Manuscript prepared for Atmos. Chem. Phys.
with version 3.2 of the LaTeX class copernicus.cls.
Date: 7 November 2013

Near-Surface Meteorology during the Arctic Summer Cloud Ocean Study (ASCOS): Evaluation of Reanalyses and Global Climate Models

Gijs de Boer[1,2], Matthew D. Shupe[1,2], Peter M. Caldwell[3], Susanne E. Bauer[4,5], Ola Persson[1,2], James S. Boyle[3], Maxwell Kelley[5], Stephen A. Klein[3], and Michael Tjernström[6]

[1]The University of Colorado, Cooperative Institute for Research in Environmental Sciences, Boulder, CO, USA
[2]NOAA Earth System Research Laboratory, Physical Sciences Division, Boulder, CO, USA
[3]Lawrence Livermore National Laboratory, Livermore, CA, USA
[4]Columbia University, The Earth Institute, New York, NY, USA
[5]NASA Goddard Institute for Space Studies, New York, NY, USA
[6]Stockholm University, Department of Meteorology, Stockholm, Sweden

Correspondence to: Gijs de Boer
(Gijs.deBoer@colorado.edu)

Abstract. Atmospheric measurements from the Arctic Summer Cloud Ocean Study (ASCOS) are used to evaluate the performance of three reanalyses (ERA-Interim, NCEP/NCAR and NCEP/DOE) and two global climate models (CAM5 and NASA GISS ModelE2) in simulation of the high Arctic environment. Quantities analyzed include near surface meteorological variables such as temperature, pressure, humidity and winds, surface-based estimates of cloud and precipitation properties, the surface energy budget, and lower atmospheric temperature structure. In general, the models perform well in simulating large scale dynamical quantities such as pressure and winds. Near-surface temperature and lower atmospheric stability, along with surface energy budget terms are not as well represented due largely to errors in simulation of cloud occurrence, phase and altitude. Additionally, a development version of CAM5, which features improved handling of cloud macro physics, is demonstrated to improve simulation of cloud properties and liquid water amount. The ASCOS period additionally provides an excellent example of the need to evaluate individual budget terms, rather than simply evaluating the net end product, with large compensating errors between individual surface energy budget terms resulting in the best net energy budget.

1 Introduction

Both modeling and observational studies demonstrate that the Arctic is warming at a rate faster than the rest of the globe (e.g. IPCC, 2007; Serreze et al., 2009; Serreze and Francis, 2006; Rigor et al., 2000). While planetary warming trends are thought to be largely the result of elevated greenhouse

gas concentrations, the "Arctic Amplification" described above is yet to be fully understood. Various ideas have been presented on possible drivers for this Arctic amplification, including feedbacks resulting from changes to snow and ice areal coverage (e.g. Curry et al., 1995), changes to clouds and atmospheric composition (e.g. Kay and Gettelman, 2009; Screen and Simmonds, 2010; Boé et al., 2009), large scale circulation patterns in both the atmosphere and ocean (e.g. Graversen et al., 2008) and natural, low-frequency variability (e.g. Chylek et al., 2009). To date, there is little consensus on which of these processes is most important to understanding Arctic amplification, or even whether Arctic Amplification as described constitutes a robust signal (Polyakov et al., 2002).

Part of the reason for this lack of understanding is a general dearth of observations at Arctic latitudes. Because of this limitation, many of the studies mentioned in the previous paragraph utilize modeling tools such as reanalysis products and global climate models to arrive at their conclusions. Ironically, the reasoning behind the need for using these modeling tools (limited observational records) also creates a challenging environment for model validation and limits the ability of data assimilation techniques to constrain models. In particular, the representation of cloud and radiation processes has been demonstrated to be problematic at high latitudes. Walsh et al. (2008) used measurements from the Atmospheric Radiation Measurement (ARM) program's North Slope of Alaska site at Barrow (71.3°N, 156.6 °W) to evaluate cloud and radiation fields in four different atmospheric reanalyses. These included the National Center for Environmental Prediction (NCEP) and National Center for Atmospheric Research (NCAR) reanalysis (hereafter R-1, Kalnay et al., 1996), the European Centre for Medium-Range Weather Forecasting 40 year reanalysis (ERA-40, Uppala et al., 2005), the NCEP-NCAR North American Reanalysis (NARR, Mesinger et al., 2006) and the Japan Meteorological Agency and Central Research Institute of Electric Power Industry 25 year reanalysis (JRA-25, Kazutoshi et al., 2007). This work illustrated that the reanalyses generally underpredicted area-weighted cloud fraction, resulting in a corresponding over-prediction of downwelling shortwave and an under-prediction of downwelling longwave radiative flux densities at the earth's surface. These differences are alarming in that the measurements were obtained at a location that includes routine radiosonde launches, which are subsequently assimilated into the reanalysis products. This direct integration of local measurements should result in simulated large-scale atmospheric conditions similar to those observed, a luxury not shared by more remote portions of the Arctic. More recently, Zib et al. (2012) used measurements from the Baseline Surface Radiation Network (BSRN) stations at Barrow and Ny-Alesund (78.9°N, 11.9°E) to evaluate cloud and radiative properties in the National Aeronautics and Space Administration (NASA) Modern-Era Retrospective Analysis for Research and Applications (MERRA, Rienecker et al., 2011), the NCEP Climate Forecast System Reanalysis (CFSR, Saha et al., 2010), the National Oceanographic and Atmospheric Administration (NOAA) Twentieth Century Reanalysis Project (20CR, Compo et al., 2011), the ECMWF-Interim reanalysis (herafter ERA-I, Dee et al., 2011), and the NCEP-Department of Energy (DOE) reanalysis (hereafter R-2, Kanamitsu et al., 2002). This next generation of reanalyses demonstrates large

