BAMS Article

Ocean Reference Stations

Long-Term, Open-Ocean Observations of Surface Meteorology and Air–Sea Fluxes Are Essential Benchmarks

Robert A. Weller, Roger Lukas, James Potemra,
Albert J. Plueddemann, Chris Fairall, and Sebastien Bigorre

ABSTRACT: There is great interest in improving our understanding of the respective roles of the ocean and atmosphere in variability and change in weather and climate. Due to the sparsity of sustained observing sites in the open ocean, information about the air–sea exchanges of heat, freshwater, and momentum is often drawn from models. In this paper observations from three long-term surface moorings deployed in the trade wind regions of the Pacific and Atlantic Oceans are used to compare observed means and low-passed air–sea fluxes from the moorings with coincident records from three atmospheric reanalyses (ERA5, NCEP-2, and MERRA-2) and from CMIP6 coupled models. To set the stage for the comparison, the methodologies of maintaining the long-term surface moorings, known as ocean reference stations (ORS), and assessing the accuracies of their air–sea fluxes are described first. Biases in the reanalyses' means and low-passed wind stresses and net air–sea heat fluxes are significantly larger than the observational uncertainties and in some case show variability in time. These reanalyses and most CMIP6 models fail to provide as much heat into the ocean as observed. In the discussion and conclusions section, long-term observing sites in the open ocean are seen as essential, independent benchmarks not only to document the coupling between the atmosphere and ocean but also to promote collaborative efforts to assess and improve the ability of models to simulate air–sea fluxes.

KEYWORDS: Climate records; Atmosphere-ocean interaction; Model evaluation/performance; Model errors; Surface observations

https://doi.org/10.1175/BAMS-D-21-0084.1
Corresponding author: Robert A. Weller, rweller@whoi.edu
Supplemental material: https://doi.org/10.1175/BAMS-D-21-0084.2
In final form 23 May 2022

AFFILIATIONS: Weller, Plueddemann, and Bigorre—Woods Hole Oceanographic Institution, Woods Hole, Massachusetts; **Lukas and Potemra**—University of Hawai'i at Mānoa, Honolulu, Hawaii; **Fairall**—NOAA/Physical Sciences Laboratory, Boulder, Colorado

Quantifying the exchanges of heat, freshwater, and momentum across the ocean's surface sheds light on how and where one forces or responds to the other. In collecting this information, observing methods should resolve energetic higher-frequency air–sea interaction processes while also capturing the lower-frequency modes of coupled variability and regional trends. However, in comparison to land, it is challenging to maintain sites in the open ocean where ongoing, high-quality observations of the air–sea exchanges of heat, freshwater, and momentum have been and continue to be made. As a result, alternatives to observed air–sea fluxes are in use. Atmospheric reanalyses, for example, provide gridded surface flux fields running typically from 1979 or 1980 to the present. In addition, climate models coupling the ocean and atmosphere provide retrospective analyses including surface fluxes covering 1850 to the present [e.g., Coupled Model Intercomparison Project phase 6 (CMIP6); Eyring et al. 2016].

We ask here how well the reanalyses and climate models succeed in simulating the means and low-passed variability in the air–sea fluxes. Atmospheric reanalyses are run with fixed model codes, and thus might be considered to yield consistent quality over the years. Climate models are looked to for their predictive capability, and how skillful they are at retrospective analyses could suggest how well they simulate coupled ocean–atmosphere variability. To examine the reanalyses and climate models we draw upon observations that we have collected using long-term surface moorings at three open-ocean sites since 2000, 2001, and 2004. To build the foundation for the comparison, the observing methodologies and estimates of the uncertainties in the observed air–sea flux time series are discussed. A number of prior studies have done comparisons between in situ and model-based surface meteorology and air–sea fluxes over shorter time periods and examined differences in state variables as well as in the fluxes. Our focus here is instead on comparison of the low-pass filtered fluxes over a period of close to two decades, looking at means and biases, and, in the case of the reanalysis fluxes, whether or not model–mooring differences vary in time. The results of the comparisons illustrate the value of sustained open observations of the fluxes in providing benchmarks for assessing fluxes from other sources.

The three surface moorings providing the data are the ocean reference stations (ORS) maintained by the Upper Ocean Processes Group at the Woods Hole Oceanographic Institution (WHOI). The ORS provide time series of surface meteorology that are used with bulk formulas to compute the air–sea fluxes of heat, freshwater, and momentum. The use of the ORS as sustained observing platforms is discussed in the second section. The technical approach to collecting the surface meteorological observations using the ORS is briefly summarized in the third section and in more detail in Weller et al. (2012) and Weller (2018). The fourth section summarizes assessments of the quality of the surface meteorological observations and of the air–sea fluxes computed using the COARE 3.0 formulas (Fairall et al. 1996, 2003). Quantifying the uncertainties associated with the ORS time series sets the stage for their use as benchmarks for the comparisons with models. The fifth section presents an overview of the surface meteorology and air–sea fluxes at each of the three ORS. Three modern

atmospheric reanalyses and recent coupled model runs were chosen to provide model-based fluxes. Model output was extracted at the grid points closest to ORS. The reanalyses are NCEP-2 (Kanamitsu et al. 2002), ERA5 (Hersbach et al. 2020), and MERRA-2 (Gelaro et al. 2017). The climate model outputs come from CMIP6 (Eyring et al. 2016) models. The sixth section focuses on low-pass filtered time series, annual and longer-term means, and comparisons of ORS time series with the models. The seventh section presents discussion and conclusions, making the case that at a number of open-ocean sites' air–sea flux time series collected with care provide an essential means to assess how well fluxes from other sources capture the long-term means and low-passed variability in the communication of heat, freshwater, and momentum across the sea surface.

Surface moorings used for sustained observing

In the open ocean, away from coasts and the equator, there are few sites where continuous, long-term, time-resolved observations in the atmospheric and oceanic surface layers are collected. Surface buoys, moored to the seafloor, provide attractive platforms to do so. In recent decades, surface mooring technology improved (Davis et al. 2018), enabling surface moorings to be deployed in equatorial waters of the Pacific Ocean (McPhaden 1995), the Atlantic Ocean (Bourlès et al. 2019), and the Indian Ocean (McPhaden et al. 2009) as key components of sustained ocean observing. Away from the equator, however, wind and wave conditions are more challenging, and the three ORSs are among the few long-term, open-ocean surface moorings maintained there. Other extraequatorial, long-term surface moorings include the Ocean Climate Stations (OCS; including KEO off Japan and PAPA in the Gulf of Alaska, www.pmel.noaa.gov/ocs/) and the Australian Southern Ocean Time Series mooring south of Tasmania (SOTS; https://imos.org.au/facilities/deepwatermoorings/sots). These long-term moorings are coordinated Internationally by OceanSITES (www.oceansites.org) as part of the Global Ocean Observing System. Efforts in the United States to maintain surface moorings as ocean reference sites are supported by the National Oceanic and Atmospheric Administration (https://globalocean.noaa.gov/Research/Ocean-Reference-Stations-OceanSITES).

Surface mooring deployments are typically one year in length, limited by mechanical wear and fatigue of components and by how long sensor calibrations last in the field. Careful intercomparisons and cross calibrations of instruments deployed on successive deployments enables merging of records into long, accurate, and consistent time series with high time resolution (once per minute). To compute the air–sea fluxes of heat, freshwater, and momentum we applied empirical bulk formulas to the observations of surface meteorology. While some surface moorings, especially weather buoys, share observations in near–real time, the ORS surface moorings discussed here are unique in the decision to withhold their data from assimilation into models. In this respect the ORS are distinct from more numerous and widely distributed surface observation sites that are specifically deployed for data assimilation purposes. They are also unique in the level of effort devoted to ensuring collection of complete surface meteorological and air–sea flux during each deployment and to assessing the quality of the observations. These steps are taken to collect independent observations of known quality that can be used to assess operational models and support collaborations on model improvement.

The three ORS sites lie in extraequatorial trade wind regions (Fig. 1). Trade winds cover ~50% of the World Ocean, flowing westward and equatorward and outward from the subsidence in subtropical high pressure centers into ascending deep convection in the intertropical convergence zones (ITCZs). The Stratus ORS site, 1,500 km west of northern Chile is located where marine stratocumulus clouds are characteristic and near the eastern boundary of the South Pacific where coupled models yield too warm SSTs. It was established in October 2000 to collect the observations needed to quantify the air–sea fluxes and improve understanding of

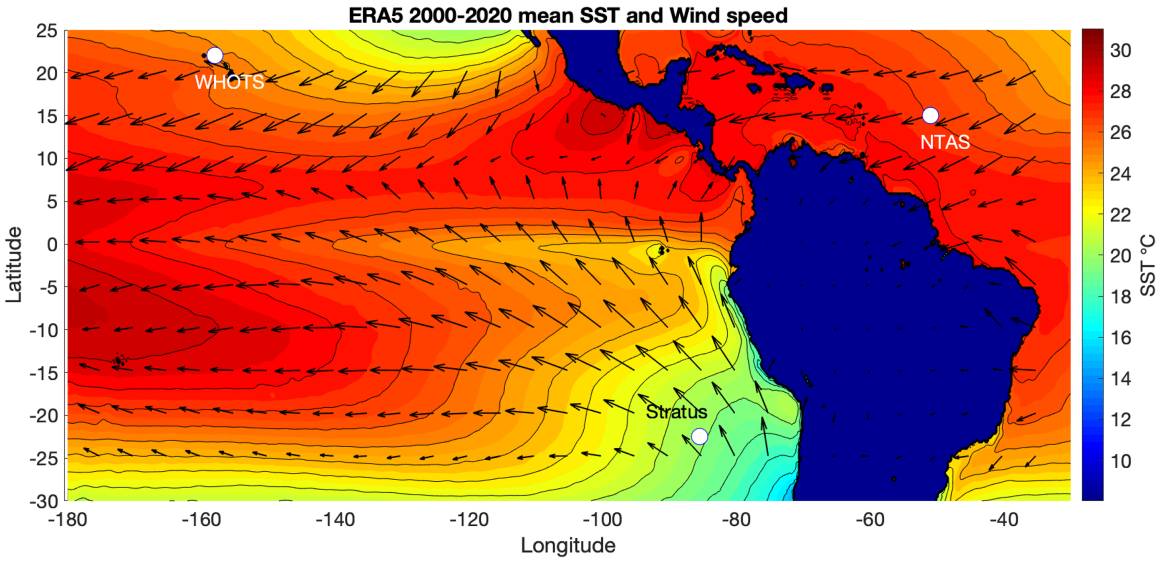

Fig. 1. Map showing the locations of the three ocean reference stations, WHOTS, Stratus, and NTAS plotted on top of color contours of the 2000–20 mean sea surface temperature field and wind vectors from ERA5.

the process that govern sea surface temperature (SST) in this data-sparse region of the eastern South Pacific. The Northwest Tropical Atlantic Station (NTAS) ORS was established in March 2001 1,100 km east of the island of Martinique in the Caribbean, in a region of interest from the perspective of oceanic heat storage and Atlantic cyclogenesis. NTAS provides data to document the air–sea fluxes and investigate connections between upper-ocean heat storage and hurricane activity. The WHOI Hawaii Ocean Time series (HOT) Station, known as the WHOTS ORS, was first deployed in August 2004 at Station ALOHA 100 km north of Oahu. ALOHA is a site of long-term ongoing oceanographic studies begun in 1988 (http://aco-ssds.soest.hawaii.edu/ALOHA/), including the ocean context for the ongoing atmospheric CO_2 time series being collected on Mauna Loa since the 1950s (https://gml.noaa.gov/obop/mlo/). Like NTAS, WHOTS is characterized by trade wind convective clouds and systems. The WHOTS ORS provides the air–sea interaction and physical oceanographic context for the other ongoing studies; it also carries a NOAA PMEL CO_2 package (Sutton et al. 2014), measuring surface ocean and atmospheric boundary layer pCO_2 and complementing the Mauna Loa time series.

