SNR, and noise) for sensible and latent heat flux for the eight global analyses and their ensemble. Parameters of the INFERS error model are obtained using all collocations from the even years between 2000 and 2007; odd years are retained for validation and yield the same ranking as below. Pairs of numbers refer to the uncalibrated (left) and calibrated (right) parameters. A striking result is that common signal/truth is quite small compared to error/noise for all products, so SNR is uniformly negative. This is the result of accommodating both correlated and uncorrelated error in (1). SNR varies between products mainly owing to the large relative variation in signal and small relative variation in noise.

Recalibration of each flux product involves subtraction of its additive bias and division by multiplicative bias. Recalculation of all metrics yields changes only in product bias and noise, and as expected (cf. Eq. 1), noise varies roughly inversely with multiplicative bias. After recalibration, ICOADS and OHF product noise is the same, by design (as SNR is the same), but common signal is unchanged (i.e., ICOADS and OHF flux estimates are quite different in spite of the OHF recalibration). Ranking by common signal reveals relatively good performance in sensible heat flux by HOAPS, J-Ofuro, and ERA, with good performance in latent heat flux by the MERRA, HOAPS, and the ensemble products. We conclude from this preliminary exercise that error models permitting a direct comparison between observations and the products that employ them is both feasible and instructive, with

the obvious caveat that inferences cannot be made where observations are not available. Clearly, however, there are observations that can be employed to address known regional product biases (cf. Fig. 2 and Josey et al, 2014), where a discussion of local error propagation (λ in Eq. 1) can now also be included.

Table 2. Performance (common signal, SNR, and noise) and nowcast calibration (additive and multiplicative bias) metrics for collocations of sensible and latent heat flux of ICOADS observations and eight global products and their ensemble. Only the ICOADS performance metrics are given as these data are taken to be calibrated already (Eq. 1). All product metrics employ collocations *from even years only* between 2000 and 2007 (odd year averages are retained for validation and are qualitatively the same; not shown). Pairs of numbers refer to pre- and post-calibration (i.e., only nowcast error and bias vary). Signal and noise (as standard deviations) and additive bias are in Wm^{-2} and SNR (Gruber et al. 2016) is in dB.

Product	Common Signal	Common SNR	ICOADS Noise	Product Noise	Product Bias Addit	Product Bias Multi
sensible heat flux						
CFSR	2.58	-18.18	20.94	15.28/20.94	4.89/0.00	0.73/1.00
ERA	**4.96**	**-12.36**	**20.57**	**14.42/20.57**	**9.42/0.00**	**0.70/1.00**
HOAPS	**5.46**	**-11.66**	**20.89**	**16.13/20.89**	**7.71/-0.00**	**0.77/1.00**
Ifremer	0.94	-26.99	20.91	17.26/20.91	5.81/-0.00	0.83/1.00
J-Ofuro	**5.45**	**-11.51**	**20.53**	**13.90/20.53**	**5.50/0.00**	**0.68/1.00**
Merra	3.08	-16.69	21.07	12.03/21.07	7.33/-0.00	0.57/1.00
OAFlux	1.89	-20.94	21.05	14.55/21.05	6.30/-0.00	0.69/1.00
SeaFlux	3.52	-15.57	21.12	14.99/21.12	12.00/-0.00	0.71/1.00
Ensemble	2.11	-19.96	21.01	13.98/21.01	6.93/0.00	0.67/1.00
latent heat flux						
CFSR	18.24	-11.88	71.61	62.20/71.61	27.76/0.00	0.87/1.00
ERA	11.47	-16.10	73.18	61.72/73.18	30.31/0.00	0.84/1.00
HOAPS	**25.35**	**-8.74**	**69.35**	**64.47/69.35**	**13.84/0.00**	**0.93/1.00**

Ifremer	16.08	-12.88	70.87	48.24/70.87	28.73/0.00	0.68/1.00
J-Ofuro	17.34	-12.36	71.90	61.95/71.90	21.04/0.00	0.86/1.00
Merra	**43.81**	**-2.62**	**59.23**	**42.99/59.23**	**26.70/0.00**	**0.73/1.00**
OAFlux	19.05	-11.51	71.71	55.37/71.71	26.15/0.00	0.77/1.00
SeaFlux	17.02	-12.53	71.97	55.35/71.97	23.86/0.00	0.77/1.00
Ensemble	**25.64**	**-8.62**	**69.15**	**52.97/69.15**	**28.22/0.00**	**0.77/1.00**

6 Buoy comparisons

6.1 Statistical results

The statistics aiming at the characterization of comparisons between buoy and flux products (and the associated bulk variables) are determined from collocated buoy (see section 2.9) and product data. Daily fluxes for each product are collocated in space with buoy estimates. The collocation criterion separating buoy and product is that the distance should be less than the product spatial resolution (Table 1). For the standardized products, the spatial criterion is 25km. The statistics are computed for different daily parameters such as 10m wind speed (U_{10}), specific air humidity, sea surface temperature, air temperature, latent heat flux and sensible heat flux. The comparisons of daily satellite (IFREMER, HOAPS, SEAFLUX, and J-OFURO) and buoy data are challenging. Each source type is estimated with a specific temporal sampling that may lead to significant differences between buoy and satellite daily data. For instance, Bentamy et al. (2011) provide a characterization of temporal sampling impact on daily wind estimation. Table 3 shows the results established for the OceanSites LHF and SHF comparisons with remote sensing data for original and standardized flux products. They indicate that comparisons based on the standardized products are very close to those based on the original data. Similar results are found for all buoy networks (not shown). The highest departures in Table 3 are associated with ERA Interim (about 12W/m²),

SEAFLUX (7W/m²), and MERRA (6W/m²). However, these bias values should be considered with caution. Indeed, the same product may have lower or higher biases depending on the mooring used as a reference. For instance, ERA Interim LHF bias estimated versus MFUK buoys is about 6W/m², but the bias is <1W/m² and not significant for NDBC comparisons. Similar inferences could be drawn for almost all products. Links between LHF biases and associated bulk variable (U_{10}, q_a, *SST*, *Ta*) biases (not shown) are not straightforward, for example although U_{10} and q_a biases are higher for IFREMER than for SEAFLUX, the resulting LHF bias is lower for IFREMER (Table 3).

Globally mean SHF bias is generally smaller (<4W/m²) than for LHF, due to a generally smaller magnitude of SHF in comparison with LHF.

The root mean square (RMS) difference values of LHF from buoy data vary between 21W/m² and 51W/m² (Figure 6). All products exhibit high RMS values at NDBC buoys moored in the western Atlantic area off the USA coast in the vicinity of the Gulf Stream. This is the region of maximal LHF variability. Indeed, LHF exceeds 180W/m² in this specific region, whereas globally mean LHF is about 87W/m². The main factor leading to the observed departures is related to the difference between buoys and product specific air humidity along the western boundaries (not shown). One should notice that most NDBC buoys do not provide q_a (or relative humidity). The specific air humidity, q_a , is estimated from air and dew point temperatures using an empirical model. The patterns of LHF difference (Figure 6) indicate that re-analyses (ERA Interim, CFSR, and MERRA) exhibit lower RMS values in comparison with satellite or synthesis products. Such results could be associated with the assimilation of buoy measurements into these reanalyses, which would make the current comparisons not truly independent. Previous studies assess the assimilation impact into numerical models (e.g. Josey *et al*, 2014). Most RMS values for IFREMER and OAFLUX, excluding those for the western Atlantic zone, are lower than 30W/m². As expected the lowest and highest RMS are

found in northern and tropical basins, respectively, in agreement with known transient LHF patterns in the tropics (Grodsky *et al*, 2009). However, both products have high RMS differences at buoys located off the Japanese coast (OceanSites KEO buoys). They are, for instance, about 55W/m² for IFREMER and 45W/m² for OAFLUX. At these specific extra-tropical locations, bulk variables experience large temporal variation related to synoptic-scale weather systems, and consequently lead to high variability in turbulent fluxes. The mean and STD values of the daily LHF time series for the OceanSites buoy located at 32°N; 145°E are about 156W/m² and 113W/m², respectively. Such high variability is not found for tropical buoys experiencing similarly high time mean LHF values (exceeding 150W/m²) such as buoys located at 15°N; 90°E, and 10°S; 10°W. The LHF STD values at these two locations are about 50W/m², and 38W/m², respectively. The main sources of KEO LHF variability, and therefore of discrepancies between buoy and OHF products, are the high variabilities of (q_s - q_a) and/or U_{10}. Indeed, high daily variability of ($q_s - q_a$) and/or U_{10} (estimated as STD from hourly buoy data) leads to high differences between KEO buoy and product LHF estimates. Such results highlight the OHF product errors associated with temporal sampling. Furthermore, it is found (not shown) that most of LHF daily maxima are underestimated by the products leading to an enhancement of root mean square.

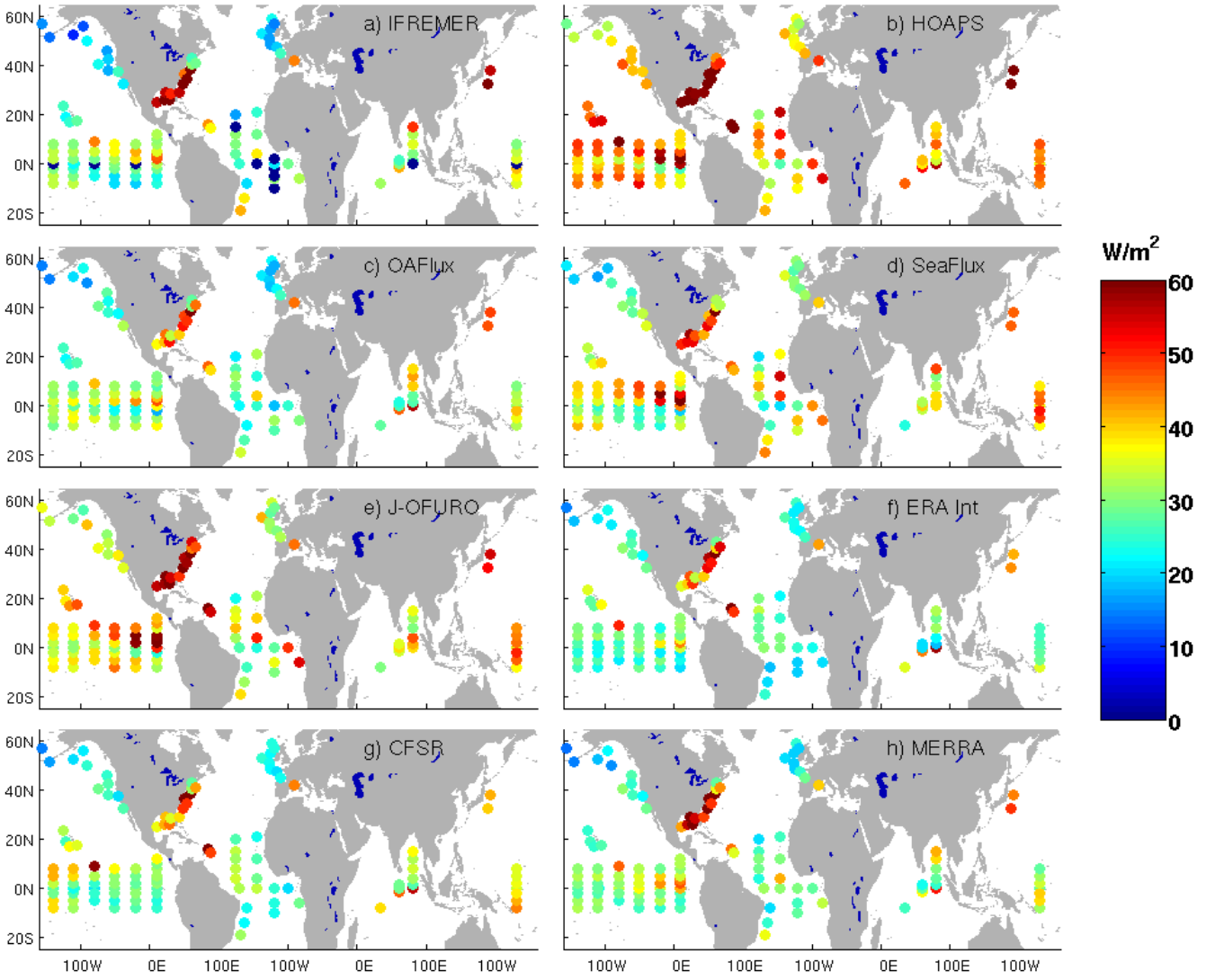

Figure 6 : Root Mean Square difference between daily buoy and OHF LHF (in W/m^2) standardized products estimated for the period 2000 – 2007. Details on the buoy measurements are introduced in section 2.9.