differences in cloud occurrence from one product to the next, including inconsistencies between relative cloud amounts from one site to the other. Radiative flux densities, while marginally better at Barrow, are still demonstrated to be problematic, with R-2 demonstrating the largest biases. Most recently, Jakobson et al. (2012) evaluated the performance of several reanalysis products over the central Arctic Ocean. Using measurements from the Tara drifting ice station (Gascard et al., 2008; Vihma et al., 2008) the authors demonstrate that the ERA-Interim reanalysis outperforms several others, including R-2 and MERRA, using a ranking system. These rankings were calculated through comparison of the analysis results from these products for fields including air temperature, specific humidity, relative humidity and wind speed.

Global climate models have similarly struggled with simulation of Arctic surface meteorology, clouds, and surface radiation. Walsh et al. (2002) demonstrated that while early climate models produced seasonal cloud cycles that were similar to observational estimates, the amount of cloud cover varied dramatically from one model to the next. As may be expected, these discrepancies led to major differences in radiation, with summertime differences of nearly a factor of three in the surface radiative budget. A broader look at models involved with the 3rd Coupled Model Intercomparison Project (CMIP3) by Svensson and Karlsson (2011) illustrated large variability among the models in their representation of net surface energy flux on an Arctic-wide scale ($> 66.6°$N) for present-day wintertime conditions. These discrepancies make deriving concrete results for the causes behind Arctic amplification challenging, particularly given the limited model validation efforts that have taken place. More recently, de Boer et al. (2012) performed an evaluation of 20th century simulations completed using the Community Climate System Model version 4 (CCSM4). They demonstrated that Arctic clouds, including their phase partitioning and liquid and ice water paths were grossly misrepresented in CCSM4, with notable impacts on the surface energy budget. Interestingly, these errors appeared to have only small impacts on the simulated surface air temperature, with model cold biases on the order of 1-2 K when compared to ERA-40, which itself has been shown to have 1.5 K warm biases compared to International Arctic Buoy Programme/Polar Exchange at the Sea Surface (IABP/POLES) measurements (Liu et al., 2007). Similar cloud and radiation discrepancies were brought out in work by Inoue et al. (2006) and Tjernström et al. (2008) while evaluating regional climate model performance over the Surface Heat Budget of the Arctic Ocean (SHEBA) Experiment site.

In the current work, we evaluate some of the tools described above using measurements obtained during the Arctic Summer Cloud Ocean Study (ASCOS, Tjernström et al., 2013). The ASCOS dataset provides us with a unique opportunity to assess the performance of models at high latitudes. Evaluations of basic surface meteorology and the surface energy budget as represented in atmospheric reanalyses and global climate models are completed using this dataset. Because this is a multi-global-model evaluation, with only limited variables and output formats available, averaging and interpolation of model results is necessary at times to ensure the most fair comparison possible.

Not all variables are handled in the same way, and therefore these actions are outlined in the individual result sections as necessary. Section 2 provides a brief overview of ASCOS, section 3 provides some background on the models, section 4 outlines results of the evaluation, and finally section 5 provides discussion of these results and a summary.

2 ASCOS

During August of 2008, the Swedish ice breaker *Oden* served as a drifting base camp for the Arctic Summer Cloud Ocean Study (ASCOS, Tjernström et al., 2013). The vessel headed north from Svalbard with a goal to spend as much time as possible in the central Arctic ice pack. Between 12 August and 2 September, Oden was moored to a 3x6 km ice floe just north of 87°N latitude (see Figure 1). Much of the analysis performed in the current work is completed for this drifting time period. During this time, the expedition first experienced a ten day period where surface air temperatures ($T_{air,sfc}$) were near, or slightly above the freezing point and low cloud cover was abundant. This was followed by a slightly colder period with $T_{a,sfc}$ between 266 and 272 K. Finally, starting around 29 August, temperatures fell further (below 263 K) and open water began to close with the initiation of the fall freeze-up of sea ice.