The goals in maintaining the ORS are to collect surface meteorological and upper-ocean data with high reliability and accuracy; carry out calibrations and at-sea intercomparisons necessary to support processing and merging data from the individual deployments and making overall data quality assessments; produce time series of the surface fluxes of heat, freshwater, and momentum with accompanying estimates of uncertainties; and make the resulting merged climate data records with documented accuracy widely available. During their deployment the ORS also serve as focal points for research and modeling activities. Multi-investigator process studies such as VOCALS-ReX (Wood et al. 2011), held around the Stratus ORS, and ATOMIC/EUREC^{4}A (Stevens et al. 2021) around NTAS, investigated ocean, atmosphere, and coupled dynamics and processes. Delayed access to hourly time series of surface meteorology has been made available to modeling centers. Exchanges of model and ORS surface meteorology have led to dialog with investigators at ECMWF about ORS–model comparisons. ORS data have also been used for validation of remote sensing products (e.g., Pinker et al. 2018). The consistent ORS time series have facilitated quantitative assessments of surface meteorology and air–sea fluxes from reanalyses (e.g., Valdivieso et al. 2017), from coupled climate models (e.g., Zuidema et al. 2016), and from hybrid combinations of model and remote sensing fields (e.g., Yu and Weller 2007).

Equipping and maintaining the ocean reference stations

Figure 2 shows the towers of the original 3-m aluminum ORS buoy hull and the more recent closed cell foam 3-m hull. The wind vane bolted to the tower orients the buoy creating an upwind location for mounting anemometers, air temperature and air humidity sensors. Goals in equipping the tower with meteorological sensors and associated electronics were accuracy; reliability; redundancy; low power consumption; low susceptibility to corrosion, radio frequency interference, and salt spray; internal storage of key metadata (sensor model and serial number, calibration information); easy digital communication; and the ability to deploy sensors either as stand-alone units or linked together to form a system. Meeting these goals lead to the development of the Improved Meteorological (IMET) system, later named the Air–Sea Interaction Meteorological (ASIMET) system (Hosom et al. 1995). To achieve the reliability necessary to collect complete records two separate and redundant modular ASIMET systems are deployed on the buoy, each sampling once per minute and running for over a year on battery power. In addition, stand-alone ASIMET modules logging internally are often added for further redundancy. ASIMET sensor modules measure downward shortwave radiation (DSWR), downward longwave radiation (DLWR), air temperature (T_a) and relative humidity (q_a), accumulated precipitation (P), barometric pressure (SLP), wind speed (WSPD) and direction (WDIR). Radiometers are mounted above all other structures to avoid shadows. Anemometers, T_a and q_a sensors are mounted on the upwind face of the tower. The P and SLP modules are mounted on the tower center. Modules measuring ocean temperature and salinity are mounted below the surface. Sensor heights/depths are included in Table 1.

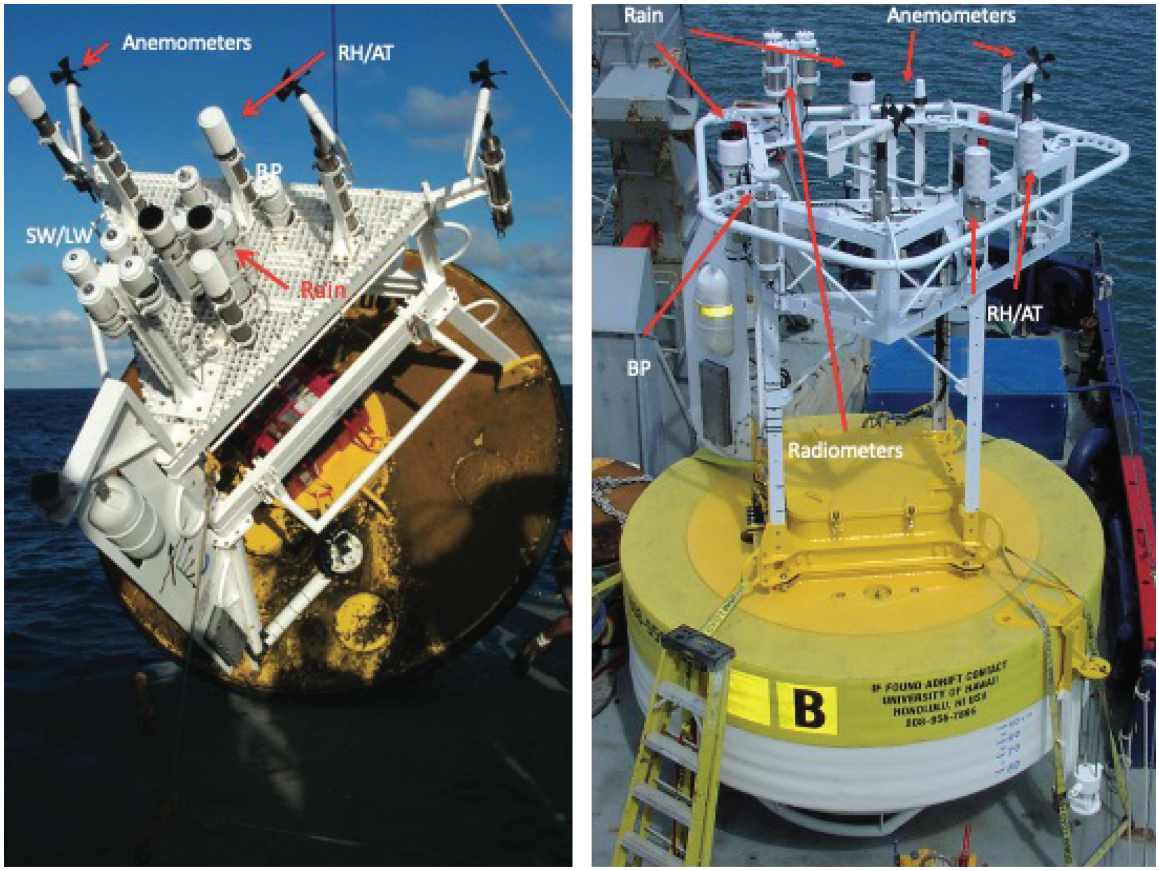

Fig. 2. ORS surface buoys with (left) older, 3-m aluminum hull and (right) present 3-m foam hull. A wind vane with radar reflector and data telemetry antenna orients the buoy with respect to the wind, and anemometers and air temperature and humidity modules (RH/AT) are placed on the upwind side of the tower. The cluster of four radiometers, one each for shortwave (SW) and longwave (LW) from two ASIMET systems, is placed above all other structures. Rain gauges and barometric pressure (BP) sensors are placed aft of the forward face.

Table 1. Present sensors in the ASIMET system. Each are configured as an ASIMET module, capable of being connected to a central datalogger and of recording internally, either to provide a record redundant to that in the central logger or as a stand-alone module.

Observable	Sensor make and model	Typical height above sea surface	Notes
Wind	R. M. Young 5103	2.82 m	Propeller-vane anemometer, stock propeller bearing upgraded
Wind	Gill Instruments WindObserver II ultrasonic anemometer	3.08 m	Sonic anemometer, used at times to mitigate data loss due to birds
Air temperature, relative humidity (T_a, q_a)	Rotronic MP-101A	2.95 m	Porous Teflon filter and multiplate radiation shield
Downward shortwave radiation (DSWR)	Eppley precision spectral pyranometer	3.43 m	Case adapted to ASIMET module tubing
Downward longwave radiation (DLWR)	Eppley precision infrared radiometer	3.43 m	Case adapted to ASIMET module tubing
Barometric pressure (SLP)	Heise DXD	3.0 m	With parallel-plate pressure port
Precipitation (P)	R. M. Young 50202	3.12 m	Self-siphoning rain gauge
Sea surface temperature and salinity (SST, SSS)	SeaBird 37	−0.85 m	Mounted on buoy bridle

Recovery of a deployed surface mooring, preceded by nearby deployment of a fresh surface mooring, is planned annually. After recovery, calibrated and quality-controlled data are used together with the COARE bulk formulas (Fairall et al. 1996, 2003) to compute heat, freshwater, and momentum fluxes. Further analysis and processing are carried out to merge surface meteorology and air–sea fluxes from successive deployments as climate data records. As the bulk formulas are updated or when new information is available about calibration or processing procedures, the deployment-by-deployment data are reprocessed and updated merged time series created. During each deployment, hourly surface meteorological data are telemetered to support checking ORS systems; telemetry also aids testing on land and at-sea comparisons between the ORS and shipboard sensors.

Fifty-four of 58 annual cruises have been conducted as planned. The research vessel is equipped with an additional set of surface meteorological sensors. For many cruises, these additional sensors have come from the NOAA Physical Sciences Laboratory (PSL), Boulder, Colorado (Fairall et al. 2008), including eddy covariance fluxes, motion-stabilized shortwave and longwave radiometers, and a temperature sensor deployed close to the sea surface. Otherwise, additional freshly calibrated, stand-alone ASIMET modules are installed on the ship. Upon arrival at the ORS site, the fresh surface mooring is deployed and one to several days of data are collected with both old and new buoys in the water. During that overlap the ship spends time on station downwind of each buoy with bow into the wind. At the new buoy, data from the shipboard sensors provide an immediate check on the quality of the freshly deployed sensors. If issues are seen, a trip to the buoy in a small boat is made to investigate; spare modules are swapped if needed to resolve problems. For both buoys, the shipboard sensors provide the means to examine differences between sensors on the two different types of platforms. A further motivation for the shipboard observations is to provide an end of deployment assessment of the old buoy's sensors against the chance they are damaged on recovery and not available for postcalibration. After the in situ intercomparison the old mooring is recovered. Recovered ASIMET modules are photographed, and data spikes induced (e.g., by putting covers on the shortwave radiometers and placing SST sensors in ice water) to allow checking and refining the time bases of the sensors. The returned ASIMET modules are postcalibrated; any performance anomalies during the deployments cause modules to be flagged for closer scrutiny. Differences between observations from the old and new buoys are resolved by comparing the overlapping buoy records to the shipboard observations at the mooring sites (Bigorre and Galbraith 2018). Overall data return has been high, but occasional

failures of the mooring have resulted in shorter-than-planned deployments. Total gap time at Stratus is 1.8 years out of 21 years planned. Total gap time at NTAS is 2.4 years out of 20 years planned. There are no gaps in the occupancy of the WHOTS ORS site.

Over the years, there has been evolution in the observing hardware. Figure 2 notes the change to a more easily maintained and shipped buoy hull. In addition, the IMET/ASIMET systems have seen evolution. Data storage media have changed; initially, PCMCIA flash memory cards were used and now microSD card storage is being phased in. Testing and evaluation of the sensors is carried out in parallel to field operations. New sensors and associated updated digital electronics are tested first on land. Then, they are deployed as stand-alone modules on an ORS buoy or on only one of the two ASIMET systems on the buoy in order to obtain coincident time series for evaluation. The practices of overlapping old and new buoy deployments and obtaining independent shipboard observations coincident with the buoy records have provided important foundations for maintaining comparability of successive deployments and merging those into longer time series.

The quality of ORS surface meteorological observations and derived air–sea fluxes

ORS meteorological sensors are exposed to solar heating, salt spray, bird guano, birds attempting to land, and contact with ships. Recovered sensors are brought back without cleaning for postdeployment calibration. Salt spray contamination of the humidity sensors is prevented by a porous Teflon filter. Radiometers returned for postcalibration have pointed to little impact from salt spray and much greater concern about bird guano and birds roosting on the sensors, leading to efforts to minimize bird landings on the buoy by adding bird spikes. In assessing such issues the comparison with fresh, attended sensors on the ship and with the fresh sensors on the new buoy aid in identifying impacts of contamination. After quality checks, data from the two redundant ASIMET systems and additional stand-alone sensors have been able to be merged to yield a complete 1-min dataset 99.9% of the on-station time.

In addition to the goal of obtaining complete time series, a focus has been on quantifying observational accuracies. Colbo and Weller (2000) assessed the accuracy of the IMET and ASIMET sensors for the basic 1-min sampling and for daily and annual averages. Accuracies in the field, when sensor drifts and environmental impacts were considered, were found to be significantly less than those cited by manufacturers and from laboratory calibrations. Analyses of the differences between redundant, coincident 1-min samples were used to determine the uncertainties of the basic, 1-min measurements. Such differences can also come from clock drift. For example, due to clock drift, the two redundant downward shortwave radiometers see, by the end of a deployment, short-lived transients in insolation at different times and show larger differences in 1-min data than in daily or annual averaged data. A 20 W m^{-2} value in instantaneous downward shortwave (Table 2) reflects finding such differences in the 1-min sampled data. Some errors, such as from radiative heating of passively ventilated sensors, buoy tilt, salt spray, flow distortion, and anemometer bearing friction, depend on time of day and on wind speed. Air and sea temperature sensors that experience radiative heating and that are subject to near-surface temperature gradients have uncertainties in the 1-min data much larger than laboratory calibrations of their sensors. Daily averages have lower uncertainty than the 1-min values due to averaging over the sensor errors due to radiative heating. Longer-term averages span diverse conditions, reducing the error contribution, for example, from low-wind periods. Rain events are transient, and it was not possible to accumulate sufficient rainfall observations in the coincident overlapping buoy and shipboard observations to support error analysis. Thus, an error of 10%, consistent with Nešpor and Sevruk (1999), is used for rainfall.