As expected, RMS SHF differences estimated at each buoy location (Figure 7) are lower than those found for LHF. They are lower than 10W/m² and may not exceed 5W/m² at most tropical locations. The highest RMS SHF differences are drawn from NDBC comparisons. These locations, moored along northeast off US coast, experience high *SST* and T_a differences as well as high wind conditions. Furthermore, the enhancement of RMS values would be due to a mismatch in *SST* derived from a buoy (local measurement) and from the gridded product. The lowest values are found for ERA Interim and CFSR estimates.

Table3: Statistics characterizing the comparisons between OceanSites and satellite product (original and standardized) daily fluxes. They are calculated for latent heat flux (LHF in W/m²), and sensible heat flux (SHF in W/m²). Statistics are calculated for each product when available during the associated time interval of the 1999-2009 period. Statistics relied on standardized products are calculated for the 1999 – 2007 period

Statistics	Products	LHF		SHF	
		Original	Standardized	Original	Standardized
Bias (W/m²)	IFREMER	-2.20	-3.34	0.09	-0.63
	HOAPS	-5.25	-3.76	-1.27	-0.64
	OAFLUX	4.26	1.76	1.31	0.65
	SEAFLUX	7.63	7.76	-1.93	-1.76
	J-OFURO	1.29	0.98	2.27	1.86
	ERA Interim	-12.01	-13.67	-2.42	-2.14
	CFSR	-0.12	-0.28	0.32	1.06
	MERRA	6.76	6.46	-0.12	-0.34
RMS (W/m²)	IFREMER	30.03	28.16	7.16	6.75
	HOAPS	42.21	41.21	9.63	9.63
	OAFLUX	31.49	29.42	5.49	4.96
	SEAFLUX	30.93	30.16	6.73	6.47
	J-OFURO	36.31	34.67	7.19	6.56
	ERA Interim	27.34	27.00	5.55	5.11
	CFSR	26.12	25.14	4.77	4.56
	MERRA	26.37	25.56	4.86	4.74
Correlation	IFREMER	0.87	0.88	0.91	0.90
	HOAPS	0.81	0.83	0.75	0.78
	OAFLUX	0.87	0.89	0.92	0.92
	SEAFLUX	0.88	0.88	0.88	0.88

	J-OFURO	0.85	0.86	0.86	0.87
	ERA Interim	0.90	0.90	0.92	0.93
	CFSR	0.90	0.90	0.91	0.91
	MERRA	0.88	0.88	0.88	0.87
Symmetrical regression coefficient	IFREMER	0.83	0.83	1.15	1.14
	HOAPS	1.12	1.06	1.14	1.08
	OAFLUX	0.88	0.86	0.95	0.99
	SEAFLUX	0.94	0.94	0.98	0.99
	J-OFURO	1.02	1.01	0.96	0.99
	ERA Interim	0.94	0.92	0.99	0.99
	CFSR	0.99	0.99	1.07	1.08
	MERRA	0.78	0.78	0.96	0.96

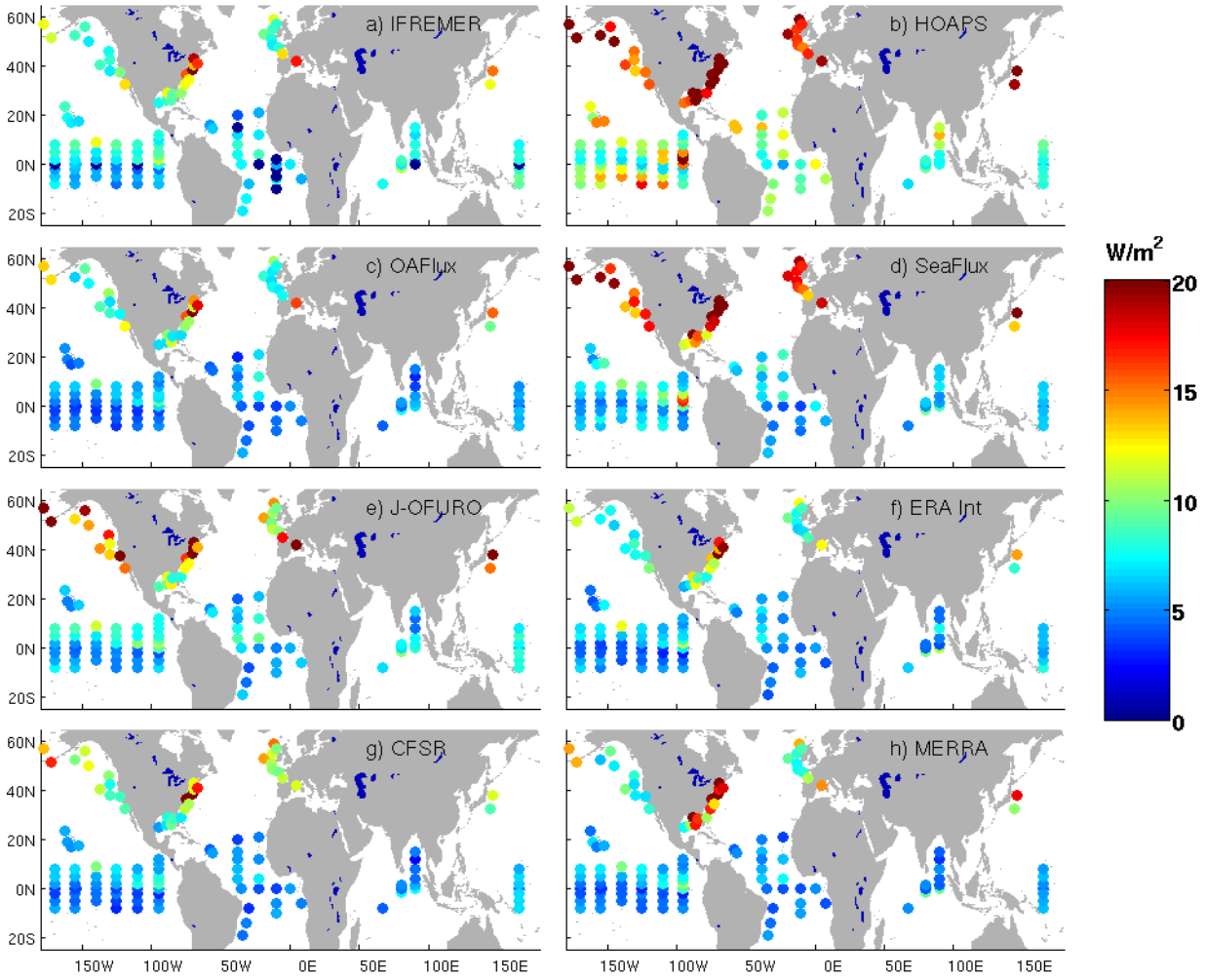

Figure 7 : Root Mean Square difference between daily buoy and OHF SHF standardized products estimated for the period 2000 – 2007.

Similar statistics are estimated from collocated buoy and the OHF/MPE flux data. For instance, the results found for OceanSites comparisons are summarized in Figure 8 showing Taylor diagrams (Taylor, 2001) of LHF and SHF. They indicate the standard deviation (STD) of LHF or SHF derived from OceanSites buoys and from the standardized and OHF/MPE products, the correlation coefficient (ρ), and the root mean square difference (RMSD) between the buoy and each product. For instance, for a buoy (used as reference) STD, ρ, and RMSD associated with LHF patterns (resp. to SHF) are of 70W/m² (resp. 20W/m²), 1 (resp. 1), and 0 (resp. 0), respectively. Both diagrams show that the OHF/MPE LHF and SHF have

the best comparison with the buoy data compared to other products (Figure 8). LHF and SHF STD are of 66W/m² and 20W/m², respectively. Furthermore, OHF/MPE exhibits the highest correlations (about 0.95) and the lowest RSMD values (24W/m² for LHF and 5W/m² for SHF).

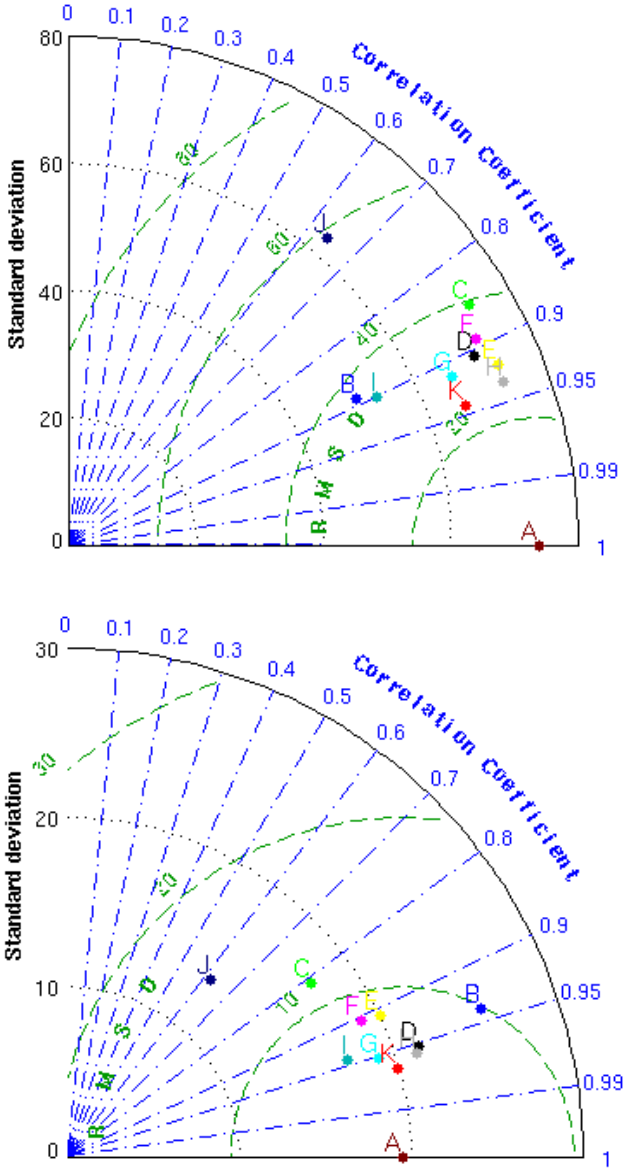

Figure 8 : Taylor diagram summarizing intercomparisons between daily OceanSites buoy and OHF LHF (top) and SHF (bottom) for the period 2000 - 2007

Statistics for the ensemble mean OHF/MPE product are also evaluated for each buoy network separately and should be compared to those obtained for the standardized products (Table 3, Figures 6 and 7). For instance, OHF/MPE RMSD generally tend to be lower than those for the standardized products (except IFREMER for MFUK, IFREMER and OAFLUX for NDBC, and ERA Interim for OceanSites and Tropical networks). One should notice that OHF/MPE always improves statistics in comparison with the MERRA, which is not used for the OHF/MPE determination. Characterization of differences between OceanSites data and OHF data is illustrated in Figures 9 and 10 for LHF and SHF, respectively. They show RMS differences at selected moorings for the nine standardized products and the OHF/MPE. The latter leads to significant decrease of RMS differences for LHF and SHF in comparison with corresponding values obtained for the standardized products. Indeed, OHF/MPE RMS differences for LHF and SHF are lower than 30W/m² and 6W/m², respectively, at all locations, except at KEO (32°N, 145°E), and at (15N 90E) where SHF bias is marginally greater that 6W/m². One interesting result that could be drawn from Figures 9 and 10, is that the OHF/MPE is not dominated by any one product. Instead, all individual products contribute to the determination and thus to the accuracy of the OHF/MPE.