A wide variety of instrumentation was deployed during ASCOS in order to comprehensively sample the atmosphere and surface. Included were a series of in-situ instruments measuring properties such as air temperature, air pressure, relative humidity, and wind speed and direction. This includes sensors at the surface as well as sensors used for profiling, with profiles obtained four times daily via radiosonde. In addition to in-situ instrumentation, surface based remote sensors were deployed to measure clouds and radiation. These included a 35-GHz millimeter cloud radar (MMCR, Moran et al., 1998), ceilometers, a scanning 60-GHz radiometer, a dual wavelength (24/31 GHz) microwave radiometer (MWR, Westwater et al., 2001), an S-Band cloud and precipitation radar, a 449-MHz wind profiler, a suite of broadband radiometers measuring surface short- and longwave radiation flux densities and turbulence masts used to estimate turbulent energy fluxes. A description of the surface energy budget measurements, along with error estimates can be found in Sedlar et al. (2011).

When combined, this suite of instruments provide an in-depth view of atmospheric processes. In addition to derivation of cloud macrophysical properties (height, thickness, fractional coverage, etc.), information on cloud phase is obtainable via methods described in Shupe (2007). Additionally, estimates of cloud liquid water path (LWP) are obtained from the MWR, with an estimated error of approximately 25 g m^{-2} (Westwater et al., 2001).

Unlike several other large measurement campaigns over the Arctic Ocean, radiosonde measurements from ASCOS were not submitted to the Global Telecommunications System (GTS, Birch et al., 2012), meaning that they were not included in the datasets used for model initialization and

data assimilation. Conversely, 6-hourly surface-based wind and pressure observations from the Oden were submitted to the GTS. While the one month observation period does not provide us with an long-term record against which to evaluate our models, this data set does represent one of the longest-lasting, comprehensive sets of high-Arctic measurements not included in the GTS, to our knowledge. Therefore, ASCOS measurements represent an independent data set that can be used to directly evaluate reanalysis model performance.

3 Models Evaluated

3.1 ERA-Interim

The European Center for Medium Range Weather Forecasting (ECMWF) has released multiple reanalysis products. Here, we evaluate the performance of the most recent version, the ERA-Interim (hereafter ERA-I, Dee et al., 2011). ERA-I provides global analyses of atmospheric and surface state variables from 1989 to the present. It builds upon the successful ERA-40 (Uppala et al., 2005) product which covered the years between 1957-2002, with notable differences. ERA-I extends the number of atmospheric pressure levels archived from 23 to 37, is run at a higher resolution (T255, approximately 0.7°) and includes a number of additional cloud and state variables in its output. Advances in data assimilation techniques include the introduction of a 12-hourly 4D-variational (4D-VAR) scheme, improved formulation of background error constraint, improved humidity analysis, improved model physics, and improved quality control, to name a few. Observationally, ERA-I utilizes all of the observational datasets from the ERA-40 project, and adds altimeter wave height information, winds and clear sky radiances from EUMETSAT, ozone profiles and radio occultation measurements. Boundary forcing fields come from a combination of ERA-40 reanalysis output (before 2001) and the ECMWF operational analysis (after 2001).

3.2 NCEP-NCAR (R-1)

The National Centers for Environmental Prediction (NCEP) teamed up with the National Center for Atmospheric Research (NCAR) to produce a reanalysis product to support research and climate monitoring communities (NCEP-NCAR, hereafter R-1, Kalnay et al., 1996). Originally planned to cover the 40-year period from 1957-1996, the project was later expanded to include years up to present day. The underpinning model's horizontal resolution (T62, approximately 1.9°) is significantly lower than that of ERA-I, as is the number of vertical levels (28 for R-1). This results in a lower-atmospheric vertical resolution ranging from 80 m (lowest grid box) to 384 m (7th grid box at 850 mb). Examples of assimilated datasets include a variety of satellite-based measurements, radiosondes from the GTS, sea ice characteristics from the ECMWF, and surface ocean data. R-1 utilizes a 3D-variational (3D-VAR) analysis scheme known as a spectral statistical interpolation. Unlike for ERA-I, forecast products are not archived for R-1 beyond the 6 hour forecast time. These

6 hour forecast fields are in addition to analysis time values provided for limited variables.

3.3 NCEP-DOE (R-2)

To improve upon R-1, NCEP teamed up with the Program of Climate Model Diagnosis and Intercomparison (PCMDI) at Lawrence Livermore National Laboratory (LLNL) to create the NCEP-DOE AMIP-II reanalysis (hereafter R-2, Kanamitsu et al., 2002). R-2 utilizes the same spatial and temporal resolution as R-1 (T62, 28 levels, 6 hours), and makes use of similar raw observational datasets. Differences in the datasets include the removal of data from the Special Sensor Microwave Imager (SSM/I), and the addition of limited additional data after 1993. Additionally, errors pertaining to bogus data in the southern hemisphere, snow cover analysis, humidity diffusion, oceanic albedo, relative humidity discontinuities and snowmelt were fixed for the R-2 product. New system components included in R-2 include rainfall assimilation over land to improve surface soil moisture, a smoothed orography and an updated treatment of snow. Other differences result from improvements to model physics made between the creation of R-1 and R-2. These include a new planetary boundary layer scheme, new shortwave radiation scheme, a retuned convective parameterization, improved cloud-top radiative cooling, updated cloud-tuning coefficients and further improvements to the radiation scheme. As with R-1, only 6-hour forecasts are distributed, along with limited analysis fields.