Colbo and Weller (2009) used coincident 1-min samples from duplicate sensors and analyzed the difference between redundant time series. They computed the spectra of the

Table 2. Summary of ASIMET sensor measurement accuracies in the field based on 1-min observations [after Colbo and Weller (2009) and Weller (2018)] in column 2, uncertainties in daily averages in column 3, and uncertainties in annual averages in column 4. For context, the mean values from Table 4 for the Stratus ORS are included in column 1; note that WHOTS and NTAS had an order-of-magnitude greater rainfall rate.

Sensor	Stratus mean	1 min	Daily	Annual
Downward longwave (W m^{-2})	376.8	7.5	4	4
Downward shortwave (W m^{-2})	205.2	20	6	5
Relative humidity (%RH)	73.4	1 3 (low winds)	1 3	1
Air temperature (°C)	19.6	0.2 (more in low wind)	0.1	0.1
Barometric pressure (hPa)	1,017.7	0.3	0.2	0.2
SST (°C)	20.5	0.1	0.1	0.004
Wind speed (m s^{-1})	5.8	1.5%, 0.1 (more in low wind)	1%, 0.1 (max of these)	1%, 0.1 (max of these)
Wind direction (°)	304.8	6 (more in low wind)	5	5
Rainfall (% under catchment)	0.007 mm h^{-1}	10%	10%	10%

difference time series to characterize some error sources (e.g., clock drift), and built empirical probability distribution functions (PDFs) to quantify the measurement error with 50% and 95% confidence limits for each ASIMET sensor. They noted that the 1-min accuracy of variables with high-frequency variability (e.g., shortwave, wind speed) estimated from their empirical PDFs may be adversely affected by clock drift. They also looked at the standard deviation of each sensor's 1-min time series on several deployments of Stratus and NTAS buoys and compared with the shipboard measurements made during the deployment/recovery of these moorings to determine uncertainty in the 15-min and annual averages. The field accuracies of the ASIMET/IMET sensors in the subtropics provided by Colbo and Weller (2009) are summarized in Table 2, which reflect the more conservative 95% confidence limit from the empirical PDF and the multiple intercomparisons with ship measurements. Our emphasis here is on the actual field accuracy of the sensors, which can be much less than the accuracies from either manufacturer specifications or laboratory calibrations. Bigorre et al. (2013) followed a similar approach using ASIMET measurements at a higher latitude with higher wind and sea state conditions than the ORS locations discussed here and characterized further the random and systematic sources of field measurement errors, in particular the impact of platform motion and flow distortion on the buoy.

Colbo and Weller (2009) used error propagation equations derived from the COARE 3.0 bulk formulas (Fairall et al. 1996, 2003) together with their sensor accuracies to estimate the uncertainties in the fluxes computed from the ASIMET/IMET sensor measurements. Table 3 summarizes the accuracies of annual averages of the bulk formulas fluxes using propagated errors listed in Table 2. The uncertainty in net heat flux is computed assuming no covariance between heat flux components. However, errors in the individual net heat flux components can offset each other resulting in a smaller net heat flux error. In light winds, for example, radiative heating errors impact sensors but buoy tilt-related errors are smaller, while for stronger winds radiative heating errors are small but tilt-related errors may increase.

The bulk formulas on which ORS estimated fluxes rely continue to be scrutinized and validated (Cronin et al. 2019). Observing packages with fast response sensors that enable direct computation of the turbulent fluxes were deployed on some buoys in parallel with ASIMET. As such they have over the years supported the development of the COARE algorithms, validated the performance of the algorithms in moderate wind regimes where the bulk of the data were acquired, and spurred updates to the COARE flux algorithms (e.g., Edson et al. 2013)

as deployments captured regimes with higher sea states and wind speeds. The ORS fluxes presented here were computed using COARE 3.0, which performs well in the trade wind regime. Updated versions 3.5 and 3.6 are available (ftp1.esrl.noaa.gov and https://github.com/noaa-psd/pyCOARE). Version 3.5 made a change in the version 3.0 drag coefficient at high wind speeds (as per Edson et al. 2013). Version 3.6 upgraded the wave-drag parameterization and was modified to accept surface salinity as an input to allow use on large lakes. Both versions preserved the specification of the sensible and latent heat transfer coefficients so the effect on the surface energy budget calculations is negligible. To verify this, the ORS fluxes were computed again using COARE 3.6 (Edson et al. 2013); the mean difference from COARE 3.0 for the ORS was 0.13 W m^{-2} for Q_{LH} and −0.08 W m^{-2} for Q_{SH}. COARE 3.6 yielded mean $|\tau|$ that was 4% lower than that from COARE 3.0. The updates to the COARE algorithm were motivated primarily by new observations in wind speed regimes above 10 m s^{-1}. They include improvements to how surface waves and surface roughness are represented and do not significantly impact the findings presented here.

Table 3. Uncertainties of annual averaged fluxes from an ORS buoy, after Colbo and Weller (2009). The accuracy of *E* − *P* was calculated computing evaporation from latent heat flux and using mean rainfall typical of NTAS and WHOTS. Stratus long-term mean values provide a context for the uncertainties.

Flux	Stratus mean	Typical
Net longwave (Q_{LW}; W m^{-2})	−42.5	3.9
Net shortwave (Q_{SW}; W m^{-2})	193.9	5
Latent (Q_{LH}; W m^{-2})	−106.9	5
Sensible (Q_{SH}; W m^{-2})	−7.6	1.5
Net heat flux (Q_{net}; W m^{-2})	36.9	8
Wind stress magnitude (N m^{-2})	0.075	0.007
$E - P$ (mm h^{-1})	0.15	0.01

An overview and the long-term means of the surface meteorology and air–sea fluxes

Using the raw 1-min ASIMET meteorological data, daily averages were computed showing a strong daily signal in DSWR (Fig. 3). The fluxes were computed using the hourly meteorology. With the focus here on means and lower-frequency variability, the analyses below use the daily averaged meteorological and flux time series. At all three ORS, annual cycles are evident in DSWR, SST, T_a, q_a, Q_{net}, SLP, and Q_{LW}. To avoid potential biases and false trends, long-term means were computed across only full calendar years (Table 4). The filtered results were truncated at the ends to eliminate end effects.

The differences in surface meteorological means in Table 4 across the three sites can be understood within the Lagrangian framework of the trade wind boundary layer (Albrecht, 1979, 1984; Albrecht et al. 1995a,b; Bretherton et al. 1999; Albrecht et al. 2019) with the sites on a virtual streamline. Stratus is in the strong subsidence of the subtropical high of the eastern basin with associated stratocumulus cloud deck. WHOTS is in the fair-weather cumulus core of the trade winds equatorward and west of the high where the surface layer becomes decoupled from the troposphere. NTAS is in the region of strong moisture convergence close to deep tropical convection in the western tropics poleward of the ITCZ. The long-term mean DSWR is smallest at Stratus (205.2 W m^{-2}) and largest at NTAS (246.7 W m^{-2}) due to decreasing latitude and cloud fraction downstream in the virtual framework. Mean DLWR increases downstream going from Stratus to WHOTS to NTAS due to warming, moistening, and thickening of the planetary boundary layer (PBL). Mean SSTs, T_a, and q_a all increased going from Stratus, to WHOTS, and then to NTAS.

Figure 4 shows the daily averaged air–sea fluxes at Stratus (wind stress, Q_{net} and its components, and net freshwater flux represented by cumulative precipitation and evaporation). The complementary figures for WHOTS and NTAS are in the online supplementary material (https://doi.org/10.1175/BAMS-D-21-0084.2). Table 5 presents the mean long-term fluxes. Mechanical forcing of the ocean surface by the mean wind stress is strongest at NTAS. All three ORSs have mean net ocean heating of between 20 and 40 W m^{-2}. Mean Q_{LW} is between −42 and −57 W m^{-2},

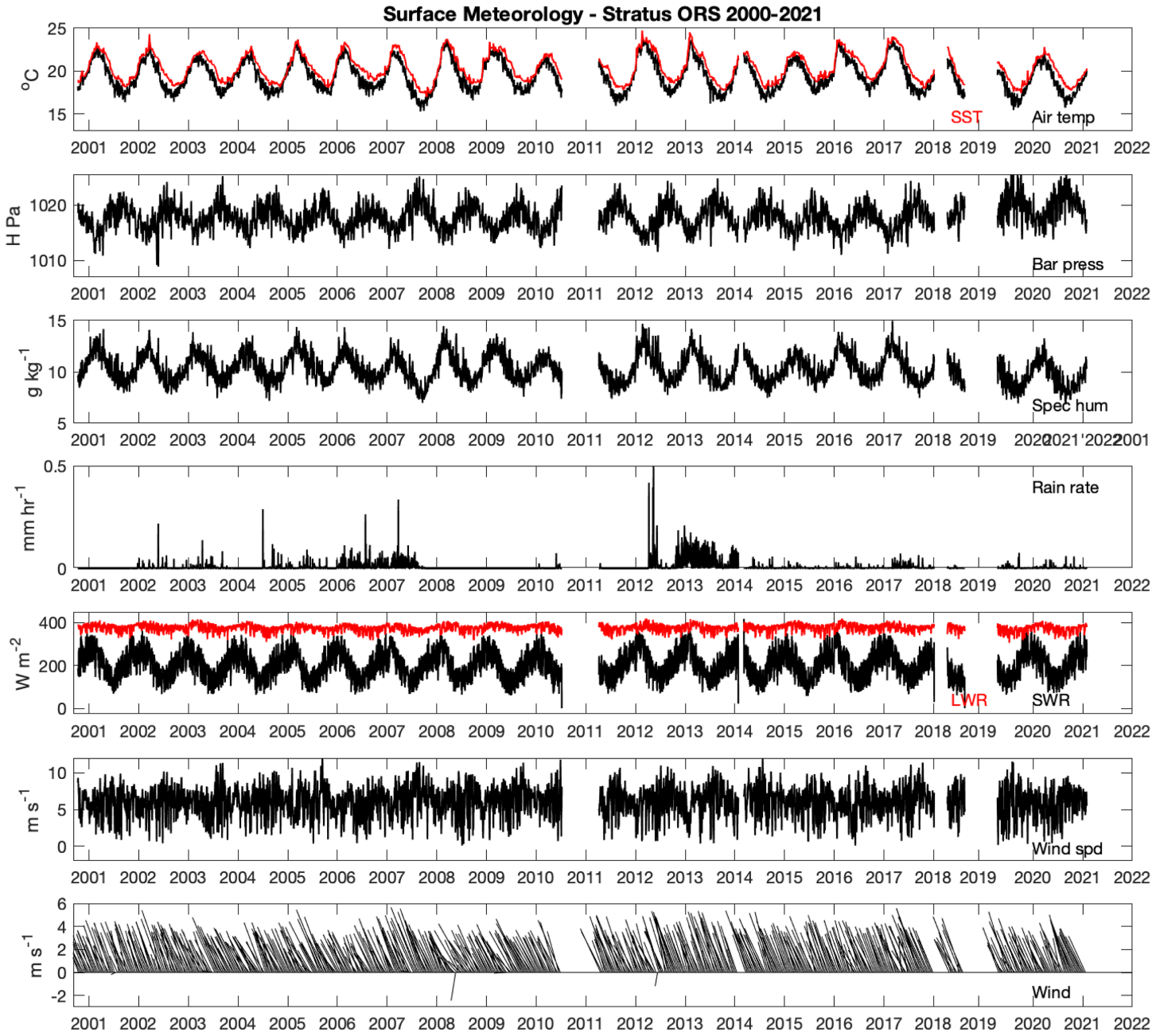

Fig. 3. Daily averaged surface meteorology as observed at Stratus. Gaps in the data resulted from breaks in the mooring. (from top to bottom) SST, T_a, SLP, q_a, rain rate, DLWR (red) and DSWR (black), wind speed, and vector wind, respectively. Matching figures for WHOTS and NTAS are provided in the supplementary material. The wind vectors (toward) are subsampled every 5 days.

largest at WHOTS corresponding to the minimum Q_{net} there. The Q_{LW} heat loss is larger than the turbulent sensible heat flux (–8 to –4 W m^{-2}). Q_{LH} is the largest component of heat loss, –106.9, –135.9, and –139.0 W m^{-2}, respectively, at Stratus, WHOTS, and NTAS, reflecting the downstream increases of wind speed and SST. Over 1.2 m yr^{-1} of freshwater is transferred to the atmosphere at each ORS site.