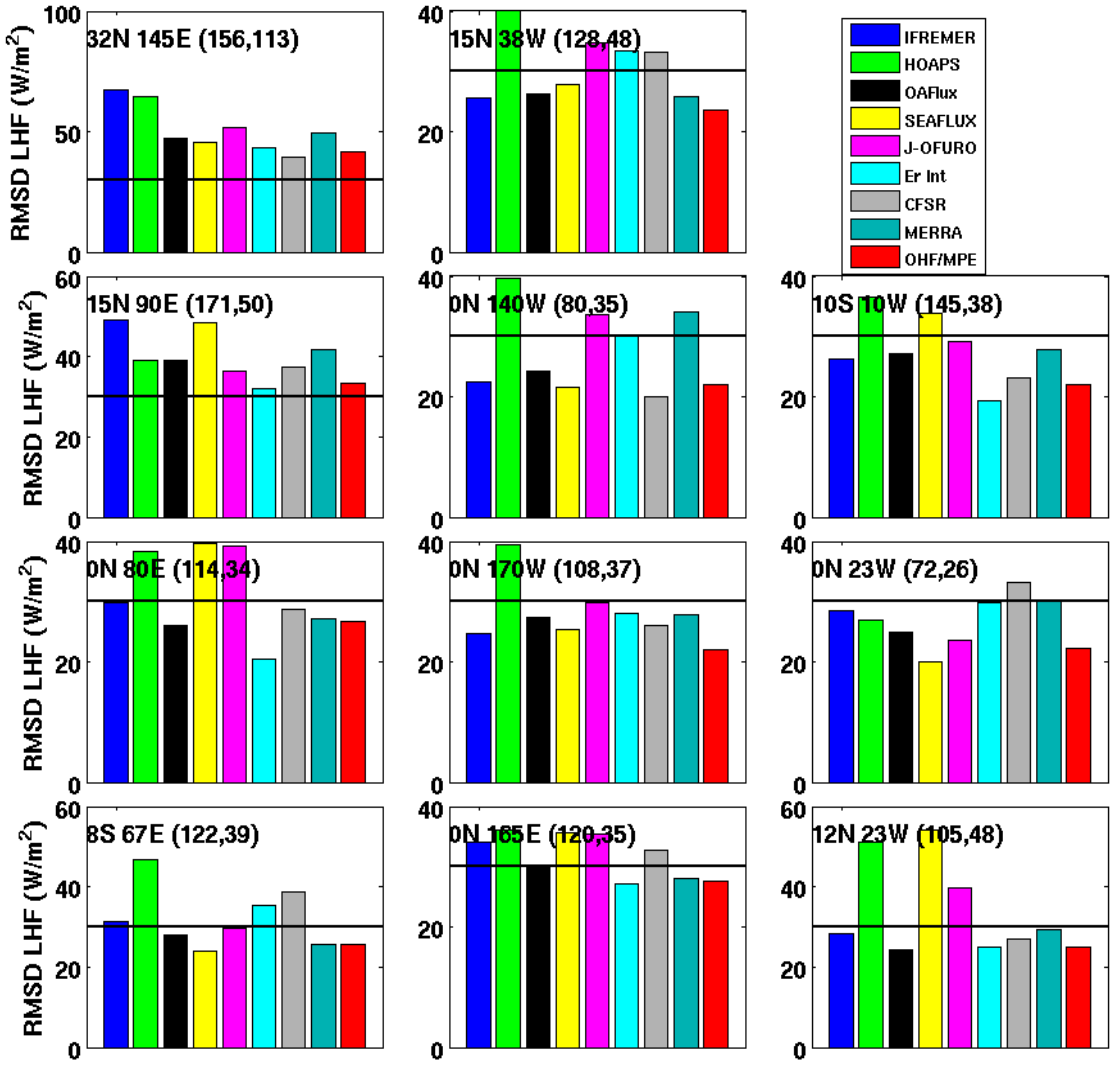

Figure 9 : LHF root mean square difference (RMSD) between individual selected OceanSites buoy and each OHF product. Buoy coordinates are provided at the top of each panel. The tow numbers within brackets in each panel indicate LHF buoy mean and standard deviation values (in W/m²), respectively. The horizontal black line indicates 30W/m². Statistics are estimated from all available collocated data during the 2000 – 2007 period.

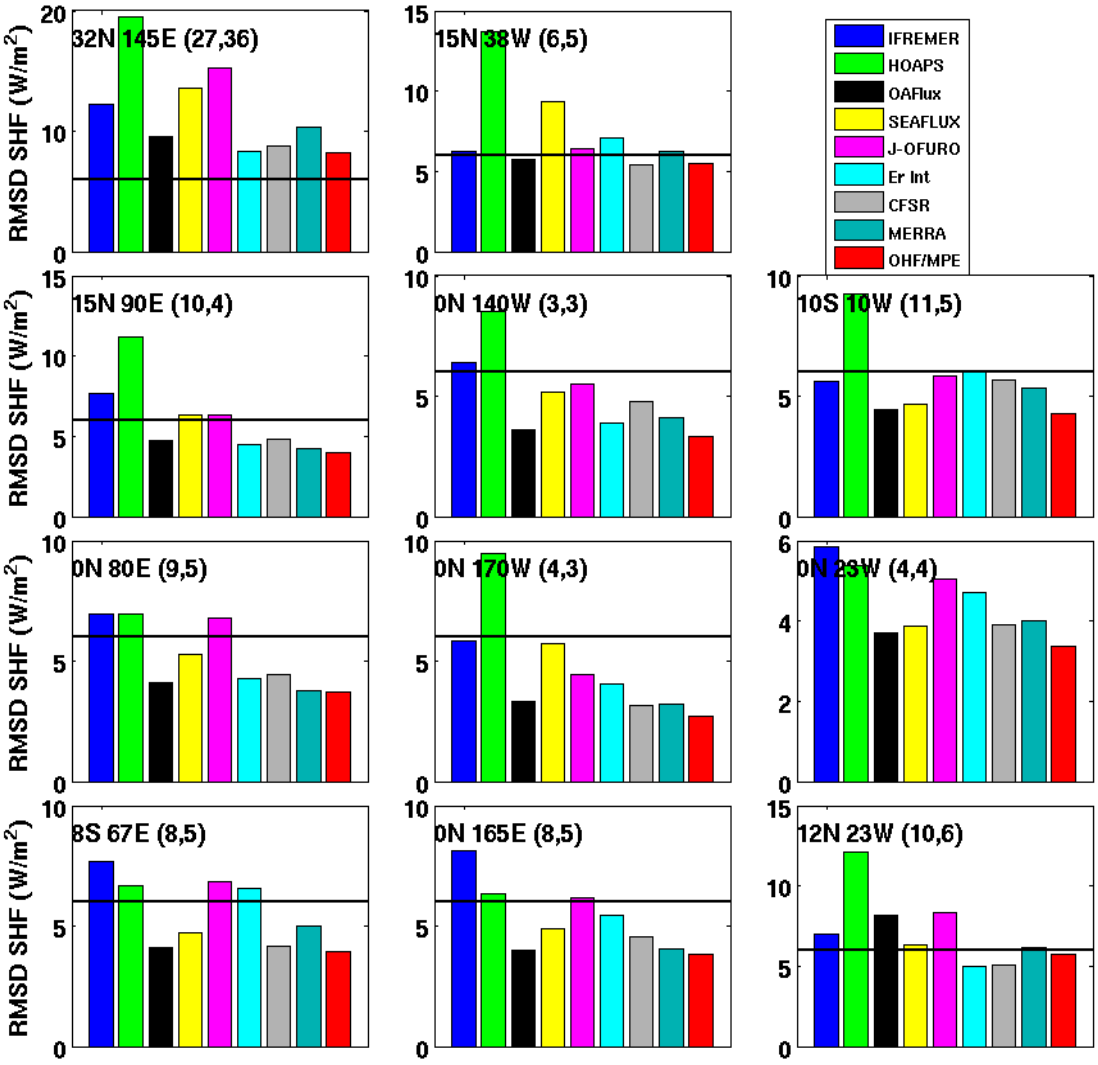

Figure 10 : SHF root mean square difference (RMSD) between individual selected OceanSites buoy and each OHF product. Buoy coordinates are provided at the top of each panel. The tow numbers within brackets in each panel indicate SHF buoy mean and standard deviation values (in W/m²), respectively. The horizontal black line indicates 6W/m². Statistics are estimated from all available collocated data during the 2000 – 2007 period.

7 Ensemble versus Standardized Products

The results of the buoy data comparisons indicate that OHF/MPE is more accurate than any of the contributing products. Hence, it is now employed for the characterization of the spatial and temporal errors of each standardized product.

The evaluation is first performed over global oceans for the period 2000 through 2007. Mean and the associated STD characterizing the difference between OHF/MPE and each product (in this order) are shown in Figure 11 for LHF and SHF. About 95% of LHF (*resp*. SHF) mean biases are within -5W/m² and 18W/m² (resp. -7W/m² and 4W/m²) for IFREMER, -22W/m² and 10W/m² (-26W/m² and 8W/m²) for HOAPS, -7W/m² and 14W/m² (-2W/m² and 4W/m²) for OAFLUX, -5W/m² and 21W/m² (-24W/m² and 2W/m²) for SEAFLUX, -16W/m² and 13W/m² (-5W/m² and 16W/m²) for J-OFURO, -19W/m² and 1W/m² (-7W/m² and 1W/m²) for ERA Interim, -20W/m² and 1W/m² (-1W/m² and 10W/m²) for CFSR, and -6W/m² and 15W/m² (-10W/m² and 3W/m²) for MERRA. Although all products show similar large scale spatial patterns (Figure 4), their differences versus OHF/MPE (Figure 11) lead to different and significant spatial distributions of mean biases and STD for LHF and SHF. IFREMER LHF error patterns found along the Atlantic and Pacific western boundary currents rely mainly on the specific air humidity being wetter (Bentamy *et al*, 2013) and along equatorial areas due to higher surface winds and dryer q_a. LHF HOAPS patterns meet those shown in (Anderson *et al*, 2010) and obtained from comparison with in-situ LHF estimates. HOAPS LHF tends to be higher along the tropical and southern oceans. Both LHF OAFLUX and J-OFURO exhibit patterns related to surface wind distributions, but with opposite signs. LHF patterns derived for SEAFLUX tends to be highly correlated with q_a spatial patterns. Indeed, the highest q_a values are localized in the tropics, being wetter in SEAFLUX. ERA

Interim and CFSR LHF exhibit systematic biases versus OHF/MPE over most oceanic regions. Their differences are largest in the tropical and sub-tropical regions. MERRA, not used in OHF/MPE determination, shows LHF bias and STD quite similar to those obtained for IFREMER.