3.4 CAM5

Another set of simulations evaluated in this study were completed using a recent version of the Community Atmosphere Model (CAM5.1, Neale et al., 2010). In order to recreate conditions from the ASCOS period, these simulations were completed using the Cloud-Associated Parameterizations Testbed (CAPT, Phillips et al., 2004). CAPT utilizes operational analyses from numerical weather prediction centers to initialize CAM5 and produce short term forecasts. In this instance, the European Center for Medium-Range Weather Forecasts (ECMWF) Year of Tropical Convection (YOTC) analysis was used to initialize forecasts within CAPT. The analysis data are interpolated from the finer-resolution analysis grid of 0.150° and 91 levels to the CAM5 grids using procedures outlined in Boyle et al. (2005). These procedures use a slightly different interpolation approach for each of the dynamic state variables (i.e. horizontal winds, temperature, specific humidity and surface pressure), along with careful adjustments to account for the difference in representation of the earths topography between models. A series of 6-day hindcasts are initialized every day at 0000 UTC from the ECMWF analysis for the entire YOTC period from 1 May 2008 to 30 April 2010. Only the atmospheric winds, temperature and moisture are initialized, while the rest of the initial variables (land and atmosphere) come from an additional ECMWF-nudged run of the same model. Skin surface temperature and sea ice are prescribed using the NOAA Optimum Interpolation (OI) Sea Surface Temperature (SST) V2. These data are weekly means on a 1°x1° grid and are interpolated in time

from weekly to the model time step. Since the model has a spin-up period to adjust to ECMWF conditions, the ASCOS time series are created by concatenating hours 24-48 from each hind cast.

CAM5 was run with the finite volume dynamic core at resolution of 0.90°x1.250° in the horizontal and utilizes 30 vertical levels. This version of CAM contains a range of significant enhancements and improvements in the representation of physical processes. Except for the deep convection scheme, most other physical parameterizations have been updated from CAM4 to CAM5 (Neale et al., 2010). In addition, a three-mode modal aerosol scheme (MAM3) has been implemented in CAM5 to provide internally mixed representations of number concentrations and mass for Aitken, accumulation and coarse aerosol modes (Liu et al., 2011). These major physics enhancements permit new research capability for assessing the impact of aerosol on cloud properties. In particular, they provide a physically based estimate of the impact of anthropogenic aerosol emissions on the radiative forcing of climate by clouds.

We also analyze simulations from a development version of CAM5 which includes several changes expected to improve model clouds (CAM5-PF). In particular, this new simulation includes a new parameterization for cloud macrophysics (which combines stratiform liquid cloud fraction, condensation, and evaporation) based on a truncated Gaussian PDF for sub-grid variability in saturation excess (defined as total water mixing ratio minus ice mixing ratio minus liquid saturation mixing ratio). This truncated Gaussian also replaces the Gamma sub-grid scale distribution previously used for microphysics. This improves inter-process consistency and slightly reduces microphysical depletion rates. Further improvements include sub-stepping macro- and microphysics to improve coupling between condensational growth and microphysical erosion of cloud and fixing an inconsistency between the liquid water content and droplet number used by microphysics. The latter change distributes liquid across more droplets (further reducing microphysical depletion), while the former keeps parameterizations from bouncing liquid between physically unrealistic states. These collective changes are expected to increase high-latitude cloud fraction and liquid water path, while handling microphysical depletion more appropriately.

3.5 NASA GISS-ModelE2

GISS-ModelE2 simulations were completed using an updated version of the NASA GISS GCM ModelE2 specifically developed for the 5th IPCC assessments (CMIP5). The CMIP5 version of the GISS-ModelE2 is improved over that used for CMIP3 (and described in Schmidt et al. (2006) and Hansen et al. (2007)) in a number of respects (Schmidt et al., in preparation). Firstly, the model has a higher horizontal and vertical resolution (2 lat x 2.5 longitude, 40 layers). The vertical layers are distributed on a non-uniform grid, with spacing of roughly 25 mb (250 m) from the surface to 850 mb, and roughly 40-50 mb (400-700 m) from 850 to 415 mb. Secondly, various physics components have been upgraded from the CMIP3 version, namely the convection scheme, stratiform cloud scheme and gravity wave drag. The simulations discussed here further include the aerosol

microphysics scheme MATRIX (Bauer et al., 2008).

For this work the model is run continuously for 2008, covering the ASCOS campaigns described above. In order to force representative meteorology in the GCM, the model uses prescribed sea surface temperatures and sea ice, and the horizontal wind components of the model are nudged towards the R-2 reanalysis. R-2 winds are available on a 6-hourly time step and are linearly interpolated to the model 30-minute timestep. It is important to recognize that this nudging may introduce systematic errors in these runs tied to any biases present in the analysis used for the nudging. However, it is believed that the nudging is gentle enough that most biases are connected to the physics of the GISS-ModelE2 itself. The aerosol scheme uses the CMIP5 emissions by Lamarque et al. (2010). This setup has previously been used by Bauer and Menon (2012) and de Boer et al. (2013) to evaluate forecast-mode simulations using ModelE.