Low-frequency variability and comparisons of the ORS time series to modern reanalyses and CMIP6 models

A particular incentive for sustaining the ORS is the development of in situ climatologies at each site, characterizing the temporal variability of the fluxes, quantifying means, and documenting low-frequency variability. The ORS are intended to provide reference sites, benchmarks against which other sources of air–sea fluxes are compared. The observed flux uncertainties

Table 4. Long-term means of surface meteorology at the three ORS formed by taking the annual means of the mean daily time series formed from the calendar years of complete data. WHOTS has 14 full years (2005–19), Stratus 15 full years (between 2000 and 2021), and NTAS 16 full years (between 2001 and 2021). Surface current data come from the current meter closest to the surface on the mooring and are used when computing the wind relative to the surface current in the bulk flux formulas. The variables are as observed at the heights given in Table 1. Both relative and specific humidity are given.

Variable	Stratus	WHOTS	NTAS
Wind speed (m s^{-1})	5.8	5.5	6.3
Wind direction (toward, °)	304.8	263.2	257.4
Air temp (°C)	19.6	24.2	26.6
SST (°C)	20.5	25.1	27.2
RH (%), SH(g kg^{-1})	73.4, 10.4	75.5, 14.2	76.6, 16.6
Barometric press (hPa)	1,017.7	1,017.0	1,014.6
Downward shortwave (W m^{-2})	205.2	237.3	246.7
Downward longwave (W m^{-2})	376.8	388.8	405.9
Rain rate (mm h^{-1})	0.007	0.063	0.057
Surface current speed (m s^{-1})	0.04	0.06	0.08
Surface current direction (°)	263.4	289.9	300.3

(Table 3) are small compared to the means in Table 5. We chose here to compare the ORS to two types of models, atmospheric reanalyses and coupled climate models. While operational atmospheric models evolve over the years as improvements are made to their physics or data assimilation schemes, atmospheric reanalyses utilize model code and assimilation methods that are fixed and thus potentially offer consistent time series of fluxes over the years. One source of potential change with time in the performance of reanalyses, however, is change in time and space of the density and types of observations assimilated. Coupled climate models do not have the same sensitivity to the availability of observations; and the historical runs of coupled climate models were included in this comparison because they seek to simulate the present state and variability of climate around the globe.

Signals in the low-passed fluxes were anticipated. Weller (2015) had found significant linear trends in the first 10 years of Stratus wind stress magnitude $|\tau|$ and Q_{LH}. Annual mean $|\tau|$ increased by 29% from 2000 to 2010 while annual mean Q_{net} decreased 39 W m^{-2}. The increasing wind stress and decreasing Q_{net} trends ended in 2010. The subsequent extension of the Stratus time series showed that linear climate trend estimates from short records can be misleading. It also showed that the atmospheric forcing of the ocean varied significantly over the years, suggesting that the temporal context for limited duration observations and

Table 5. Means of the long-term records of the air–sea fluxes at the three ORS formed by averaging the annual means of the mean daily time series from only the full calendar years used to make the means in Table 4. Freshwater flux is evaporation plus precipitation; evaporation is negative, precipitation is positive; and heat flux and freshwater flux are negative for ocean loss.

Flux	Stratus	WHOTS	NTAS
Wind stress magnitude (N m^{-2})	0.075	0.076	0.081
Wind stress direction (toward, °)	304.9	261.3	256.2
Net heat flux (Q_{net}; W m^{-2})	36.9	24.7	36.2
Latent heat (Q_{LH}; W m^{-2})	–106.9	–135.9	–139.0
Sensible heat (Q_{SH}; W m^{-2})	–7.6	–7.0	–3.9
Net shortwave (Q_{SW}; W m^{-2})	193.9	224.3	231.9
Net longwave (Q_{LW}; W m^{-2})	–42.5	–56.5	–52.8
Freshwater (m yr^{-1})	–1.3	–1.2	–1.3

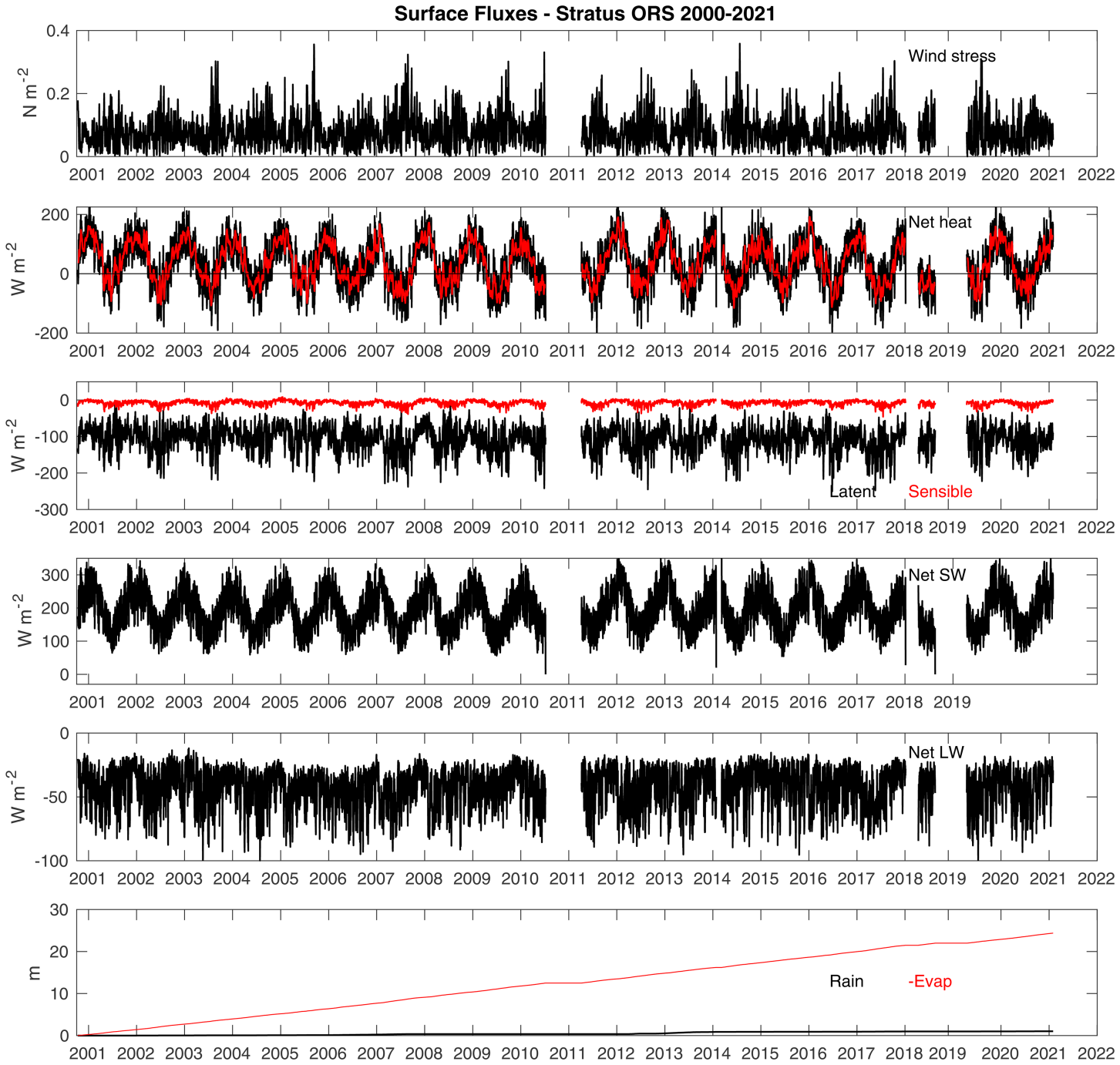

Fig. 4. Daily averaged air–sea fluxes as observed at Stratus. Gaps in the data resulted from breaks in the mooring. (from top to bottom) The magnitude of the wind stress, Q_{net} (black: daily; red: 10-day running mean), Q_{LH} (black) and Q_{SH} (red), Q_{SW}, Q_{LW}, and an overplot of cumulative rainfall and evaporation, respectively.

process studies should be considered in the context of the low-frequency variability in atmosphere–ocean coupling.

In the absence of observed air–sea fluxes, can the reanalysis fluxes provide that context and, more generally, provide accurate low-passed air–sea fluxes with realistic temporal variability? To answer this, reanalysis fluxes were extracted at the grid point nearest to each ORS and daily time series assembled. The three reanalyses were ERA5 (1/4° grid; Hersbach et al. 2020), MERRA-2 (2/3° × 1/2° grid; Gelaro et al. 2017), and NCEP-2 (T62, close to 1/2° grid; Kanamitsu et al. 2002). Figures 5a and 5c show the comparisons of the wind stress and net heat flux time series and the mean values while Figs. 5b and 5d show the model minus ORS

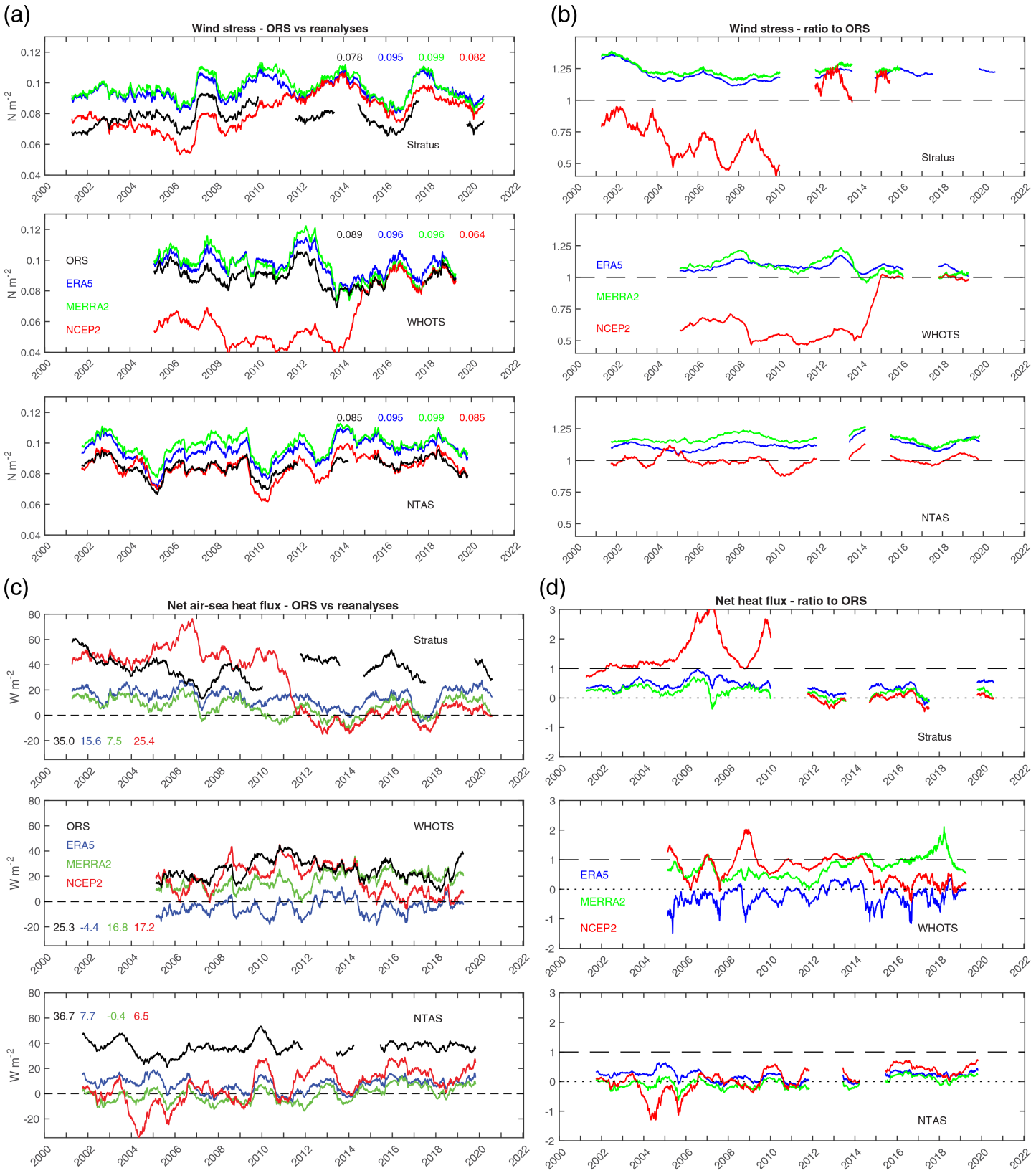

Fig. 5. (a) Three panels comparing low-passed wind stress magnitude at each of the ORS (black) with that at ERA5 (blue), MERRA-2 (green), and NCEP-2 (red). ORS traces start/end half the width of the 365-day running-mean filter from the beginning/end of data segments. Means over the common periods are given in color-coded text. (b) Three panels showing the ratios of the difference between the low-passed reanalyses wind stress and the low-passed ORS wind stress, normalized by the ORS low-passed wind stress. (c) Three panels comparing low-passed net air–sea heat flux at each ORS with the three reanalyses. ORS traces start/end half the width of the 365-day running-mean filter at the beginning/end of data segments. Means over common periods are given in color-coded text. (d) Plots of the ratios of the low-passed net heat fluxes from the reanalyses minus the respective ORS, normalized by the ORS. The dashed line marks a ratio of 1.0; dotted line marks 0.0 to highlight when the reanalyses have the opposite sign to the ORS.