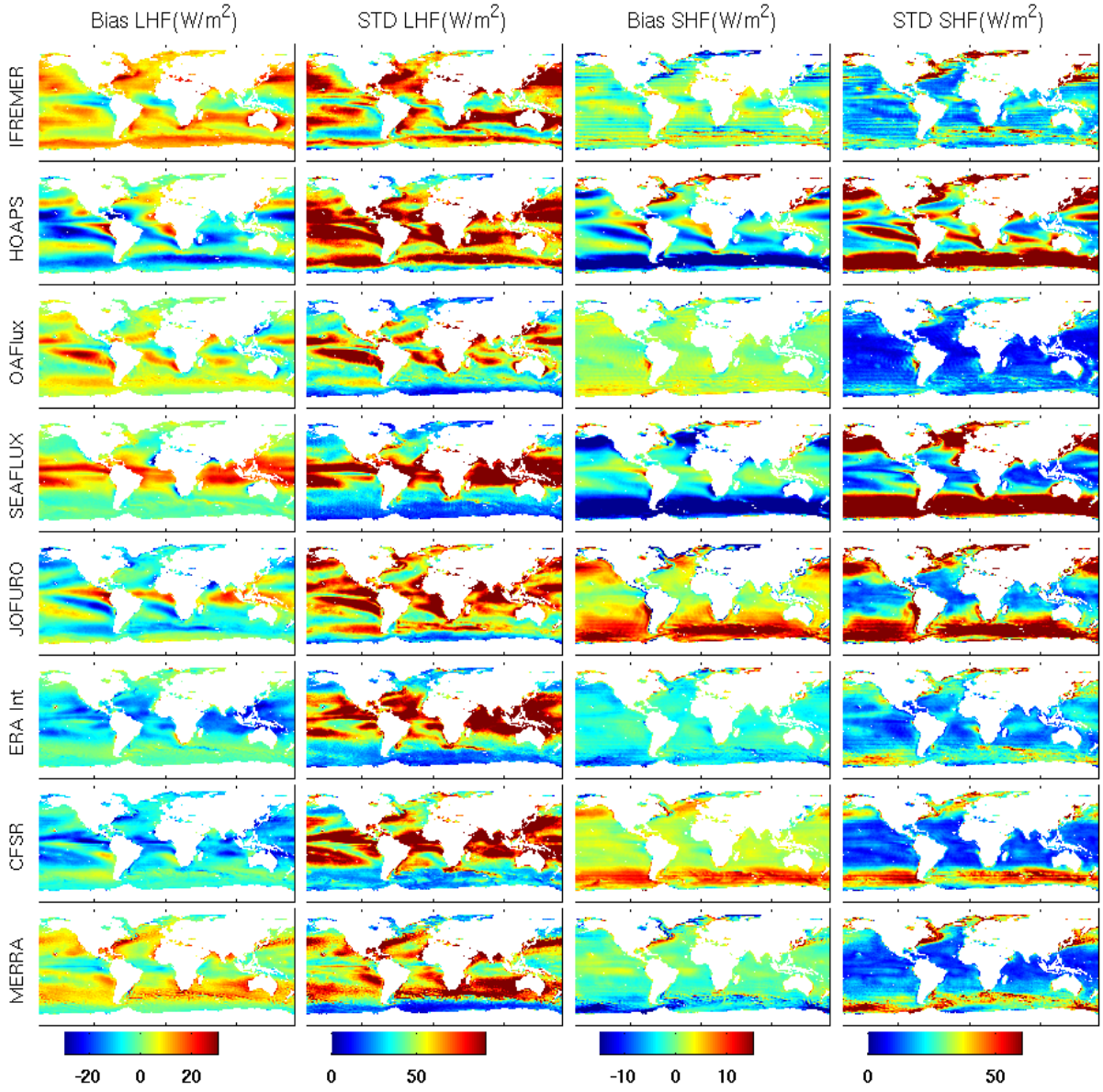

Figure 11 : Spatial distributions of mean bias (first and third column) and STD (second and fourth column) in LHF and SHF difference between OHF/MPE and each standardized product (rows 1 through 8). They are estimated for the 2000 – 2007 period.

Overall, SEAFLUX SHF has larger bias and STD in mid and high latitudes, whereas OAFLUX and ERA Interim both show quite small biases, but with different signs, and quite similar STDs. MERRA SHF and STD features show fairly good agreement with OHF/MPE.

Differences between OHF/MPE and each individual product are shown as monthly mean averaged over the global ocean and over four selected Atlantic latitude bands (Figure 12). Over the global oceans most products show a smooth decrease or increase over time. For instance, IFREMER LHF bias changes slightly over time, decreasing from 9.47W/m² to 4.77W/m². Time variation in J-OFURO is more pronounced with a factor of about 6 between beginning and end of study period. Time changes of LHF biases relied on LHF time changes. For instance, IFREMER LHF bias is found associated with a positive trend characterizing IFREMER LHF time features (not shown). A similar positive trend is also found from OHF/MPE (not shown). IFREMER LHF biases estimated over (40°N-60°N) and (20°N-40°N) latitude bands, both exhibit a seasonal signal mainly related both to atmospheric and oceanic seasonal features. High bias values are seen in (20°N–40°N) mainly due to the seasonal changes of LHF occurring over western boundary currents eg. the Gulf Stream. The IFREMER LHF bias shows a peak during the winter season when LHF is maximized and reduces to small values in summer (<1W/m²). Similar seasonal features are seen in all products, but with lower magnitudes, except for J-OFURO. In the tropical zone (20°S-20°N) SEAFLUX and re-analysis models (ERA Interim and CFSR) LHF are consistently lower and higher, respectively. For SHF time series, the main departures are found for SEAFLUX and HOAPS which both exhibit increasing negative biases, and for J-OFURO and CFSR which are consistently biased lower than OHF/MPE. Significant departures are also found for SEAFLUX over the sub-tropical area (20°N-40°N), and for HOAPS and ERA Interim over the tropical zone.

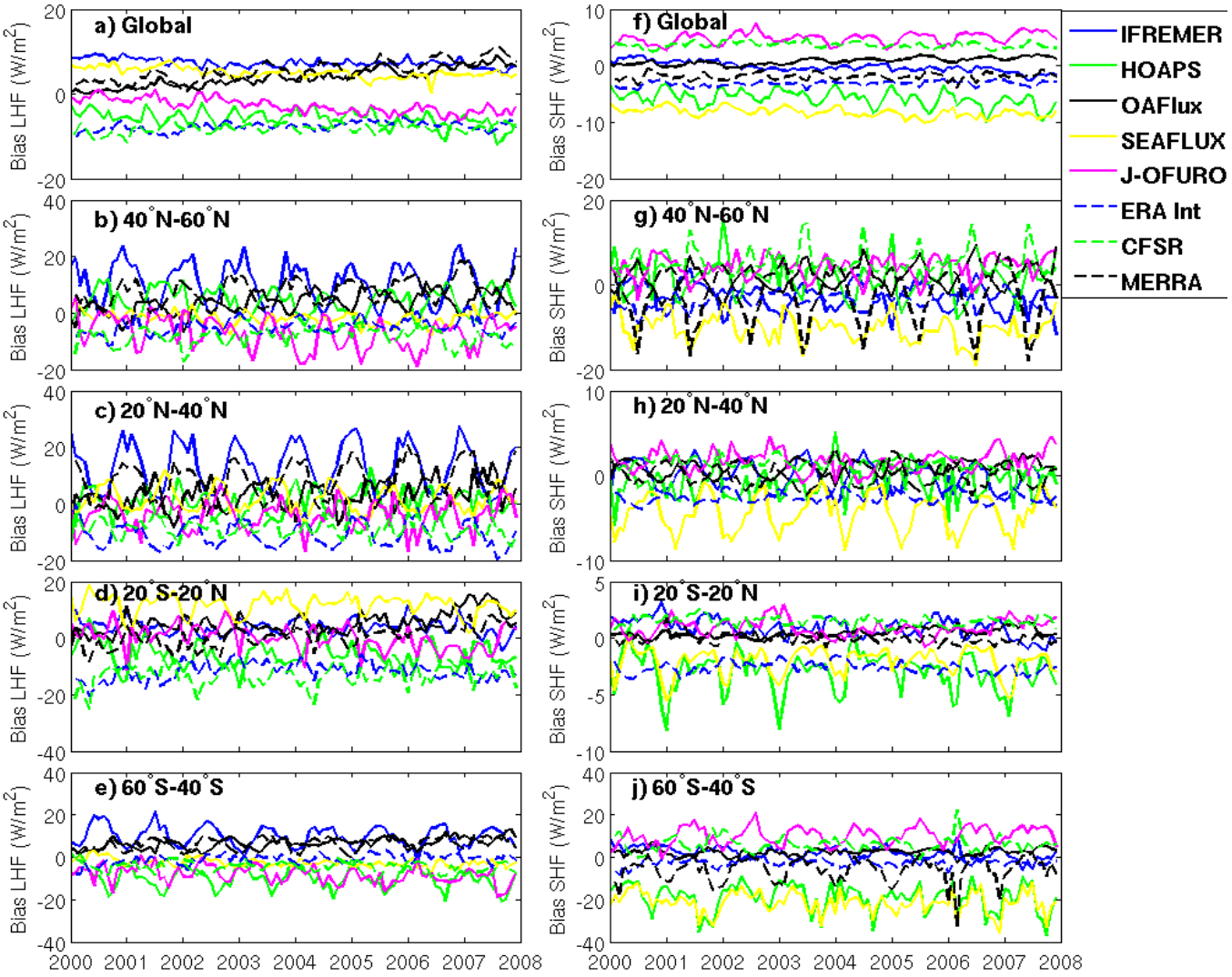

Figure 12 : Time series of monthly-averaged differences between LHF OHF/MPE and standardized products. The biases are shown for LHF (left panels) and SHF (right panels). They are calculated for global oceans (a) and f)) and for four Atlantic latitude bands 40°N-60°N (b) and g)), 20°N-40°N (c) and h)), 20°S-20°N (d) and i)), and 60°S-40°S (e) and j)) .

7 Probability distribution results

For the validation of different surface turbulent flux products against buoy measurements we also applied an approach developed by Gulev and Belyaev (2012) focused on the analysis of probability distributions of surface turbulent fluxes. In this approach,

probability distributions of surface turbulent heat fluxes are approximated by the 2-parameteric MFT (Modified Fisher-Tippett) distribution which allows for the analysis of the probability density functions (PDFs) and high percentiles of surface fluxes. Details of the derivation of the MFT PDF are presented in Gulev and Belyaev (2012). In this study, application of this framework allows for identification of flux products which demonstrate significant differences in the PDFs with the buoys. For instance, it is possible to identify those products which are comparable with each other in terms of mean values but demonstrate significant differences in surface flux extremes. Satellite products may have problems to effectively capture extreme surface fluxes due to representativeness error inherent in the procedure of pre-processing of satellite products and the impact of this can be effectively studied by analyzing PDFs of surface fluxes.