4 Notes on Sampling

A major consideration in evaluations such as in this study is how to best analyze available measurements to appropriately represent the scales inherent to model grid boxes (McComiskey and Feingold, 2012). While in-situ and surface-based observations such as those obtained during ASCOS have the potential to capture process-level relationships, they do not necessarily capture the spatial variability included within a model grid box without alteration of obtained measurements. A simple approach is aggregation (averaging) of data over time scales that begin to capture the spatial variability in the model grid. Naively, it may be assumed that it would be appropriate to average over a time period that covers the full scale of the grid box assuming some advective velocity (e.g., 10 m s^{-1}). At 2°, this requires averaging of periods on the order of 6-7 hours. Using this technique is not practical, however, as it blurs the evolution of quantities occurring within the diurnal cycle. An alternative approach entails averaging over shorter periods (e.g., one hour) in order to capture some of the sub-grid scale variability in the measurements, while maintaining signals inherent in an evolving atmosphere. This short aggregation timescale is very appropriate for time periods featuring consistent large scale meteorological conditions and a relatively homogeneous surface, but may fail during frontal passages or at coastal sites. In this work, comparison is additionally complicated by the way in which parameters are presented within the reanalysis products. For example, while most variables are provided on a 6-hourly timescale, some of the values presented (e.g. liquid water path, cloud fraction) represent instantaneous values for that time (but still averaged across the model grid box), other variables (e.g. precipitation, radiation) are provided as average values over the 6-hour period. Complicating things further, the observational datasets used for this study represent different time scales as well, with most quantities aggregated into hourly averages, and others represented as 3-hourly averages. In this evaluation, we have done our best to be consistent in our comparisons. For variables represented in the reanalyses as averages, we have taken the time period over which this

average is taken and computed averages from the measurements and GCM output as well. Likewise, for instantaneous values, we have interpolated the measurements and GCM results to most closely match the time of the reanalysis dataset. The (at least) hourly averaging used on the measurements ensures that some spatial variability is accounted for in those datasets, even when "instantaneous" evaluations are done since the highest temporal resolution used is a one-hour average. Further details are provided in the individual sections in Section 5.

5 Results

5.1 Surface Meteorology

Figure 2 gives an overview of the performance of different modeling products in simulating basic surface meteorological quantities for the time that Oden was drifting with the pack ice north of 87°N latitude. Included are (from top to bottom) 2-meter air temperature ($T_{air,sfc}$, K), surface air pressure ($P_{air,sfc}$, mb), 2-meter air relative humidity ($RH_{air,sfc}$, %), the 10-meter zonal wind component (U, m s^{-1}) and the 10-meter meridional wind component (V, m s^{-1}). For all variables, the model results have been linearly interpolated in space to the exact location of the Oden at a given time, reducing (but not eliminating) the influence of resolution that would result from a nearest grid cell comparison, and eliminating the influence of jumping between grid boxes with the movement of Oden during the campaign. The time series represented in Figure 2 provide a comparison between the observations, reanalyses and GCMs. Biases are calculated and presented as distributions in the right hand column of Figure 2, with black circles representing the median difference, the box representing the interquartile range (IQR) and the whiskers representing the extent of the 5th and 95th percentiles of the biases. Additionally, values for median biases and correlations for all evaluated variables are presented in Table 1. For the variables compared here, the reanalyses provide these variables as instantaneous values at the 6-hourly intervals. In order to most closely compare these quantities between the GCMs (which provide 1 hour averages and 3-hourly instantaneous output for CAM5 and GISS, respectively), the observations (which are provided as one-minute averages) and the reanalyses, observations are averaged on a one-hour scale to account for sub grid variability within the models, and then interpolated, along with the GCM output, to the 6-hourly reanalysis times.

Looking first at $T_{air,sfc}$ (Figure 2[a]) , it is apparent that there is quite a bit of variability from one product to the next, and that there does not appear to be a clear "best" model in terms of agreement with observations. Statistically, R-2 outperforms other models in terms of the median bias, though it is clear that this bias is derived from a distribution with significant variability. ERA-I is correlated much more closely with the observations (correlation of 0.82) than the other models are. As noted above, the first portion of this period (roughly 12-17 August) features $T_{air,sfc}$ near the freezing point. This is captured by the models, with most within 1-2 degrees of the observations. There is a

short period of cooler temperatures on 16 August which only the CAM5 models appear to simulate well. Unlike the original CAM5 version, CAM5-PF $T_{air,sfc}$ does not recover back to the freezing point as quickly as in the observations, taking nearly two days to recover. Beyond 17 August, the observations, ERA-I and R-1 remain in this near-freezing point state for another four days. The ModelE2 has a cooler period from 17-19 August, while CAM5, CAM5-PF, and R-2 all cool off quickly on 18 August. Both versions of CAM5 cool to temperatures around 265 K by 20 August, which is cooler than the observed $T_{air,sfc}$ decrease to 267 K on 21 August. R-2 captures this cool period most successfully, with ERA-I, R-1 and ModelE2 all remaining too warm. The time period from 24 August until 31 August is an interesting one in that the observations along with R-1, ModelE2 and ERA-I reanalyses all feature temperatures that are only slightly below the freezing point, while CAM5 and R-2 generally have much colder (around 265 K) temperatures and CAM5-PF oscillates between these states. Physical mechanisms driving these differences will be explored later in the paper.