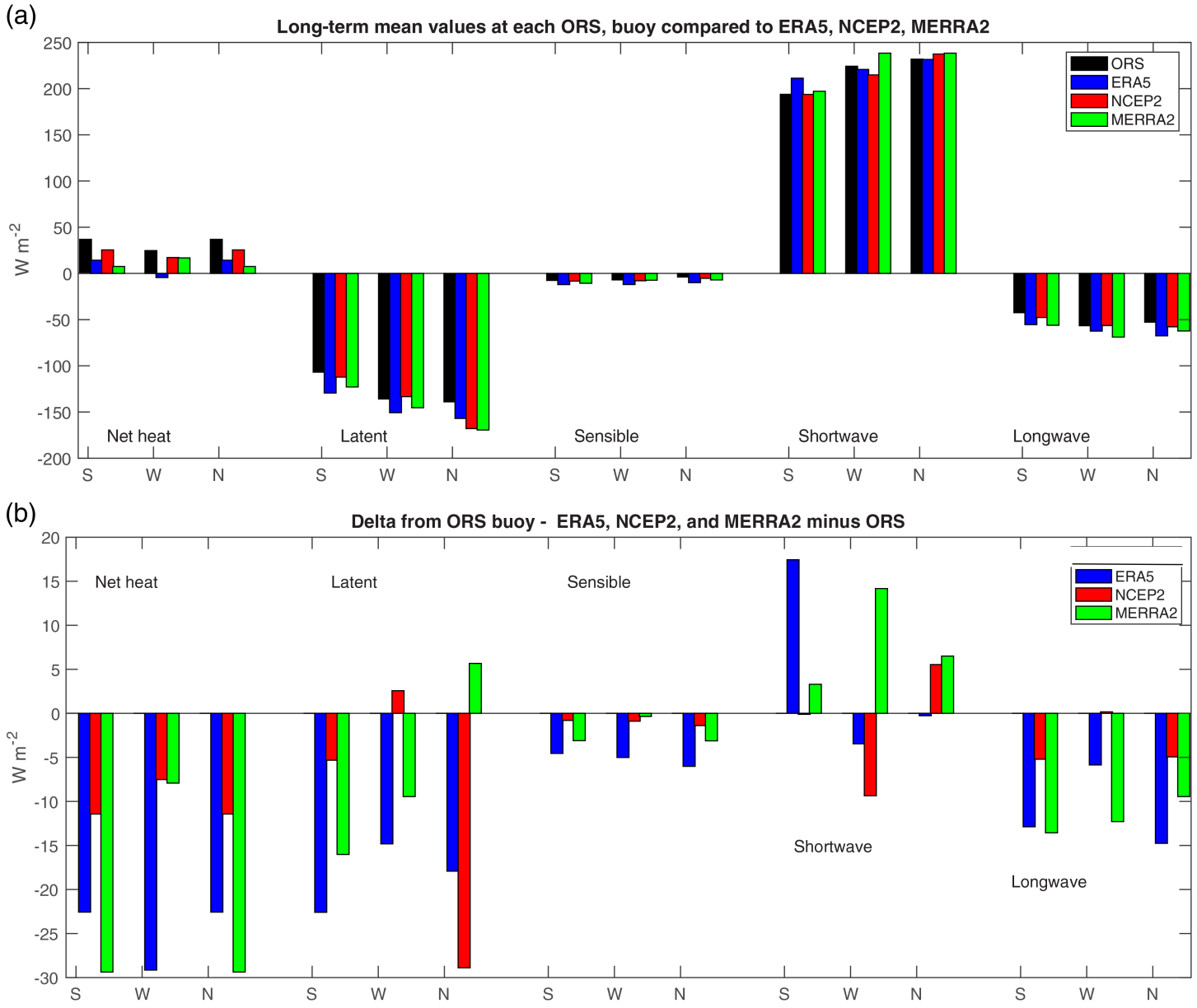

Fig. 6. (a) Comparisons of the long-term means of Q_{net}, Q_{LH}, Q_{SH}, Q_{SW}, and Q_{LW} at the three ORS (black) with the means from ERA5 (blue), NCEP-2 (red), and MERRA-2 (green). On the *x* axis Q_{net} and each component have a cluster of four bars labeled as W (WHOTS), S (Stratus), and N (NTAS). (b) Comparisons of the differences (reanalysis minus ORS) in the long-term means of Q_{net}, Q_{LH}, Q_{SH}, Q_{SW}, and Q_{LW} at the three ORS with the means from ERA5 (blue), NCEP-2 (red), and MERRA-2 (green). On the *x* axis the Qnet and each component have a cluster of three bars labeled as W (WHOTS), S (Stratus), and N (NTAS).

differences normalized by the ORS. Means computed over the time periods common to each ORS and the models are shown.

ERA5 and MERRA-2 stress means are 22% and 27% higher, respectively, at Stratus, 8% higher at WHOTS, and 12%–16% higher at NTAS. NCEP-2 stress mean is 5% higher at Stratus and in agreement at NTAS but its mean is of questionable significance at WHOTS due to a large reduction in bias around 2014. Except for the 2014–15 increase in stress in the NCEP-2 record for WHOTS, much of the low-passed temporal variability of $|\tau|$ is common to the reanalyses and the ORS records. The temporal variability in the model–ORS differences can be seen in the ratio of (model – ORS)/ORS (Fig. 5b). The ability of the models to capture much of the low-passed variability is reflected in the smoother time series of normalized model–ORS difference. ERA5 and MERRA-2 stresses are close to each other, both higher than the ORS. However, though not as dramatic as at WHOTS, NCEP-2 stress moved closer to the other models at Stratus after 2010.

Figure 5c compares the means and low-passed Q_{net} time series, and Fig. 5d shows the ratios of (model − ORS)/ORS. ERA5 mean Q_{net} is biased low at all three ORSs, between 20 and 30 W m^{-2} low in the mean. MERRA-2 is also biased low at all ORS, between 9 and 38 W m^{-2} in the mean. NCEP-2 mean Q_{net} is between 8 and 31 W m^{-2} low. The reanalyses have significantly less heat entering the ocean; at WHOTS the ERA5 mean of −4.4 W m^{-2} and at NTAS the MERRA-2 mean of −0.4 W m^{-2} indicate heat being removed from the ocean. In contrast to wind stress, there is less coherence in the low-frequency variability at each ORS across the models and the observations (Fig. 5d). There is thus more temporal variability in the normalized model–ORS Q_{net} difference time series than seen in the stress difference time series (Fig. 5b). At Stratus, as for stress, the model net heat fluxes converge to each other after 2010, but NCEP-2 shows years-long periods of large (up to >60 W m^{-2}) positive biases relative to the Stratus ORS. The spread among the model–ORS net heat flux differences is greatest at WHOTS with each model showing excursions. The three reanalysis net heat fluxes are closest together at NTAS, and all track with negative bias across the record, with NCEP-2 model–ORS difference showing, early in the record, the largest excursions.

The net heat flux Q_{net} sums the heat gain from Q_{SW} with the Q_{LH}, Q_{SH}, and Q_{LW} losses, and comparing ORS and reanalysis heat flux components shows the source of the Q_{net} differences. Figure 6a shows a bar graph of the long-term mean flux components and the long-term mean Q_{net} for the three ORS. Mean Q_{net} values are small in comparison to the magnitudes of mean Q_{LH} and Q_{SW}, comparable to those of mean Q_{LW}, and larger than those of Q_{SH}. Differencing the means of the reanalyses from the ORS (Fig. 6b) highlights how the components contribute to the low bias in reanalysis Q_{net}. With some exceptions, reanalysis Q_{LH} and Q_{LW} are too large and overestimation of Q_{SW} by reanalyses does not fully offset the too large heat loss of $Q_{LH} + Q_{LW} + Q_{SH}$. There are a few exceptions, such as NCEP-2 and ERA5 Q_{SW} smaller than measured at WHOTS.

Air–sea fluxes can also be extracted from coupled climate models. Climate model runs under different greenhouse gas scenarios are used to predict the evolution, for example, of surface temperature. The model historical runs from which projections are made might be expected to realistically simulate the current era means of the fluxes between the atmosphere and ocean. To examine this, coincident time periods of ORS observations and output from CMIP6 (Eyring et al. 2016) historical runs were selected: for Stratus (18 October 2000–8 April 2018), for WHOTS (13 August 2004–15 September 2018), and for NTAS (31 March 2001–12 June 2018). While the reanalyses assimilate SST data, the coupled models predict SST. To set the context before examining Q_{net}, CMIP6 mean SSTs from each of 51 model runs were compared to the ORS mean SSTs (Fig. 7). The observed ensemble-mean SSTs for the CMIP6 models were 0.72°C warmer than the observed Stratus mean, 0.52°C cooler than the observed WHOTS mean, and 0.65°C cooler than the observed NTAS mean (Fig. 7). The individual model-mean SST biases are generally consistent across the individual models at each site—too cool at WHOTS and NTAS, too warm at Stratus, but there is a spread of several degrees among the model SST biases at each ORS. Extreme individual model biases were 3.26°, −3.02°, and −2.62°C at Stratus, WHOTS, and NTAS, respectively.

The bottom panel of Fig. 7 combines both the net heat flux and SST comparisons to show the different scatter diagram space the models occupy compared to the ORS means. Net air–sea heat flux was only available from 15 models. The means for each model and the model ensemble mean at each ORS were computed. The CMIP6 models' ensemble-mean net air–sea heat fluxes are biased low at all three sites compared to the ORS means. The 15-model-mean Q_{net} (and all but one model) shows ocean cooling at WHOTS and is biased low by −40 W m^{-2} compared to the WHOTS mean. The 15-model-mean Q_{net} is near zero at NTAS, biased low by −36.1 W m^{-2} compared to the NTAS observed mean. The 15-model-mean Q_{net} shows small ocean heating at Stratus and is −30.4 W m^{-2} lower than the observed Stratus mean. There is a scatter of ~20 W m^{-2} in the individual model-mean Q_{net} values at each site. They range from

−18 to −55 W m^{-2} too low and none of the individual model means comes closer than 18 W m^{-2} to the observed ORS means. These 15 models also show ensemble-mean SSTs that are too warm at Stratus and too cool at WHOTS and NTAS. The consistent underestimation of air–sea heat flux at all three sites while two sites have a cold SST bias and one site a warm SST bias suggests a regionality to the performance of the CMIP6 models.

Discussion and conclusions

The purpose of ORS is to provide independent evaluation of reanalysis and climate model rendering of surface forcing of the oceans. The independence is an essential aspect that is achieved by withholding the data from the Global Telecommunication System (GTS) and by obtaining fluxes via direct measurement (radiation and precipitation) or via parameterizations based on independent direct measurement (COARE turbulent flux parameterization). The operational and climate model centers are free to choose how they derive the surface fluxes but ORS fluxes are never tuned to model outputs. The independent ORS fluxes are accompanied by well-documented uncertainty estimates.