To apply the analysis of PDFs to the OHF products and buoy data we used co-located time series of surface fluxes for the 9 satellite and reanalysis products (including the OHF/MPE ensemble product) and buoy time series. Here we present the results for the Northern Hemisphere winter season (JFM) when high frequency variability of turbulent fluxes is the strongest and, thus, surface flux extremes are most pronounced. Co-location of the gridded products with buoy time series at daily resolution was applied as in Section 6.1. Because the computation of PDF parameters implies quite strict sampling requirements, for this analysis we used only buoys providing sufficient data; specifically we required that at least 2 winter seasons were present in the buoy record and each seasonal record has at least 10 daily values of surface fluxes. This resulted in elimination from the analysis of a considerable subset of the buoy array used in the previous sections. Thus we used 11 of 12 MFUK buoys, 29 of 96 NDBC buoys, 62 of 68 TAO buoy and 7 of 13 PIRATA buoys. Also most of the OceanSites buoys, specifically in the Indian Ocean, were excluded from the analysis. Estimation of the MFT parameters and derivation of PDFs from daily time series was similar

to that in Gulev and Belyaev (2012) for 6-hourly data. According to Gulev and Belyaev (2012) the 2-parameter MFT PDF is given by:

$$P(x) = \alpha\beta \exp(-\beta x) \exp(-\alpha \exp(-\beta x)), \quad (2)$$

Where probability density function $P(x)$, x being turbulent flux, is modelled by the non-dimensional location parameter α and the dimensional scale parameter β. Of these, β controls the squeeze of the MFT distribution and α determines the modal value of the distribution (under fixed β). In most locations, except for a few tropical locations, goodness of fit of the MFT PDF was higher than 95%, according to both the K-S test and the Michael's test. Figure 13 shows MFT PDFs for several buoy locations over the global ocean. The subpolar location in the Bering Sea (Figure 13a) is an example when buoy latent heat flux is smaller on average compared to most satellite and reanalysis products. Nevertheless, it demonstrates much stronger extreme values, with PDF characterized by a heavier tail than the global products. Here the strongest winter mean flux of 70 W/m^2 revealed by HOAPS exceeds the buoy mean by about 20 W/m^2. At the same time flux values corresponding to 99.9th percentile give 306 and 265 W/m^2, respectively for buoy and HOAPS. This reflects strong synoptic and mesoscale atmospheric variability in the subpolar regions, which is not reproduced by reanalyses or satellite products, but is well captured by buoy measurements.

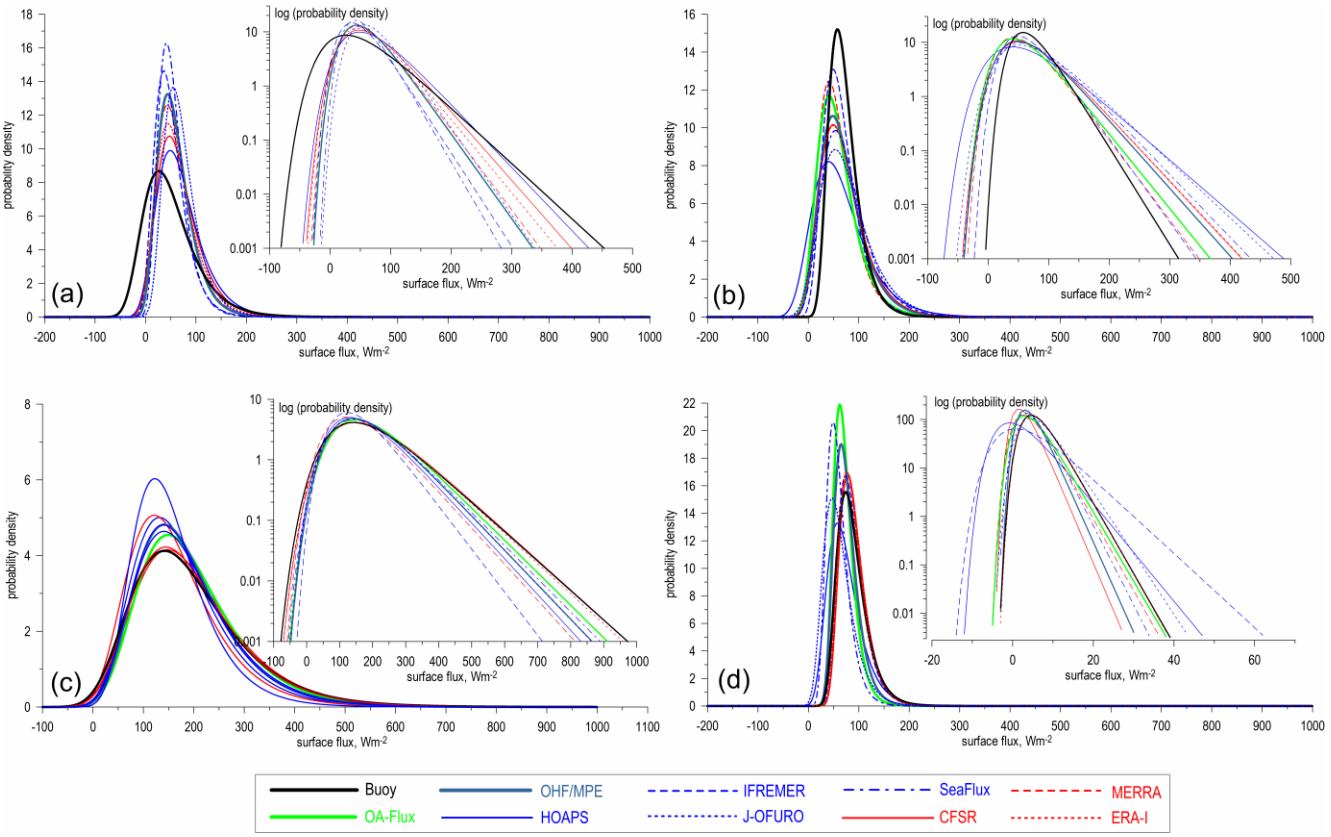

Figure 13: Examples of MFT PDF fitted to the winter turbulent heat fluxes revealed by different OHF products for the four buoy locations: 177.58° W, 57.05° N (a), 130.00° W, 38.03° N (b), 78.48° W, 28.95° N (c) and 95.00° W, -5.00° N (d). Inlays show PDFs in log-scale (for probability density) in order to better indicate differences in the tails of distributions.

The California buoy (Figure 13b) demonstrates the opposite effect: Buoy seasonal mean flux was considerably stronger compared to all other products (72 vs 65 W/m^2 for buoy and HOAPS respectively), while for the 99.9th percentile the buoy shows a latent flux value of 220 W/m^2 that is about 170 W/m^2 smaller than HOAPS and also smaller than all other products with the differences from 20 (for MERRA and IFREMER) to 150 (for J-OFURO) W/m^2. Importantly, this change in the PDF cannot be attributed to inadequate sampling in the buoy data, as all co-located products were sub-sampled in time according to the buoy data sampling characteristics (see section 6.1). Inadequate sampling may indeed impact on PDF characteristics (see, e.g. Gulev et al. 2007a,b; Gulev and Belyaev 2012), however for this purpose non co-located data sets should be considered. According to Gulev et al. (2007a,b)

sampling errors in turbulent fluxes derived from VOS data in poorly sampled areas may amount to several tens W/m^2, with about 60% of this being attributed to the random sampling error. Gulev and Belyaev (2012) demonstrated that sampling error in VOS based fluxes seriously affects extreme fluxes, and to a lesser extent mean values. In our study sampling may affect PDFs of fluxes derived from satellite products with the impact being two-fold. First, there are very few missed daily grid values in some satellite products, however this effect does not result in significant error with respect to the fully sampled satellite product. Secondly, even when daily grid data are provided in satellite products, these gridded values may be affected by interpolation procedures and gap-filling algorithms employed in every product (see Section 2). In this respect our further analysis of PDFs is focused on quantifying this effect. The satellite products, and to a lesser extent the reanalyses, reveal stronger flux extremes even when the mean values in these products are smaller than those revealed by the buoy data. The Gulf Stream area (Figure 13c) gives an example of generally consistent differences between the buoy data and the different products for both the means and the high order percentiles. In this area the OAFLUX product performs remarkably better than in the other regions in capturing extreme fluxes. In the tropical location (Figure 13d) the buoy shows the mean value very close to the CFSR and somewhat smaller compared to the HOAPS. At the 99.9th percentile the buoy and CFSR values remains consistent with the relation between the means, which is not the case for the HOAPS showing 18 W/m2 stronger extreme fluxes than the buoy.

We have to note here, that for this study original buoy data available for most locations at 10-minute resolution were averaged to obtain daily time series consistent with a temporal resolution of satellite products. Similarly, reanalyses time series were also converted to daily values while their original resolution varies from 1 to 6 hours. This fits to our focus on validation of satellite-based fluxes, developed at daily resolution in most products. However,

this puts aside the analysis of sub-daily variability in turbulent fluxes, which is quite large and may seriously affect probability distributions. Thus, monthly maxima in the latent heat flux derived from the original 10-minute data may exceed estimates derived from daily data by about 50-70% in the tropics and several times in subpolar latitudes. Further analysis of this issue relates to the temporal scaling problem for surface fluxes, which was addressed for synoptic scales (see, e.g. Gulev 1994, Wu et al. 2016), but still requires understanding in the minute-to-hourly sub-range. For this problem buoy data (and other high resolution in-situ measurements are very useful, which is not the case for most satellite datasets (at least with the present performance).

The inconsistencies in representation of mean and extreme fluxes in different products is illustrated in Figure 14 showing the 2-dimensional diagram of the MFT PDF in the coordinates of the distribution parameters of which α is a location parameter and β is the scale parameter (eq. 2). In this space mean and extreme fluxes may have different relations as shown by Gulev and Belyaev (2012). For the location in the Bering Sea the buoy flux is associated with relatively small values of α and β, implying strong decrease in the mean value, but holding quite strong extreme values, exceeding those for all satellite and reanalysis products whose means were larger than reported by the buoy. For the location of the California buoy, on the other hand, the buoy fluxes are associated with high values of α and β, implying relatively high mean values (exceeding or comparable with the means reported by the other products), but much weaker extreme fluxes.

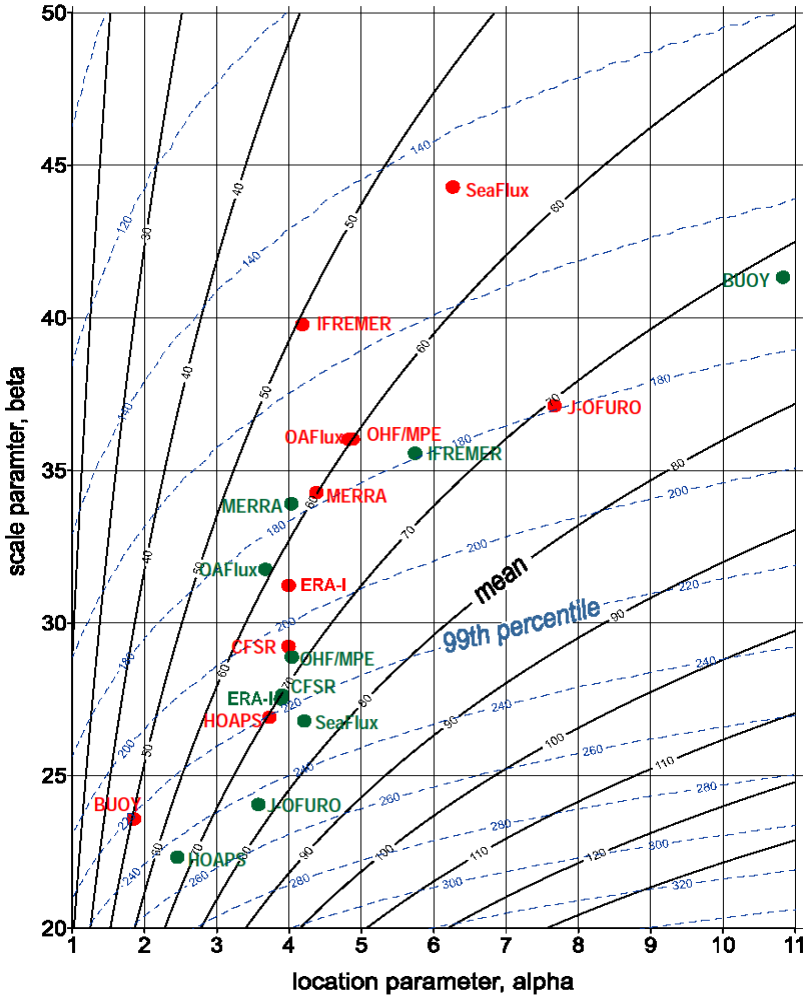

Figure 14. Estimates of winter latent heat flux from different products for the buoy locations 177.58° W, 57.05° N (red) and 130.00° W, 38.03° N (green) shown in the coordinates of the location and scale parameters of the MFT distribution. Black solid lines correspond to mean flux values and blue dash lines stand for 99th percentile of MFT distribution. Diagram shows that under the same differences in mean flux values differences in extreme fluxes may have different signs dependent on the tendencies in the location and scale parameters.