All of the model products generally perform well in simulating synoptic scale weather phenomena. Figure 2[b] illustrates a comparison of $P_{air,sfc}$ at Oden between the different products. With the exception of R-1 and R-2, differences from the observations are generally small. All products capture the general pattern, with $P_{air,sfc}$ generally increasing between 13 August and 2 September. Looking at the right-hand column of Figure 2[b], the R-2 6-hour forecast demonstrates a clear high bias in $P_{air,sfc}$ (4.93 mb), while the R-1 6-hour forecast demonstrate a low bias (-3.01 mb). Biases at all forecast scales for ERA-I , GISS ModelE2 and CAM5 are generally at 1 mb or less, which should not be surprising since surface pressure measured aboard Oden was submitted to the GTS. One interesting point to note is that CAM5-PF does feature larger variability in the $P_{air,src}$ errors than CAM5. With $P_{air,sfc}$ generally represented well, it is not surprising that simulation of 10 meter winds was also quite good. Both the zonal and meridional winds (Figure 2[d,e]) produced by CAM5, GISS ModelE2 and reanalysis products follow the observed winds closely. The only clear exceptions to this are the CAM5-PF errors between 17 August and 20 August and 24 August to 30 August which correspond with $P_{air,sfc}$ biases in that model. GISS ModelE2 meridional winds also are biased high, likely the result of forcing from R-2, which demonstrates a similar high bias. The right-hand column of Figure 2[d,e] illustrates that with the exception of CAM5-PF and GISS ModelE2 meridional winds, differences between simulated and observed winds were generally around 1 m s^{-1} or smaller, and that mean differences fall very close to the zero line. ERA-I featured the highest correlations in both the zonal and meridional winds (.92 and .91, respectively), and also was very well correlated to the surface pressure (.99).

Finally, looking at the simulation of $RH_{air,sfc}$ (Figure 2[c]), the models produce values that are at times very different from observed values. ERA-I produces $RH_{air,sfc}$ values near 100 % for the entire observation period. There are periods of time where this is the correct solution, though the observations feature significantly more variability than ERA-I does. R-1 tends to be drier than

the observations, particularly during the period from 21-24 August where the model has $RH_{air,sfc}$ values between 75-90 % and observations hover between 95-100 %. R-2 features a similar drop in $RH_{air,sfc}$, though it is not as large, and, unlike R-1, is close to observations during the period from 24 August to 2 September. CAM5 and CAM5-PF both feature less variability than the observations, but feature values that fall in the middle of the observed variability. Ultimately, this works to reduce the median bias for these models (Figure 2[c], right-hand column), and ultimately result in some of the lowest biases of all of the models with poor correlation to the observations (Table 1).

5.2 Clouds and Precipitation

ASCOS was in general very cloudy, a fact reflected in the observed cloud fractions (CF) in Figure 3[a]. Three-hourly averaged values were at 100% for most of the campaign, with only a few time periods in early September where values dropped significantly. As with surface meteorology, the models and observations provide different quantities for reporting cloud fraction. ERA-I, R-1 and R-2 all provide an instantaneous value every six hours, while GISS ModelE2 provides a 3-hourly instantaneous value. The observations are provided as a 3-hourly average, while CAM5 provides 1-hourly averages. In order to attempt to make a more consistent comparison, CAM5, GISS and observational values are interpolated to the reanalysis times without averaging further. While CF is only a rough means for evaluating model performance in cloud simulation, the model results do not compare favorably with the observations. ERA-I featured CFs that were closest to observations, with only a couple of periods where the CF fell below 100%. The fact that ERA-I only reports instantaneous values at the 6-hourly times, while the other models (and observations) report 6-hourly averages, does make it likely that some differences will occur. During times where ERA-I CF did stray from 100%, it was not by much, falling only to values around 70-80%. Towards the end of the measurement period, ERA-I retains 100% CF, while the observations feature less clouds. In comparison, the other models generally feature less clouds than both the observations and ERA-I. Both R-1 and R-2 consistently produce far too few clouds. This is particularly true for the period between 26-29 August, where low, stratiform clouds persisted in the observations but both NCEP products have cloud fractions near 10 %. ModelE2 features significant variability in it's representation of cloud fraction, spanning values between 0-100 %. The two CAM5 versions both under-predict cloud cover, and while the original CAM5 version has fewer very low values, CAM5-PF performs better during the previously mentioned stratiform cloud period during the end of August. As will be explored in more detail in the following section, this increased cloud cover is generally responsible for CAM5-PF's warmer surface temperatures during this time period when compared with CAM5 (Figure 2[a]). Biases for the GCMs are shown to be comparable, with all three models underestimating cloud fraction by roughly 20 %, though the timing of these differences becomes quite important in the overall impact of this bias on surface energy balance and meteorology.