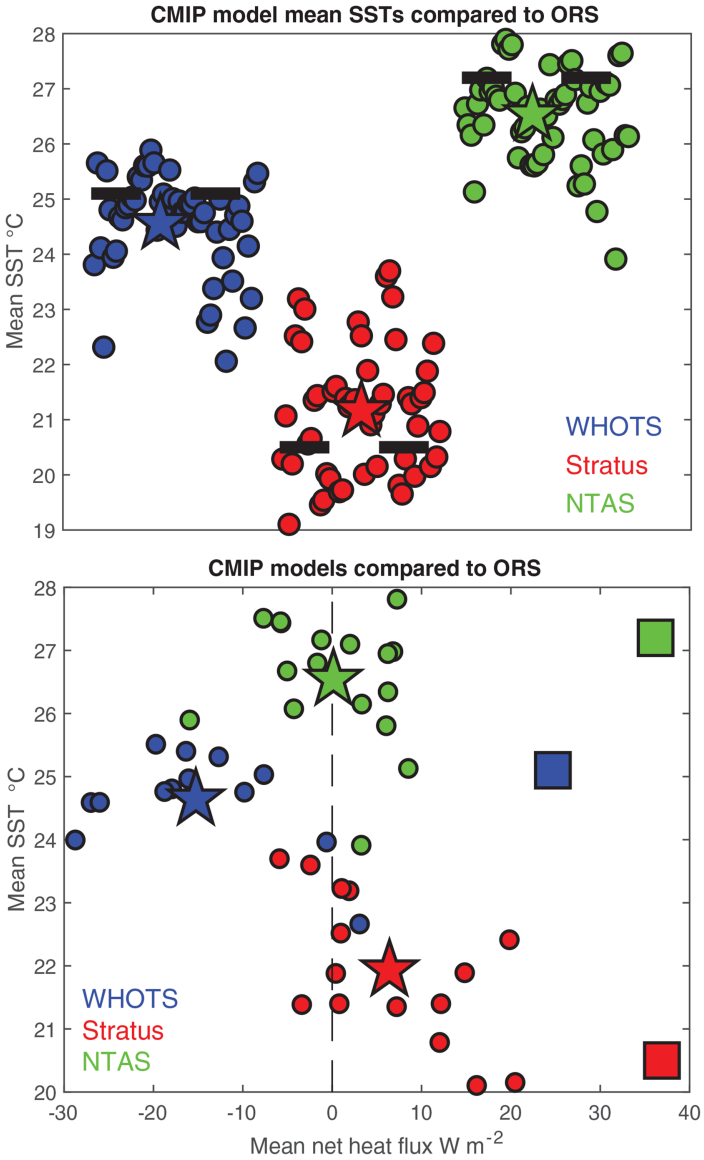

Fig. 7. (top) Comparison of mean SST at the three ORS with the CMIP6 models, historical simulation run. The horizontal axis reflects successive lateral offset displacements when plotting the SSTs from the different CMIP6 models (filled circles) grouped together for each ORS. The dashed lines are the ORS mean SSTs, and the stars are the mean of all model SSTs for that ORS site. (bottom) Comparison of both mean net air–sea heat flux and mean SST between ORS and the 15 CMIP6 models that provided both Q_{net} and SST. Squares are ORS means; stars are means of the individual CMIP6 models (each marked by a filled circle). For both, the ORS site is given by the color.

One of the findings is that reanalysis model–ORS flux differences of long, record-length means and between the low-passed time series are significant, larger than the uncertainties associated with the ORS values. ERA5 and MERRA-2 have wind stresses that are too strong. The ERA5 and MERRA-2 as well as the CMIP6 models provide too little net heat into the ocean. The negative reanalysis Q_{net} bias is characteristic of uncoupled atmospheric models when SST is specified (cf. Bretherton and Battisti 2000). Other investigators have found significant biases in reanalysis fluxes. Zhou et al. (2020) found that Q_{net} from an ensemble of 20 atmospheric general circulation models (AGCM) was biased 31 W m^{-2} low compared to PIRATA and TAO buoys, stemming largely from Q_{LH} and Q_{SH} that were 15 and 7 W m^{-2} too large, respectively. Brunke et al. (2011), using shipboard covariance flux data not dependent on the COARE algorithm,

found NCEP-2 Q_{LH} to be 20.2 W m^{-2} too large, Q_{SH} 0.7 too large, and stress -0.1×10^{-3} N m^{-2} low. MERRA-2 biases were 2.6 W m^{-2} too large in Q_{LH}, −0.8 W m^{-2} smaller in Q_{SH}, and -6.2×10^{-3} N m^{-2} low in stress. Cronin et al. (2006) used observations from the eastern end of the TAO array and COARE 3.0 to compute fluxes and compare them to NCEP-2 and ERA40. Both models had Q_{LH} that was too large in convective regions, with NCEP-2 biased by up to 100 W m^{-2}, leading to Q_{net} with a large negative bias. The CMIP6 coupled model SST bias found here is consistent with the CMIP5 ensemble-mean SST error relative to SST climatology in Zuidema et al. (2016).

A second notable finding is that the annual-averaged reanalysis models' differences from the ORS can vary greatly in time over long time series. Figures 5b and 5d highlight the large change in the NCEP-2 stress bias at WHOTS in 2014–15 accompanied by an increase in the net heat flux bias there, the shift in NCEP-2 Q_{net} at Stratus from a large positive bias before 2010 to a negative bias after 2012, and the different time varying annual biases across the three reanalyses in Q_{net} over the years. As mentioned earlier, the performance of reanalysis models can vary in time due to changes in the types and locations of observations that are assimilated. Robertson et al. (2011) noted how changes in the availability of satellite data impacted MERRA heat fluxes and precipitation. Brunke et al. (2011) also raised the issue that reanalysis performance varies as the satellite data ingested changes. NCEP-2 does not assimilate satellite scatterometer data but ERA5 and MERRA-2 do. Hersbach et al. (2020) summarize the data assimilated into ERA5: scatterometer winds from ASCAT over 2007–14 and OceanSat over 2012–14. Drifting and moored buoy data were assimilated by ERA5 beginning in ~2016. Discussion with NCEP staff (J. Woolen 2022, personal communication) pointed to the possibility that NCEP-2, which does not utilize satellite winds, may be sensitive to additional sources of in situ observations being introduced into the assimilation. The Li et al. (2016) discussion of Hawaii wave hindcasting notes that additional weather buoys were added in the region during 2000–12.

There are issues to be aware of when considering these findings. First, are there other sources of uncertainty in the bulk formulas fluxes other than those that propagate through from the observables? Second, are the ORS sites representative of their regions or possibly located where there are spatial gradients in surface ocean and meteorological fields? Third, to what extent should in situ bulk formulas fluxes be comparable to gridded fluxes from models?

Flux algorithms themselves introduce uncertainty in representing the effects of turbulence near the surface. However, the heritage of the COARE algorithm is that it was developed to quantify the surface fluxes during successful efforts to close the heat budget of the upper ocean (Weller et al. 2004). Since then, it has been validated and improved during joint field deployments of turbulent flux and mean meteorological sensors. The algorithm includes all seven second-order physical effects evaluated by Brodeau et al. (2017). Thus, the COARE algorithm, particularly in the moderate wind wave regime of the trade winds, is not believed to be the source of additional uncertainties in the fluxes. Brunke et al. (2003) used shipboard covariance flux observations along with bulk met variables to evaluate different flux algorithms and gave COARE 3.0 a positive assessment. It agreed within 5% for wind speeds 0–18 ms^{-1} with an ensemble of 26,000 direct flux observations made by three different research groups from ship and buoy platforms (Fairall et al. 2011). Godfrey et al. (1999) in a painfully careful Lagrangian ocean mixed layer study with a ship and a drogued buoy found the net surface flux obtained with COARE and radiative flux measurements balanced the ocean heating within 5 W m^{-2}. Mean bias of a few percent in the algorithm combined with unresolved biases in the ORS measured variables (propagated through the algorithm) implies that differences from mean ORS fluxes of sensible plus latent heat flux of more than about 8 W m^{-2} are significant.

The flux algorithms used in the models differ from the COARE algorithm. Brodeau et al. (2017) noted that the differences stemming from these algorithms can be significant. For example, the National Center for Atmospheric Research (NCAR) flux algorithm (Large and

Yeager 2009) enhances evaporation by 12%–15% compared to COARE 3.0 leading to extra cooling of 13 W m^{-2} on global average and up to 20–30 W m^{-2} in upwelling areas in eastern basins. Some studies have separated the differences between model and in situ fluxes into those stemming from the surface meteorological variables and those stemming from the flux algorithms. Bigorre and Plueddemann (2021) compared 11 years of NTAS surface meteorology and air–sea fluxes with three reanalyses (ERA-Interim, MERRA-2, and NCEP-2). They found long-term mean Q_{net} at NTAS of greater than −40 W m^{-2} in MERRA-2 and NCEP-2 and −34.2 W m^{-2} in ERA. The MERRA-2, NCEP-2, and ERA mean wind stresses were 24%, 4%, and 14% larger than the NTAS mean. However, when the reanalyses surface meteorology was used with the COARE 3.0 algorithm, the model Q_{net} and stresses were in most cases much closer to the ORS. A more detailed diagnosis of the disagreements in surface flux components and bulk variables would likely further illuminate model issues, but that was beyond the scope of this study. Similarly, in the interest of brevity we also limited our analysis to annual and record-length mean variables rather than embark on more sophisticated statistical evaluations (e.g., in Gulev and Belyaev 2012).

Are the ORS representative of their regions? For Stratus, over the years there has been meteorological and air–sea flux observing on the ship steaming to and from the site. De Szoeke et al. (2010) found little spatial change in surface meteorology and fluxes over 10° around Stratus at 20°S, 85°W. De Szoeke at al. (2010) found that coupled models' Q_{net} were typically biased negative compared to observations both east and west of 20°S, 85°W by tens of watts per square meter along with too-warm model SSTs. During a multi-investigator, multiplatform process study, VOCALS-ReX, that focused on an east–west region centered roughly along 20°S extending from the coast to west of Stratus (Wood et al. 2011), modeling and multiplatform sampling confirmed the representativeness of the Stratus ORS for the region away from the coast. Bigorre and Plueddemann (2021) examination of satellite and model fields around NTAS pointed to strong spatial correlation over the region; flux autocorrelations remained above 0.8 to almost 1,000 km in the meridional direction and 2,000 km in the zonal direction. The WHOTS ORS is located at the ALOHA site about 100 km north of Oahu, almost always windward of the Hawaiian Islands given the prevailing trades. Analysis of nine years of AVHRR SST (Dousset 2003) shows no evidence of island effects on seasonal SST averages and cloud frequency in the vicinity of WHOTS. In some cases, as during Kona storms with flow from the southwest, the effect of the orography of Kauai on the winds may affect SST around WHOTS, but these events are very infrequent. Li et al. (2016) show the downstream wake in wind perturbations southwest of the islands, with no discernable upstream influences at ALOHA. Kidd and McGregor (2007) analysis of rainfall over Hawaii and the region around the islands shows that with the dominant trade wind regime there is significant reduction of rainfall to the southwest out to 100 km, in the lee of the islands, while the enhancement of rain on the upwind side is limited to a smaller area, perhaps up to 45 km east and north of the Hawaiian Islands, far from WHOTS.

The comparison of mean fluxes (both turbulent and radiative) from ORS with model outputs raises the normal issues of point-vs-grid-average variables (Hakuba et al. 2013; Wild et al. 2015; Schwarz et al. 2018; Cronin et al. 2019). Time averaging on monthly or longer scales will effectively remove all significant random subgrid-scale variability leaving residual biases of a few watts per square meter at marine sites (Wild et al. 2015). While time averaging will remove random variability, some residual differences can be caused by station placement in a region of local mean gradients. The discussion above supports the representativeness of each ORS of their region. Checks were also done of spatial gradients. Surface solar irradiance is the most inhomogeneous field and Schwarz et al. (2018) show the three ORS stations located in regions of very low spatial sampling bias. While WHOTS is near the Hawaiian Islands, its location under trade winds on the upwind side mitigates

impacts of the nearby islands. Bigorre and Plueddemann (2021) checked the sensitivity of computing and comparing fluxes at NTAS by using the next to closest as well as the closest grid point data from the reanalyses.