Figure 15 presents comparisons of means and 99.9th percentiles of sensible and latent turbulent heat fluxes between the buoys on one hand and CFSR and HOAPS on the other. Selection of CFSR and HOAPS for this comparison was justified by the fact that both are locations for which the differences between the means and between the 99th percentile values are consistent with each other (i.e. have the same sign) and are also locations where the differences in extreme fluxes are qualitatively different from the differences in mean values. Importantly, the relations between the mean and extreme fluxes are not everywhere consistent even qualitatively. For CFSR some subpolar locations, as well as a few locations in the equatorial Pacific Ocean, show stronger extremes recorded by buoys even if the mean fluxes are higher in the reanalysis. The California buoys and the east Atlantic Ocean buoys are also characterized by stronger extremes in the reanalysis but with means being higher in the buoy fluxes. This is especially evident for the sensible heat flux. HOAPS (Figure 15 b, d) tends to demonstrate stronger extremes than those for the buoy fluxes in many locations, even if the means are higher at the buoys. Note that even in the case when the signs of differences between the HOAPS and the buoy fluxes are consistent, the extreme fluxes in HOAPS always differ from the buoy values quite significantly (not shown). Selection of CFSR and HOAPS for this comparison was justified by the fact that in most locations CFSR and HOAPS show the highest extreme fluxes among reanalyses and satellite products respectively, and also show different sign of differences with each other at most buoy locations. Similar analysis was performed for the other data sets (figure not shown). Generally, IFREMER and SEAFLUX are tending to underestimate surface flux extremes (see Figure 12), but also show more consistency with reanalyses in e.g. the East Atlantic and East Pacific regions where CFSR and HOAPS demonstrate strong differences in the shape of PDFs.

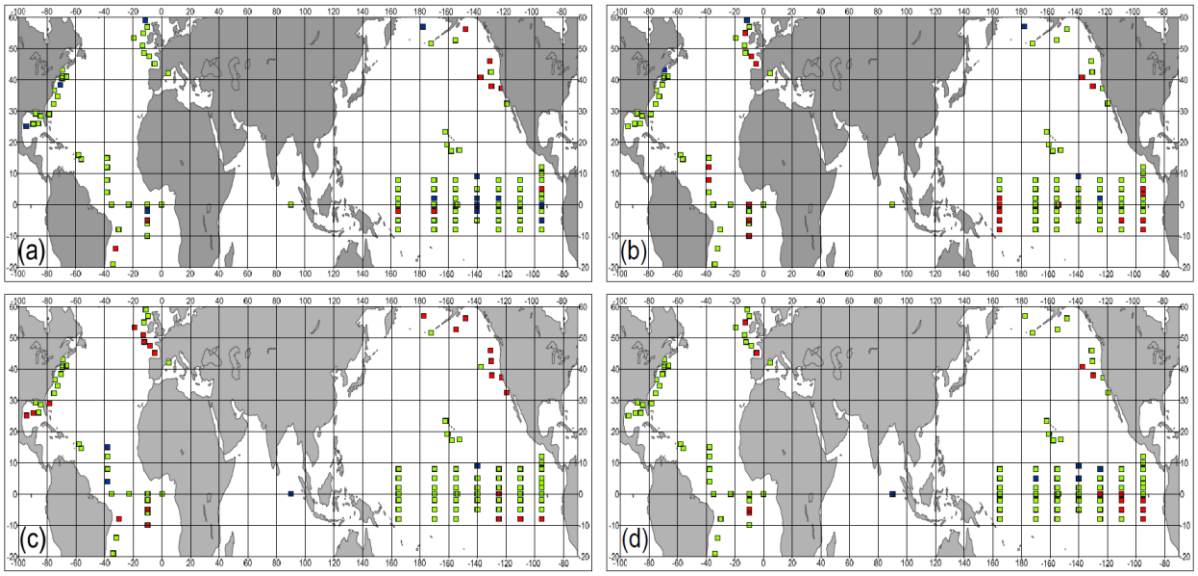

Figure 15. Characteristics of the consistency of differences between CFSR and buoy (a, c) and between HOAPS and buoy (b, d) for the mean and extreme latent (a, b) and sensible (c, d) heat fluxes. Green squares show the locations where the differences between the means and between the 99.9th percentiles hold the same sign and their PDFs are matching each other at 90% significance (k-s test). Blue (red) squares indicate the locations where buoy shows smaller (higher) mean fluxes compared to CFSR or HOAPS, but extreme fluxes are higher (smaller) in buoy time series.

In Figure 16 we show comparisons of different percentiles of the latent heat flux for the different products for the mid latitude East Pacific Ocean and the Tropical Atlantic. For this comparison we considered averaged PDFs for all buoys in the selected regions. Remarkably in the North East Pacific all satellite products tend to demonstrate larger values of high fluxes and smaller values for lower percentiles compared to the buoys. This is also clearly evident for the OHF/MPE product, showing the closest to buoy values at 50th and 75th percentiles. OAFLUX generally follows the buoy values with somewhat weaker values for lower percentiles and somewhat stronger flux extremes. Among the satellite-based products HOAPS and the J-OFURO demonstrate the strongest flux extremes while IFREMER shows the lowest extremes, although consistent at most percentiles with buoy data. The CFSR and ERA-Interim performance in the East Atlantic Ocean is comparable to the satellite products, showing stronger flux extremes than the buoys and OAFLUX, and considerably smaller

fluxes of lower percentiles. In comparison, MERRA demonstrates good consistency with the OAFlux in absolute values and with both OAFLUX and the buoys in the PDF behavior. In the Tropical Atlantic only the HOAPS satellite-based products show significant differences from the buoys and from the OAFLUX PDF, with weaker lower percentiles and stronger high percentile values. The OHF/MPE product fits well to both the buoy and OAFLUX at all percentiles. Comparison of the reanalysis fields for the tropical Atlantic Ocean shows general consistency with the buoy and the OAFLUX PDFs, with the differences for the different percentiles being primarily due to deviations in the means.

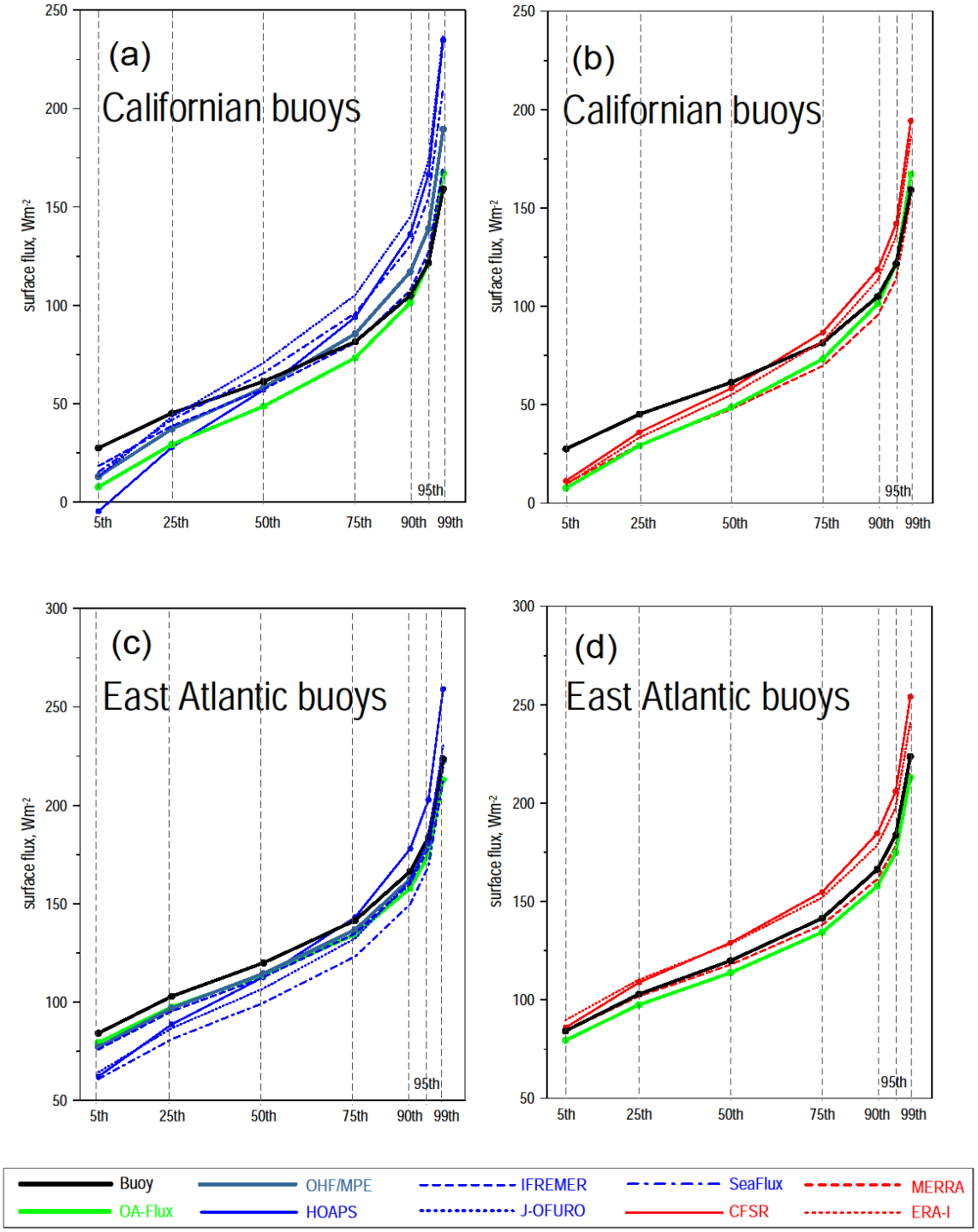

Figure 16. Comparison of different percentiles of the MFT distribution of the latent heat flux for different flux products for the East Pacific region (a, b) and the tropical Atlantic region (c, d). Panels (a) and (c) show satellite products in comparison with buoys and OAFLUX, panels (b) and (d) show comparisons for reanalyzes.

8 Summary and Conclusion

Over the last twenty years, there have been various attempts to estimate accurate LHF and SHF over the global ocean with high space and time resolution. LHF and SHF are estimated based on the use the aerodynamic bulk approach requiring the knowledge of variables such as surface wind speed, specific air and surface humidities, and air and surface temperatures. The LHF and SHF characteristics taking into account space and time resolutions may lead to significant differences between available heat flux products derived from remotely sensed observations or from atmospheric models.