Cloud fraction only provides us with a limited perspective on the performance of models in sim-

ulating clouds. Perhaps more important, at least from the point of view of surface radiation, is the amount of liquid water contained within these clouds. Unfortunately, not all of the models include the simulated liquid water path in the publicly available output. For those that do, ERA-I provides 6-hourly instantaneous values, GISS ModelE2 provides 3-hourly instantaneous values, CAM5 provides 1-hourly averages and the observations are available as 15-second averaged values. To improve the comparison, observational estimates were averaged on a 1-hourly basis to account for spatial variability within model grid boxes. These averages, along CAM5 1-hourly averages and GISS ModelE2 3-hourly instantaneous values were interpolated to the times available for the reanalyses. Figure 3[b] presents the time series of liquid water path for the models that provide this quantity (ERA-I, CAM5, CAM5-PF and ModelE2) as well as the observations. The observations presented here represent a lower bound on LWP due to the need to correct data for time periods with liquid precipitation. Here, we replace the unrealistically high values reported during these times with the value reported directly before precipitation starts. This likely results in an observational underestimate of the LWP during these times. As with cloud fraction, ERA-I generally matches the observations closely. While differences still exist, ERA-I and GISS ModelE2 feature liquid water for most of the observation period. Both versions of CAM5 have multiple periods where very little liquid water exists. Improvements made for CAM5-PF present themselves clearly in the representation of LWP from 23-28 August, with CAM5-PF featuring liquid-containing clouds during this stratiform cloud period. As will be shown in the following section, this increase in LWP during those dates results in improved simulation of short- and longwave radiation during those dates as well, ultimately resulting in a more accurate depiction of the surface energy budget and near surface temperatures.

Precipitation rates are in general challenging to model and observe. Precipitation during the period of observation for this quantity was generally very light, with rates between 0-0.2 mm hr^{-1}. With the exception of GISS ModelE2, which provides 3-hourly instantaneous precipitation rates, all of the other models provide averaged quantities. All of the reanalyses provide 6-hourly mean precipitation, while the observations and CAM5 are both available as 1-hourly means. In order to bring these estimates closer together, the observations and the GCMs are averaged over the same 6-hourly period for which averages are provided in the reanalyses. The models fail, in general, to reproduce the more significant precipitation events observed (e.g. 17 August, 20 August, 23 August, 29 August). There does not appear to be a clearly superior model in terms of reproducing these light precipitation events. The right hand column of Figure 3[c] demonstrates the challenges models appear to have with accurate representation of precipitation. While all of the mean biases are relatively small, the variability in the biases is quite large. This indicates not only the inability of models to correctly simulate the magnitude of the precipitation, but also the issues occurring with respect to correctly timing the precipitation as demonstrated by the relatively low (and statistically insignificant) correlations in Table 1. Most of the models feature light precipitation throughout the

observation period that is greater than the observed precipitation for much of that time. Additionally, some of the models (R-1 on 18 August; CAM5-PF on 23 and 25 August) produce more significant precipitation events during times where very little precipitation was observed.

5.3 Surface Energy Budget

Figures 4 and 6 provide an overview of the surface energy budget terms as governed by:

$$Q_{SFC} = F_{LW} + F_{SW} - F_{SH} - F_{LH} \quad (1)$$

where F_{LW} and F_{SW} represent the net longwave and shortwave radiative fluxes, respectively, F_{SH} represents the surface sensible heat flux, F_{LH} represents the surface latent heat flux and Q_{SFC} represents the residual flux, including subtracted conduction terms. The influence of the conduction terms on the overall energy budget can be significant, but because we are evaluating the atmosphere these terms are not discussed here. The signs of radiative terms are in line with their impact on the surface, with positive values acting to heat the surface and negative values acting to cool the surface, while the turbulent latent and sensible heat flux terms follow their traditionally applied sign convention, with negative values warming the surface and positive values cooling the surface. In all panels of Figures 4 and 6 the ASCOS observations are presented in black, with various model results indicated in colored lines. For the surface energy budget terms, ERA-I, R-1 and R-2 all provide 6-hourly averaged values, while CAM5 and the observations provide 1-hourly averages and GISS ModelE2 provides 3-hourly instantaneous values. To best compare these different values, the 1-hourly averages from CAM5 and the observations, as well as the 3-hourly instantaneous GISS ModelE2 values are averaged over the 6-hour period represented by the reanalyses.