How many such ORS are needed and where? A World Climate Research Programme (WCRP) report of a workshop on ocean syntheses and surface flux evaluation (Yu et al. 2013) noted that there were only about 20 fully equipped surface buoys at that time and that that number was insufficient to assess the accuracy of global gridded flux products. That paper also supported maintaining reference stations where data would be withheld from assimilation, including in reanalyses, thus allowing comparisons with independent observations. To do this, they suggested that a specific World Meteorological Organization (WMO) identifier be associated with reference station data to facilitate it being excluded and that the WMO identifiers of all data assimilated into numerical weather prediction models be archived and available. Our outreach to the reanalysis centers returned assurances that the ORS data, which are withheld from the GTS, were not assimilated, though at the same time they noted that a significant level of effort was required at present for retrospective investigations of which observations were assimilated.

The value of long records of high-quality in situ observations has been recognized in other studies. Zhang et al. (2020) pointed to three decades of long-term observations (clouds, aerosols, and radiation) made by the U.S. Department of Energy's Atmospheric Radiation Measurement program that have been used to calibrate climate models and argue that there should be greater utilization of the data by the climate modeling community in diagnosing and improving models and their parameterizations of unresolved processes. Valdivieso et al. (2017) compared monthly fluxes from global ocean and coupled reanalyses with those from moorings, including the ORS. At Stratus they found the ensemble of the models had less ocean heating (~16 W m^{-2}) due largely to excess (~−14 W m^{-2}) Q_{LH}. The excess Q_{LH} stemmed from higher wind speeds, and biases were also found in SST, T_a, and q_a. They advocate for more extratropical buoys to be deployed with cross calibrations between successive deployments to yield long records for use as reference time series.

In recommendations in their *BAMS* paper on coupled climate models in the eastern tropical Pacific and Atlantic Oceans, Zuidema et al. (2016) recommended that buoys be maintained for climate model validation, including at least one in areas characterized stratocumulus clouds, and that they should be equipped to measure all components of the surface energy budget. The buoy data would be used to examine the biases in individual coupled climate models introduced by their difficulties in representing the radiative and turbulent flux components. Cronin et al. (2019) presented a map of 22 existing OceanSITES surface flux sites (now down to 21), recommending that surface flux sites be added in eight new key regions, especially where there is large heat loss and evaporation from the ocean; creation of super sites that would include turbulent covariance flux measurements; and that the ships servicing the moorings be enhanced to monitor the coupled boundary layer. A previous *BAMS* article (Beal et al. 2020) discusses the evolution of the Indian Ocean Observing System and provides an example of both the need for additional surface reference buoys and a rationale for selecting sites. Using six different current surface flux products a map of the standard deviation of the 6 products showed the locations where, reflected by large standard deviation, there were large uncertainties in Q_{net}. These regions of large uncertainty were chosen for deployment of six flux reference sites in order to "provide essential benchmark time series for validating and improving air–sea flux parameterizations in models and for improving uncertainty quantification in air–sea flux products" (Beal et al. 2020).

In our case the deployment of the three ORS reflects our belief that the extraequatorial trade wind regions play an important role in the coupled ocean–atmosphere system. Easterly trade winds cover roughly 50% of the ocean surface, in a wide belt spanning the equator and

extending into the subtropics. Various studies of reanalysis fields and with coupled climate models have highlighted long-term increases in the width, but not strength, of the Hadley cell (Staten et al. 2020) and with mixed results regarding the strength of the near-equatorial Walker circulation component of the trade winds in the Pacific (Zhao and Allen 2019). The data from the three ORS in the trade winds are presented here to make known their availability to supporting assessments of models in the region. The ORS time series allow identification of significant biases in reanalyses and in coupled models. Reanalysis products show temporal changes not seen in withheld observations. Greater care is recommended in using these gridded reanalysis surface fields in the trade wind regions.

Significant biases in SST and Q_{net} were found in the historical run of the CMIP6 models, also suggesting care should be taken in using these models to examine the surface energy balance over the ocean and the evolution of SST. If consistent across the trade winds the biases in model fluxes must accumulate over thousands of kilometers within the westward flowing ocean. Given the models' underestimations of net heat flux into the ocean, the accumulating heat budget errors could only be offset in coupled models by underestimating ocean cooling processes. The SST and Q_{net} biases found in the CMIP6 models points to opportunities to improve those models and understand how they maintain upper-ocean heat balances. We highlight Q_{net} as an essential metric that should be a required variable in all future CMIP efforts.

The oceanographic community is working together including under the UN Decade of Ocean Science for Sustainable Development (www.oceandecade.org) as Observing Air–Sea Interactions Strategy (OASIS; https://airseaobs.org) to increase sampling of surface meteorology and air–sea fluxes over the global ocean. Cronin et al. (2019) discuss at length the need for more ORS-type sites and suggest the present OceanSITES network of 22 fluxes sites should be expanded to cover a total of 19 key regions. As surface meteorological and air–sea flux data become more available, synergistic efforts among the observing, modeling, and remote sensing communities should improve utilization of observations to improve models. We made the choice to withhold ORS data from real-time telemetry and availability on the GTS and have strong beliefs that independent, nonassimilated reference observations are required. However, we have also set up time-delayed file transfers with some modeling centers making ORS data available and in turn obtaining their forecast model output at nearby grid points. We hope the descriptions here of the ORS observing methods, of the production of quality-controlled time series of surface meteorology and air–sea fluxes, the quantification of associated uncertainties, and the example comparisons with atmospheric reanalysis and coupled climate models will motivate investigators to use the ORS data, lead to further interactions between observing and modeling of air–sea exchanges, and to improvements to models.

Acknowledgments. The ocean reference stations have been supported by the National Oceanic and Atmospheric Administration (NOAA) from their inception. Present support for the ORS and RAW, AJP, and SB is from NOAA's Global Ocean Observing and Monitoring Program under CINAR Cooperative Agreement NA14OAR4320158. CWF is also supported by NOAA Global Ocean Observing and Monitoring Program. The three ORS are part of the international OceanSITES program that coordinates collection of time series observations in the open ocean. The WHOTS ORS and RL and JP have been partially supported by the National Science Foundation (NSF), under the Hawaii Ocean Time-series program grants (OCE-0327513, 0926766, and OCE-126014) and by the State of Hawaii. The development of the IMET/ASIMET system reflects the talents of Ken Prada, Geoff Allsup, Dave Hosom, and others. The maintenance of the three ORS depends critically on the skill of the technical staff, now Emerson Hasbrouck, Ray Graham, and others at WHOI and F. Santiago-Mandujano and J. Snyder, Dan Fitzgerald, and students at the University of Hawai'i. The WHOI meteorological sensor calibration facility and sensor testing are done by Jason Smith. Data quality control and analysis support by Kelan Huang has

been essential. Email exchanges with Jack Woolen and Wesley Ebisuzaki at NCEP and with ECMWF staff including Anton Beljaars about differences between the ORS and the reanalyses are acknowledged. The use of the ERA5 dataset from the European Centre for Medium-Range Weather Forecasts is gratefully acknowledged as is the serving of those data by the Copernicus Programme. We wish to thank Xuanyu Chen of PSL and three anonymous reviewers for thoughtful comments on the paper.

Data availability statement. The three ORSs are part of OceanSITES, a component of the Global Ocean Observing System. The data are submitted to the OceanSITES Global Data Assembly Center at the NOAA National Data Buoy Center (http://dods.ndbc.noaa.gov/oceansites), which is mirrored by the Coriolis system at Ifremer (ftp://ftp.ifremer.fr/ifremer/oceansites/). The ORS data (hourly, daily, monthly surface meteorology and air–sea fluxes) are available as merged time series under the Reference Data Sets menu item at http://uop.whoi.edu/ReferenceDataSets/index.html; each of the three ORS also have their own pages on the uop.whoi.edu website. ERA5 hourly data on pressure levels from 1979 to the present were downloaded from Copernicus: https://cds.climate.copernicus.eu/cdsapp#!/dataset/reanalysis-era5-pressure-levels?tab=form. NCEP-2 6-hourly data (from dataset ds094.0), NCEP Climate Forecast System version 2 (CFSv2) 6-hourly products, were downloaded from UCAR: https://rda.ucar.edu/datasets/ds094.0/. NASA MERRA-2 hourly data are available at https://goldsmr4.gesdisc.eosdis.nasa.gov/opendap/MERRA2/M2T1NXOCN.5.12.4, https://goldsmr4.gesdisc.eosdis.nasa.gov/data/MERRA2/M2T1NXOCN.5.12.4, and https://disc.gsfc.nasa.gov/datasets/M2I6NVANA_5.12.4/summary. CMIP6 output was accessed via the NCAR ESM using scripts modified from those provided by the PANGEO project (https://raw.githubusercontent.com/NCAR/intake-esm-datastore/master/catalogs/pangeo-cmip6.json).

References

Albrecht, B. A., 1979: A model of the thermodynamic structure of the trade-wind boundary layer: Part II. Applications. *J. Atmos. Sci.*, **36**, 90–98, https://doi.org/10.1175/1520-0469(1979)036<0090:AMOTTS>2.0.CO;2.

——, 1984: A model study of downstream variations of the thermodynamic structure of the trade winds. *Tellus*, **36A**, 187–202, https://doi.org/10.1111/j.1600-0870.1984.tb00238.x.

——, C. S. Bretherton, D. Johnson, W. H. Schubert, and A. S. Frisch, 1995a: The Atlantic Stratocumulus Transition Experiment—ASTEX. *Bull. Amer. Meteor. Soc.*, **76**, 889–904, https://doi.org/10.1175/1520-0477(1995)076<0889:TASTE>2.0.CO;2.

——, M. P. Jensen, and W. J. Syrett, 1995b: Marine boundary layer structure and fractional cloudiness. *J. Geophys. Res.*, **100**, 14209–14222, https://doi.org/10.1029/95JD00827.

——, and Coauthors, 2019: Cloud System Evolution in the Trades (CSET): Following the evolution of boundary layer cloud systems with the NSF–NCAR GV. *Bull. Amer. Meteor. Soc.*, **100**, 93–121, https://doi.org/10.1175/BAMS-D-17-0180.1.

Beal, L. M., and Coauthors, 2020: A road map to IndOOS-2: Better observations of the rapidly warming Indian Ocean. *Bull. Amer. Meteor. Soc.*, **101**, E1891–E1913, https://doi.org/10.1175/BAMS-D-19-0209.1.

Bigorre, S. P., and N. Galbraith, 2018: Sensor performance and data quality control. *Observing the Oceans in Real Time*, R. Venkatesan et al., Eds., Springer, 243–261, https://doi.org/10.1007/978-3-319-66493-4_12.

——, and A. J. Plueddemann, 2021: The annual cycle of air-sea fluxes in the northwest tropical Atlantic. *Front. Mar. Sci.*, **7**, 612842, https://doi.org/10.3389/fmars.2020.612842.

——, R. A. Weller, J. B. Edson, and J. D. Ware, 2013: A surface mooring for air–sea interaction research in the Gulf Stream. Part II: Analysis of the observations and their accuracies. *J. Atmos. Oceanic Technol.*, **30**, 450–469, https://doi.org/10.1175/JTECH-D-12-00078.1.

Bourlès, B., and Coauthors, 2019: PIRATA: A sustained observing system for tropical Atlantic climate research and forecasting. *Earth Space Sci.*, **6**, 577–616, https://doi.org/10.1029/2018EA000428.

Bretherton, C. S., and D. S. Battisti, 2000: An interpretation of the results from atmospheric general circulation models forced by the time history of the observed sea surface temperature distribution. *Geophys. Res. Lett.*, **27**, 767–770, https://doi.org/10.1029/1999GL010910.

——, S. K. Krueger, M. C. Wyant, P. Bechtold, E. Van Meijgaard, B. Stevens, and J. Teixeira, 1999: A GCSS boundary-layer cloud model intercomparison study of the first ASTEX Lagrangian experiment. *Bound.-Layer Meteor.*, **93**, 341–380, https://doi.org/10.1023/A:1002005429969.

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, https://doi.org/10.1175/JPO-D-16-0169.1.

Brunke, M. A., C. W. Fairall, X. Zeng, L. Eymard, and J. A. Curry, 2003: Which bulk aerodynamic algorithms are least problematic in computing ocean surface turbulent fluxes? *J. Climate*, **16**, 619–635, https://doi.org/10.1175/1520-0442(2003)016<0619:WBAAAL>2.0.CO;2.