One of the main motivations of this study is to further assess the quality of the products widely used by the scientific community based on the same validating method regardless of space and time characteristics of each product. This is achieved through the calculation of standardized heat flux products from the original IFREMER, HOAPS, OAFLUX, SEAFLUX, J-OFURO, ERA Interim CFSR, and MERRA data. All standardized data are estimated as daily averages on a regular 0.25° grid over the global ocean. The accuracy of the resulting LHF and SHF, as well as of the associated bulk variables, is mainly determined through comprehensive comparisons with daily data at more than 200 moorings in different areas of the world ocean during the period 2000-2007 when all products are available. In particular, the results for LHF and SHF quality indicate that the buoys and products compare well, with cross correlations all above 95% significance level, ranging between 0.83 (HOAPS) and 0.90 (ERA Interim) for SHF, and 0.78 (HOAPS) and 0.92 (OAFLUX) for LHF. The lowest RMS differences, about 25W/m² for LHF and 5W/m² for SHF, are found for numerical models. Such result would be related to the assimilation of bulk variable measurements into the models. Similar RMS differences are found for satellite-based

IFREMER, SEAFLUX and J-OFURO products. They are about 30W/m² and 7W/m² for LHF and SHF, respectively. However systematic bias between standardized product and the buoys were small to moderate (less than 10% of the mean values) across all moorings locations. Somewhat stronger biases in means were found for the buoy locations in the Gulf Stream and Kurishio boundary currents, these likely result from inaccurate product surface wind speed and/or specific air humidity.

The quality of each product was used for the determination of the multi-product ensemble (OHF/MPE) from the standardized LHF and SHF estimates. The ensemble product has the best agreement with the buoys in both LHF and SHF, with RMS differences not exceeding 25W/m² and 5 W/m², respectively. Even though the availability of in-situ heat flux datasets suitable for an extensive validation of flux products is quite limited, the use of OceanSites buoy estimates allowed the determination of flux characteristics in the tropical as well in extra-tropical locations. The derived results demonstrate that OHF/MPE exhibit the best statistics at almost all mooring locations. They also demonstrate that the combination of OHF flux products into a multi-product ensemble is a useful tool for further investigating the quality of each product at various spatial and temporal scales. The main result is that each product exhibits specific and significant regional departures that are varying in time.

A model for regression with correlated error, with sufficient information to constrain a relationship between two datasets by multiple samples from large gridded analyses, has been proposed. The model does not suffer from a neglect of autocorrelated errors, as in ordinary regression and triple collocation, but instead requires them (cf. Su *et al*. 2014). Statistical comparisons between observations and the products that employ these observations yield a complementary view on product performance. Relative performance among OHF products has been determined by the extent to which they share variations in a true flux that is common to ICOADS estimates. However, the main conclusion is that this common truth is quite small

and that the combination of correlated and uncorrelated error is large. This conclusion is subject to the caveat that extreme fluxes (greater than a few hundred Wm^{-2}) are omitted, but there is nevertheless good evidence that quantification of correlated (and total) error requires more attention (e.g., Gruber et al. 2016).

Analysis of the statistical distributions of surface turbulent fluxes performed using MFT distribution shows that differences in mean flux values across different products may be qualitatively different from the differences in flux extremes, implying differences in PDFs of fluxes. Generally, the behavior of PDFs of turbulent fluxes at buoys, in satellite products and reanalyzes is more consistent in the tropics compared to the mid and subpolar latitudes. Importantly, buoy values do not always demonstrate stronger surface flux extremes as might be anticipated for the high resolution point measurements. This is particularly observed in the eastern Pacific where most satellite products, as well as two of the three reanalyzes (CFSR and ERA-Interim), show stronger surface flux extremes compared to buoys, while buoy data report higher mean values. Differences in extreme fluxes (99.9^{th} percentile) at some locations between HOAPS and J-OFURO on the one hand, and buoy records on the other, may amount to 200-300 W/m^2 with the mean values being smaller in satellite products compared to buoy data. The reasons why satellite products may demonstrate stronger surface flux extremes compared to in situ observations should be discussed in terms of satellite retrievals of surface humidity and wind speed under extreme conditions. The OHF/MPE product designed and evaluated in this study is quite consistent with buoy data in terms of PDFs in the tropics and demonstrates a change in PDF in mid and subpolar latitudes towards underestimation of lower percentiles and overestimation of high percentiles.

We looked in this paper into the characteristics of satellite and reanalysis surface fluxes and standardized products compared to the buoy data mostly without going into details of the impact of individual variables and bulk algorithms used. While the analyzed satellite products

and OAFLUX employ the COARE-3.0 algorithm for the flux computations (as did also the buoy and ICOADS data), reanalysis fluxes are based on application of different algorithms. In this respect further comparison of the effect of individual variables in reanalyses by applying a standard algorithm (e.g. COARE-3.0) to the reanalyses state variables will be useful. For instance, recently Brodeau et al. (2017) reported significant differences between the fluxes derived by 3 algorithms applied to the same set of state variables. Such a comparison may help to establish more truth in understanding of the regional biases with respect to the buoy data. Another challenging task is to analyze the impact of data assimilation onto fluxes in reanalyses by considering output from the reanalyses models without data assimilation. In this respect results of comparisons of reanalyzes with buoys should be taken with the caveat that surface air temperature and humidity are constrained to a much lesser extent that surface winds and sea surface temperatures which can be well observed by satellite. Moreover, where surface air data are available at buoy locations this can lead to inhomogeneities as noted by Josey et al. (2014) who shows that the effect of dual assimilation of the humidity from TAO buoys in ERA-Interim can results in up to 50 Wm^2 flux anomalies during 1990s.

Based on these study results some general conclusions can be drawn. Further research aiming at the improvement and validation of bulk variables (surface wind speed, specific air humidity, and sea surface and air temperatures) at various space and time scales are highly recommended. More specifically, deeper calibration and validation of remotely sensed U_{10} and q_a should be performed over regions where most of the flux products show significant discrepancies in both LHF and SHF, such as near western boundaries and tropical regions. The expected results are to obtain improved characterization of all required bulk variable errors. Methods and algorithms dealing with bulk variable retrievals should be applied to generate consistent long time series of remotely sensed measurements (backscatter coefficients, brightness temperatures). Future improvements in the algorithms and the

associated flux product development are critically dependent on longer time series of in-situ data from different global sites in order to be representative of the full range of oceanic and atmospheric conditions..

Acknowledgments: The authors are grateful to ESA, EUMETSAT, CERSAT, JPL, ECMWF, NCEP, NASA, Météo-France, NDBC, PMEL, and UK MetOffice for providing numerical, satellite, and in-situ data used in this study. We would like to thank D. Croizé-Fillon and IFREMER/CERSAT team for data processing support. The authors would also like to thank the strong support of the reviewers to improve this study. This study is supported by the ESA Support to Science Element (STSE) program under contract 4000111424/14/I-AM. SKG also benefited from the Russian Science Foundation grant # 14-17-00697-II.

References

Akima H., 1970: A new method of interpolation and smooth curve fitting based on local procedures, J. ACM, October 1970, 17(4), 589-602.

Andersson, A., K. Fennig, C. Klepp, S. Bakan, H. Grassl, and J. Schulz, 2010: The Hamburg Ocean Atmosphere Parameters and Fluxes from Satellite Data - HOAPS-3. Earth Syst. Sci. Data Discuss., 3, 143-194, doi:10.5194/essdd-3-143-2010.

Andersson, A., Klepp, C., Fennig, K., Bakan, S., Grassl, H., &Schulz, J. , 2011: Evaluation of HOAPS-3 ocean surface freshwater flux components. Journal of Applied Meteorology and Climatology, 50, 379-398.

Andersson, E. and Järvinen, H., 1999: Variational quality control. Q. J. R. Meteorol. Soc., 125, 697-722

Atlas, R., Hoffman, R. N., Ardizzone, J., Leidner, S. M., Jusem, J. C., Smith, D. K., Gombos, D., 2011: A cross-calibrated, multiplatform ocean surface wind velocity product for meteorological and oceanographic applications. Bulletin of the American Meteorological Society, **92**, pp 157-174

Ayina L. H., A. Bentamy, A. Mestas-Nunez, G. Madec, 2006: The impact of satellite winds and latent heat fluxes in a numerical simulation of the tropical Pacific Ocean. *Journal of Climate*, 19(22), 5889-5902. http://dx.doi.org/10.1175/JCLI3939.1

Beljaars, A. C. M., 1995: The parametrization of surface fluxes in large-scale models under free convection. Q.J.R. Meteorol. Soc., 121: 255–270. doi: 10.1002/qj.49712152203

Bentamy, A., K. B. Katsaros, A. M. Mestas-Nuñez,W.M.Drennan, E. B. Forde, and H. Roquet, 2003: Satellite estimates of wind speed and latent heat flux over the global oceans. J. Climate, 16, 637–656.

Bentamy A., D. Croizé. Fillon, 2011: Gridded Surface Wind Fields from Metop/ASCAT Measurements. Inter. Journal of Remote Sensing, 33, pp 1729-1754.

Bentamy, A., S. A. Grodsky, K. B. Katsaros, A. M. Mestas-Nuñez, B. Blanke and F. Desbiolles , 2013: Improvement in air–sea flux estimates derived from satellite observations, *International Journal of Remote Sensing*, 34 (14), DOI:10.1080/01431161.2013.787502.

Bentamy A. S. A. Grodsky, A. Elyouncha, B. Chapron, F. Desbiolle, 2016 : Homogenization of Scatterometer Wind Retrievals, *Int. J. Climatol*. doi:10.1002/joc.

Berry D and E. C. Kent, 2009 : A new air-sea interaction gridded data set from ICOADS with uncertainty estimates. *Bulletin Of The American Meteorological Society 90: 645–656, DOI: 10.1175/2008BAMS2639.1.*

Bosilovich, M. G., cited, 2008: NASA's modern era retrospective analysis for research and applications: Integrating Earth observations. Earthzine. (http://www.earthzine.org/2008/09/26/nass as modern-era-retrospective-analysis/

Bradley, E. F. and C.W Fairall, 2007: A Guide to Making Climate Quality Meteorological and Flux Measurements at Sea. NOAA Technical Memorandum OAR PSD-311, NOAA/ESRL/PSD, Boulder, CO, 108 pp.

Brodeau, L., B. Barnier, S. K. Gulev, and C. Woods, 2017: Climatologically significant effects of some approximations in the bulk parameterizations of turbulent air-sea fluxes. *J. Phys. Oceanogr.*, 47, 5-28, doi: 10.1175/JPO-D-16-0169.1.

Casey, K.S., T.B. Brandon, P. Cornillon, and R. Evans, 2010: The Past, Present and Future of the AVHRR Pathfinder *SST* Program, in *Oceanography from Space: Revisited*, eds. V. Barale, J.F.R. Gower, and L. Alberotanza, Springer. doi: 10.1007/978-90-481-8681-5_16.

Charnock, H., 1955: Wind-stress on a water surface, Q. J. Roy. Meteorol. Soc., 81, 639–640

Chou, S.-H., E. Nelkin, J. Ardizzone, R. M. Atlas, and C.-L. Shie, 2003: Surface turbulent heat and momentum fluxes over global oceans based on the Goddard satellite retrieval, version 2 (GSSTF2). Journal of Climate, 16, 3256–3273.

Chou, S.-H., E. Nelkin, J. Ardizzone, and R.M. Atlas, 2004: A comparison of latent heat fluxes over global oceans for four flux products, *J. Climate*, *17*, 3973-3989.

Clayson, C. A., J. B. Roberts, and A. Bogdanoff, 2013: SEAFLUX Version 1: a new satellite-based ocean-atmosphere turbulent flux dataset. *International Journal of Climatology*, in revision

Danielson, R. E., 2017. Collocations of ICOADS and Ocean Heat Flux (centered five-day) latent and sensible heat flux estimates. SEANOE online data archive at http://doi.org/10.17882/_____, in preparation.

Danielson, R. E., J. A. Johannessen, M.-H. Rio, G. Quartly, B. Chapron, F. Collard, C. Donlon, 2017: Exploitation of error correlation in a large analysis validation: GlobCurrent case study, *Remote Sens. Environ.*, in preparation.Dee DP, Uppala SM, Simmons AJ, Berrisford P, Poli P, Kobayashi S, Andrae U, Balmaseda MA, Balsamo G, Bauer P, Bechtold P, Beljaars ACM, van de Berg L, Bidlot J, Bormann N, Delsol C, Dragani R, Fuentes M, Geer AJ, Haimberger L, Healy SB, Hersbach H, Holm EV, Isaksen L, K allberg P, Kohler M, Matricardi M, McNally AP, Monge-Sanz BM, Morcrette J-J, Park B-K, Peubey C, de Rosnay P, Tavolato C, Thepaut J-N, Vitart F., 2011 : The ERA-Interim reanalysis: configuration and performance of the data assimilation system. *Q. J. R. Meteorol. Soc*. 137 : 553 – 597. DOI:10.1002/qj.828

Fairall CW, Bradley EF, Hare JE, Grachev AA, Edson JB. 2003. Bulk parameterization of air–sea fluxes: updates and verification for the COARE algorithm. Journal of Climate 16: 571–591, DOI:10.1175/1520- 0442(2003)016<0571:BPOASF>2.0.CO;2.

Freeman, E., Woodruff, S. D., Worley, S. J., Lubker, S. J., Kent, E. C., Angel, W. E., Berry, D. I., Brohan, P., Eastman, R., Gates, L., Gloeden, W., Ji, Z., Lawrimore, J., Rayner, N. A., Rosenhagen, G. and Smith, S. R. , 2017 : ICOADS Release 3.0: a major update to the historical marine climate record. *Int. J. Climatol*., 37: 2211–2232. doi:10.1002/joc.4775

Gille, S., S. Josey, and S. Swart, 2016 : New approaches for air-sea fluxes in the Southern Ocean, *Eos, 97,* doi:10.1029/2016EO052243. Published on 13 May 2016.

Global Climate Observations System (GCOS), 2016: The Global Observing System for Climate: Implementation Needs. http://unfccc.int/files/science/workstreams/systematic_observation/application/pdf/gcos_ip_10oct2016.pdf

Grodsky S. A., A. Bentamy, J. A. Carton, R. T. Pinker, 2009: Intraseasonal Latent Heat Flux Based on Satellite Observations. *Journal of Climate*, 22(17), 4539-4556. http://dx.doi.org/10.1175/2009JCLI2901.1

Gruber, A., Su, C.-H., Zwieback, S., Crow, W. T., Dorigo, W., Wagner, W., 2016. Recent advances in (soil moisture) triple collocation analysis. Int. J. Appl. Earth Obs. Geoinform., 45, 200–211, doi:10.1016/j.jag.2015.09.002.

Gulev, S.K., 1994: Influence of space-time averaging on the ocean-atmosphere exchange estimates in the North Atlantic mid-latitudes. *J. Phys. Oceanogr*., 24, 1236-1255.

Gulev, S.K., T. Jung, and E. Ruprecht, 2007a: Estimation of the impact of sampling errors in the VOS observations on air-sea fluxes. Part I. Uncertainties in climate means. *J. Climate*, 20, 279-301.

Gulev, S.K., T. Jung, and E. Ruprecht, 2007b: Estimation of the impact of sampling errors in the VOS observations on air-sea fluxes. Part II. Impact on trends and interannual variability. *J. Climate*, 20, 302-315.

Gulev, S.K., and K.P. Belyaev, 2012: Probability distribution characteristics for surface air-sea turbulent heat fluxes over the global ocean. *J. Climate*, 25, 184-206, doi: 10.1175/2011JCLI4211.1

Gulev S. K, M. Latif, N. Keenlyside, W. Park, P.K. Koltermann, 2013: North Atlantic Ocean control on surface heat flux on multidecadal timescales. *Nature*, 499, 464–467, doi:10.1038/nature12268

Hubert, M., P. J. Rousseeuw, and T. Verdonck, 2012: A Deterministic Algorithm for Robust Location and Scatter, J. Comp. Grap. Stats, Vol. 21, pp. 618–637, doi: 10.1080/10618600.2012.672100.

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. Climate , 12 , 2856–2880, doi: 10.1175/1520-0442(1999)012 , 2856:NIITOH .= 2.0.CO;2

Josey S. A., L. Yu, S. Gulev, X. Jin, N. Tilinina, B. Barnier, and L. Brodeau, 2014: Unexpected impacts of the Tropical Pacific array on reanalysis surface meteorology and heat fluxes Geophys. Res. Lett., 41, 6213–6220, doi:10.1002/2014GL061302, 2014

Kurihara, Y., T. Sakurai, and T. Kuragano, 2006: Global daily sea surface temperature analysis using data from satellite microwave radiometer, satellite infrared radio meter and in-situ observations (in Japanese), Weather Bull., 73 , s1–s18.

Lock, A. P., A. R. Brown, M. R. Bush, G. M. Martin, and R. N. B. Smith, 2000: A new boundary layer mixing scheme. Part I: Scheme description and single-column model tests. Mon. Wea. Rev., 128, 3187–3199.

Louis, J. F., M. Tiedtke, and J. Geleyn, 1982: A short history of the operational PBL parameterization at ECMWF. Proc. Workshop on Planetary Boundary Layer Parameterization, Reading, United Kingdom, ECMWF, 59–80.

McColl, K. A., J. Vogelzang, A. G. Konings, D. Entekhabi, M. Piles, and A. Stoffelen, 2014: Extended triple collocation: Estimating errors and correlation coefficients with respect to an unknown target, Geophys. Res. Lett., 41, 6229–6236, doi:10.1002/2014GL061322.

Mestas-Nuñez A., A. Bentamy, K. B. Katsaros, 2006: Seasonal and El Niño variability in weekly satellite evaporation over the global ocean during 1996-98. *Journal of Climate*, 19(10), 2025-2035. http://dx.doi.org/10.1175/JCLI3721.1

Mestas-Nuñez, A.M., F.J. Kelly**,** A. Bentamy, and K.B. Katsaros, 2013. The ENSO footprint in monthly satellite evaporation over the global ocean during 1993-2007. *Remote Sensing Letters*, **4,** 706-714, doi:10.1080/2150704X.2013.788259

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

Reynolds, R.W., T.M. Smith, C. Liu, D.B. Chelton, K.S. Casey, M.G. Schlax , 2007: Daily High-Resolution-Blended Analyses for Sea Surface Temperature. *J. Climate*, 20, 5473-5496.

Rienecker, M. M., and Coauthors, 2011: MERRA: NASA's Modern-Era Retrospective Analysis for Research and Applications. *J. Climate*, **24**, 3624–3648, doi:10.1175/JCLI-D-11-00015.1.

Roberts, J. B., C. A. Clayson, F. R. Robertson, and D. Jackson, 2010: Predicting near - surface characteristics from SSM/I using neural networks with a first guess approach. J. Geophys. Res. , 115 , D19113, doi: 10.1029/2009JD013099

Saha S. *et al* , 2010: The NCEP climate forecast system reanalysis. *Bull Am Met Soc* 91:1015–1057

Sapiano, M. R. P., W. K. Berg, D. S. McKague, and C. D. Kummerow, 2012: Toward an Intercalibrated Fundamental Climate Data Record of the SSM/I Sensors. IEEE Trans. Geosci. Rem. Sens., Vol. 51 , Issue 3, doi 10.1109/TGRS.2012.2206601

Schlussel, P., L. Schanz, and G. Englisch, 1995: Retrieval of latent-heat flux and longwave irradiance at the sea-surface from SSM/I and AVHRR measurements, *Adv. Space Res.* , 16 , 107–116, doi:10.1016/0273- 1177(95)00389-V.

Smith, S.R., P.J. Hughes, and M. A. Bourassa, 2011: A comparison of nine monthly air-sea flux products. *International Journal of Climatology*, **30**, 1002-1027.

Simmons A, Uppala S, Dee D, Kobayashi S. 2006. ERA-Interim: New ECMWF reanalysis products from 1989 onwards. ECMWF Newsletter 110 : 26 – 35.

Stoffelen, A., 1998: Toward the true near-surface wind speed: Error modeling and calibration using triple collocation, J. Geophys. Res., 103(C4), 7755–7766, doi:10.1029/97JC03180.

Su, C.-H., D. Ryu, W. T. Crow, and A. W. Western, 2014: Beyond triple collocation: Applications to soil moisture monitoring, J. Geophys. Res. Atmos., 119, 6419–6439, doi:10.1002/2013JD021043.

Taylor, K.E.: Summarizing multiple aspects of model performance in a single diagram, J. Geophys. Res.,106, 7183-7192, 2001 (also see PCMDI Report 55, http://www-pcmdi.llnl.gov/publications/ab55.html)

Tomita, H. and M. Kubota, 2006: An analysis of the accuracy of Japanese Ocean Flux data sets with Use of Remote sensing Observations (J-OFURO) satellite-derived latent heat flux using moored buoy data,J. Geophys. Res, 111, C07007, doi:10,1029/2005 JC003013, 2006.

Trenberth K. E., J. T. Fasullo, J. Kiehl, 2009: Earth's global energy budget. Bulletin of the American Meteorological Society, **90**, 311–323, http://dx.doi.org/10.1175/2008BAMS2634.1

von Schuckmann, K., A. Cazenave, D. Chambers, J. Hansen, S. Josey, Y. Kosaka, N. Loeb, P.-P. Mathieu, B. Meyssignac, M. Palmer, K. Trenberth, M. Wild, 2016: An

imperative to monitor Earth's energy imbalance, Nature Climate Change 6, 138–144, doi:10.1038/nclimate2876

WGASF, 2000: Intercomparison and validation of ocean-atmosphere energy flux fields - Final report of the Joint WCRP/SCOR Working Group on Air-Sea Fluxes. WCRP-112, WMO/TD-1036. P. K. Taylor, Ed., 306 pp., http://eprints.soton.ac.uk/69522/1/wgasf_final_rep.pdf

Wentz, F. J, 1997: A well calibrated ocean algorithm for special sensor microwave/imager, J. Geophys. Res.,102(C4), 8703–8718.

Wentz, F. J., 2013: SSM/I Version-7 Calibration Report, report number 011012, Remote Sensing Systems, Santa Rosa, CA, 46pp.

Woodruff, S. D., S. J. Worley, S. J. Lubker, Z. Ji, J. E. Freeman, D. I. Berry, P. Brohan, E. C. Kent, R. W. Reynolds, S. R. Smith, and C. Wilkinson, 2011: ICOADS Release 2.5: extensions and enhancements to the surface marine meteorological archive. Int. J. Clim., 31(7), 951-967 (DOI: 10.1002/joc.2103).

Wu, Y., X. Zhai, and Z. Wang, 2016: Impact of synoptic atmospheric forcing on the mean ocean circulation. J. Climate, 29, 5709-5724.

Yu, L., X. Jin, and R. A. Weller, 2008: Multidecade global flux datasets from the objectively analyzed air–sea fluxes (OAFLUX) project: Latent and sensible heat fluxes, ocean evaporation, and related surface meteorological variables. Tech. Rep. OA- 2008-01, Woods Hole Oceanographic Institution, OAFLUX Project, 64 pp.