——, Z. Wang, X. Zeng, M. Bosilovich, and C.-L. Shie, 2011: An assessment of the uncertainties in ocean surface turbulent fluxes in 11 reanalyses, satellite-derived, and combined global datasets. *J. Climate*, **24**, 5469–5493, https://doi.org/10.1175/2011JCLI4223.1.

Colbo, K., and R. A. Weller, 2009: Accuracy of the IMET sensor package in the subtropics. *J. Atmos. Oceanic Technol.*, **26**, 1867–1890, https://doi.org/10.1175/2009JTECHO667.1.

Cronin, M. F., C. W. Fairall, and M. J. McPhaden, 2006: An assessment of buoy-derived and numerical weather prediction surface heat fluxes in the tropical Pacific. *J. Geophys. Res.*, **111**, C06038, https://doi.org/10.1029/2005JC003324.

——, and Coauthors, 2019: Air-sea fluxes with a focus on heat and momentum. *Front. Mar. Sci.*, **6**, 430, https://doi.org/10.3389/fmars.2019.00430.

Davis, R. E., and Coauthors, 2018: 100 years of progress in ocean observing systems. *A Century of Progress in Atmospheric and Related Sciences: Celebrating the American Meteorological Society Centennial*, *Meteor. Monogr.*, No. 59, Amer. Meteor. Soc., https://doi.org/10.1175/AMSMONOGRAPHS-D-18-0014.1.

de Szoeke, S. P., C. W. Fairall, D. E. Wolfe, L. Bariteau, and P. Zuidema, 2010: Surface flux observations on the southeastern tropical Pacific Ocean and attribution of SST errors in coupled ocean–atmosphere models. *J. Climate*, **23**, 4152–4174, https://doi.org/10.1175/2010JCLI3411.1.

Dousset, B., 2003: Analysis of decadal time series of NOAA-AVHRR images over the Hawaiian Islands. *Proc. 30th Int. Symp. on Remote Sensing of the Environment*, Honolulu, HI, ISRE.

Edson, J., and Coauthors, 2013: On the exchange of momentum over the open ocean. *J. Phys. Oceanogr.*, **43**, 1589–1610, https://doi.org/10.1175/JPO-D-12-0173.1.

Eyring, V., S. Bony, G. A. Meehl, C. A. Senior, B. Stevens, R. J. Stouffer, and K. E. Taylor, 2016: Overview of the Coupled Model Intercomparison Project phase 6 (CMIP6) experimental design and organization. *Geosci. Model Dev.*, **9**, 1937–1958, https://doi.org/10.5194/gmd-9-1937-2016.

Fairall, C. W., E. F. Bradley, J. S. Godfrey, G. A. Wick, J. B. Edson, and G. S. Young, 1996: Cool-skin and warm-layer effects on sea surface temperature. *J. Geophys. Res.*, **101**, 1295–1308, https://doi.org/10.1029/95JC03190.

——, ——, J. E. Hare, A. A. Grachev, and J. B. Edson, 2003: Bulk parameterization of air–sea fluxes: Updates and verification for the COARE algorithm. *J. Climate*, **16**, 571–591, https://doi.org/10.1175/1520-0442(2003)016<0571:BPOASF>2.0.CO;2.

——, T. Uttal, D. Hazen, J. Hare, M. F. Cronin, N. Bond, and D. E. Veron, 2008: Observations of cloud, radiation, and surface forcing in the equatorial eastern Pacific. *J. Climate*, **21**, 655–673, https://doi.org/10.1175/2007JCLI1757.1.

——, and Coauthors, 2011: Implementation of the Coupled-Ocean-Atmosphere Response Experiment algorithm with CO_2, dimethyl sulfide, and O_3. *J. Geophys. Res.*, **116**, C00F09, https://doi.org/10.1029/2010JC006884.

Gelaro, R., and Coauthors, 2017: The Modern-Era Retrospective Analysis for Research and Applications, version 2 (MERRA-2). *J. Climate*, **30**, 5419–5454, https://doi.org/10.1175/JCLI-D-16-0758.1.

Godfrey, J. S., E. F. Bradley, P. A. Coppin, L. F. Pender, T. J. McDougall, E. W. Schulz, and I. Helmond, 1999: Measurements of upper ocean heat and freshwater budgets near a drifting buoy in the equatorial Indian Ocean. *J. Geophys. Res.*, **104**, 13269–13302, https://doi.org/10.1029/1999JC900045.

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

Hakuba, M. Z., D. Folinbi, A. Sanchez-Lorenzo, and M. Wild, 2013: Spatial representativeness of ground-based solar radiation measurements. *J. Geophys. Res. Atmos.*, **118**, 8585–8597, https://doi.org/10.1002/jgrd.50673.

Hersbach, H., and Coauthors, 2020: The ERA5 global reanalysis. *Quart. J. Roy. Meteor. Soc.*, **146**, 1999–2049, https://doi.org/10.1002/qj.3803.

Hosom, D. S., R. A. Weller, R. E. Payne, and K. E. Prada, 1995: The IMET (improved meteorology) ship and buoy systems. *J. Atmos. Oceanic Technol.*, **12**, 527–540, https://doi.org/10.1175/1520-0426(1995)012<0527:TIMSAB>2.0.CO;2.

Kanamitsu, M., W. Ebisuzaki, J. Woollen, S.-K. Yang, J. J. Hnilo, M. Fiorino, and G. L. Potter, 2002: NCEP–DOE AMIP-II reanalysis (R-2). *Bull. Amer. Meteor. Soc.*, **83**, 1631–1643, https://doi.org/10.1175/BAMS-83-11-1631.

Kidd, C., and G. McGregor, 2007: Observation and characterization of rainfall over Hawaii and surrounding region from the tropical rainfall measuring mission. *Int. J. Climatol.*, **27**, 541–553, https://doi.org/10.1002/joc.1414.

Large, W. G., and S. G. Yeager, 2009: The global climatology of an interannually varying air–sea flux data set. *Climate Dyn.*, **33**, 341–364, https://doi.org/10.1007/s00382-008-0441-3.

Li, N., K. F. Cheung, J. E. Stopa, F. Hsiao, Y.-L. Chen, L. Vega, and P. Cross, 2016: Thirty-four years of Hawaii wave hindcast from downscaling of Climate Forecast System Reanalysis. *Ocean Modell.*, **100**, 78–95, https://doi.org/10.1016/j.ocemod.2016.02.001.

McPhaden, M. J., 1995: The Tropical Atmosphere Ocean array is completed. *Bull. Amer. Meteor. Soc.*, **76**, 739–741, https://doi.org/10.1175/1520-0477-76.5.739.

——, and Coauthors, 2009: RAMA: The Research Moored Array for African–Asian–Australian Monsoon Analysis and Prediction. *Bull. Amer. Meteor. Soc.*, **90**, 459–480, https://doi.org/10.1175/2008BAMS2608.1.

Nešpor, V., and B. Sevruk, 1999: Estimation of wind-induced error of rain gauge measurements using a numerical simulation. *J. Atmos. Oceanic Technol.*, **16**, 450–464, https://doi.org/10.1175/1520-0426(1999)016<0450:EOWIEO>2.0.CO;2.

Pinker, R. T., B. Zhang, R. A. Weller, and W. Chen, 2018: Evaluating surface radiation fluxes observed from satellites in the southeastern Pacific Ocean. *Geophys. Res. Lett.*, **45**, 2404–2412, https://doi.org/10.1002/2017GL076805.

Robertson, F. R., M. G. Bosilovich, J. Chen, and T. L. Miller, 2011: The effect of satellite observing system changes on MERRA water and energy fluxes. *J. Climate*, **24**, 5197–5217, https://doi.org/10.1175/2011JCLI4227.1.

Schwarz, M., D. Folini, M. Z. Hakuba, and M. Wild, 2018: From point to area: Worldwide assessment of the representativeness of monthly solar radiation records. *J. Geophys. Res. Atmos.*, **123**, 13 857–13 874, https://doi.org/10.1029/2018JD029169.

Staten, P. W., and Coauthors, 2020: Tropical widening: From global variations to regional impacts. *Bull. Amer. Meteor. Soc.*, **101**, E897–E904, https://doi.org/10.1175/BAMS-D-19-0047.1.

Stevens, B., and Coauthors, 2021: EUREC^{4}A. *Earth Syst. Sci. Data*, **13**, 4067–4119, https://doi.org/10.5194/essd-13-4067-2021.

Sutton, A. J., and Coauthors, 2014: A high-frequency atmospheric and seawater pCO_2 data set from 14 open-ocean sites using a moored autonomous system. *Earth Syst. Sci. Data*, **6**, 353–366, https://doi.org/10.5194/essd-6-353-2014.

Valdivieso, M., and Coauthors, 2017: An assessment of air-sea heat fluxes from ocean and coupled reanalyses. *Climate Dyn.*, **49**, 983–1008, https://doi.org/10.1007/s00382-015-2843-3.

Weller, R. A., 2015: Variability and trends in surface meteorology and air–sea fluxes at a site off northern Chile. *J. Climate*, **28**, 3004–3023, https://doi.org/10.1175/JCLI-D-14-00591.1.

——, 2018: Observing surface meteorology and air-sea fluxes. *Observing the Oceans in Real Time*, R. Venkatsen et al., Eds., Springer, 17–35, https://doi.org/10.1007/978-3-319-66493-4_2.

——, F. Bradley, and R. Lukas, 2004: The interface or air–sea flux component of the TOGA Coupled Ocean–Atmosphere Response Experiment and its impact on subsequent air–sea interaction studies. *J. Atmos. Oceanic Technol.*, **21**, 223–257, https://doi.org/10.1175/1520-0426(2004)021<0223:TIOAFC>2.0.CO;2.

——, S. P. Bigorre, J. Lord, J. D. Ware, and J. B. Edson, 2012: A surface mooring for air–sea interaction research in the Gulf Stream. Part I: Mooring design and instrumentation. *J. Atmos. Oceanic Technol.*, **29**, 1363–1376, https://doi.org/10.1175/JTECH-D-12-00060.1.

Wild, M., and Coauthors, 2015: The energy balance over land and ocean: An assessment based on direct observations and CMIP5 climate models. *Climate Dyn.*, **44**, 3393–3429, https://doi.org/10.1007/s00382-014-2430-z.

Wood, R., and Coauthors, 2011: The VAMOS Ocean-Cloud-Atmosphere-Land Study Regional Experiment (VOCALS-Rex): Goals, platforms, and field operations. *Atmos. Chem. Phys.*, **11**, 627–654, https://doi.org/10.5194/acpd-10-20769-2010.

Yu, L., and R. A. Weller, 2007: Objectively analyzed air–sea heat fluxes (OAFlux) for the global ice-free oceans. *Bull. Amer. Meteor. Soc.*, **88**, 527–540, https://doi.org/10.1175/BAMS-88-4-527.

——, and Coauthors, 2013: Towards achieving global closure of ocean heat and freshwater budgets: Recommendations for advancing research in air-sea fluxes through collaborative activities. WCRP Informal/Series Rep. 13/2013, 42 pp.

Zhang, C., and Coauthors, 2020: The ARM data-oriented metrics and diagnostics package for climate models: A new tool for evaluating climate models with field data. *Bull. Amer. Meteor. Soc.*, **101**, E1619–E1627, https://doi.org/10.1175/BAMS-D-19-0282.1.

Zhao, X., and R. J. Allen, 2019: Strengthening of the Walker circulation in recent decades and the role of natural sea surface temperature variability. *Environ. Res. Commun.*, **1**, 021003, https://doi.org/10.1088/2515-7620/ab0dab.

Zhou, X., P. Ray, B. S. Barrett, and P.-C. Hsu, 2020: Understanding the bias in surface latent and sensible heat fluxes in contemporary AGCMs over tropical oceans. *Climate Dyn.*, **55**, 2957–2978, https://doi.org/10.1007/s00382-020-05431-y.

Zuidema, P., and Coauthors, 2016: Challenges and prospects for reducing coupled climate model SST biases in the eastern tropical Atlantic and Pacific Oceans: The U.S. CLIVAR Eastern Tropical Oceans Synthesis Working Group. *Bull. Amer. Meteor. Soc.*, **97**, 2305–2327, https://doi.org/10.1175/BAMS-D-15-00274.1.