Looking first at surface shortwave radiation (Figure 4[a-c]), there are substantial differences between the different models. ERA-I appears to be the only model that comes close to resembling observed values, with CAM5, R-1 and R-2 featuring excessive downwelling shortwave radiation and ModelE2 featuring too little downwelling shortwave radiation, particularly in the first of the observational period. While some of this may be the result of differences in atmospheric chemical composition and the radiative transfer codes applied, this result should generally not be considered surprising if we look back at the cloud properties contained in the different models. R-1 and R-2, which generally featured the lowest CFs, also demonstrate the largest positive biases in downwelling shortwave radiation. CAM5 CFs, which, while closer to the observations than the NCEP products, were still low and also allow for excessive solar radiation to reach the surface. CAM5-PF had improved cloud properties, and generally features better agreement with the observed downwelling shortwave radiation, though it does have larger variability than CAM5. ERA-I CF was generally comparable to that observed, and correspondingly, the downwelling shortwave is also comparable

to that observed. While ERA-I LWP appears to be lower at times than observed, these differences occur during periods where the LWP is high enough for most sunlight reaching the surface to be diffuse anyway, resulting in reduced differences in the downwelling shortwave radiation. The GISS ModelE2, which generally slightly underestimates CF and LWP features downwelling shortwave radiative flux densities that are generally biased slightly high.

Upwelling shortwave radiation demonstrates a similar pattern, with ERA-I most closely resembling the observations, and most other models featuring excessive outgoing radiation. ModelE2 has too little upwelling shortwave at the surface which should not be surprising considering its underestimate of incoming shortwave radiation. The upwelling shortwave radiative flux density is governed in part by the surface albedo produced in each of the models (Figure 5[a]), a quantity that varies substantially from one model to the next. These differences can be explained by differences between how the model products handle snow and sea ice. For example, the large difference between R-1 and R-2 is attributed to differences in specified sea-ice cover between the two products, which results in a 10-year zonally-averaged difference at northern high latitudes that is roughly comparable to that detected for the ASCOS period (Kanamitsu et al., 2002). ERA-I does not allow for snow to collect on sea ice surfaces, and therefore precipitation does not directly impact surface albedo as it does in nature. Instead, surface albedo of sea ice is prescribed to vary seasonally, with monthly values based on estimates taken from Ebert and Curry (1993), with a bare ice value used for summer and dry snow values used for winter months. This climatological estimate is responsible for the gradual increase in albedo demonstrated to be present in ERA-I is the result of changes in local sea ice concentration as well as a seasonal shift to higher albedo values with time. The smaller variability is a result of the impact of clouds and precipitation on the shortwave spectrum, which results in small variations in the broadband shortwave radiation at the surface. Because the albedos is sensitive to these changes in the spectrum of shortwave radiation, there appears to be a correlation between precipitation and the shortwave albedo (Figure 5[b]). CAM5 albedos are in general the closest to observed values, with a transition from lower values at the beginning of the observational period to a value close to that of snow for the latter portion of the campaign. Albedo values over ice in CAM5 are temperature dependent, and it becomes obvious when comparing CAM5 and CAM5-PF that the influence of cloud cover on near surface temperatures impacts albedo dramatically, with near surface warming between 26-30 August resulting in lower CAM5-PF albedos during those dates. Interestingly, the CAM5 albedo appears to be negatively correlated to precipitation events which implies that the model is producing rain rather than snow. Figure 5[b] illustrates this behavior for the CAM5, R-2 and GISS ModelE simulations, with the 6 hour period directly after the precipitation in these models negatively correlated with precipitation. The negative correlation is not directly related to the precipitation, but both the GISS ModelE2 and CAM5 simulations produce rain for some of the precipitation events. GISS ModelE2 features significantly lower albedos than observed or produced by other models and also appears to have a weak negative correlation to precipitation events at the time

485 of the precipitation event and the time periods directly thereafter. The surface albedo values used in the GISS ModelE2 simulation are based on observations for this time of year from SHEBA and appear to include a high melt pond fraction which acts to reduce the surface albedo. This underestimation of surface albedo exacerbates the high bias in downwelling shortwave radiation in the GISS ModelE2, resulting in too little solar radiation leaving the earth's surface.

490 Interestingly, the net impact of the issues discussed above is relatively small due to the presence of compensating errors and the resulting general agreement in net solar radiation at the surface. F_{SW} biases for most of the models are smaller than may be expected given the relatively large biases in the upwelling and downwelling terms (see box plots on right hand side of Figure 4), with median values of 5.33 (R-2), 7.11 (ERA-I), 9.21 (CAM5), and 9.47 (CAM5-PF) W m^{-2}. The only major exceptions
495 to this are R-1 and GISS ModelE2, which are shown to be biased high by 48.08 and 30.26 W m^{-2}, respectively. This comparison demonstrates the extent to which surface albedo values can result in compensating errors that nearly cancel out in the net shortwave radiative flux density. Both R-1 and GISS ModelE have relatively low albedos resulting in large positive imbalances in shortwave radiation, while R-2 and the two CAM5 versions have higher albedos which help to cancel out their
500 excessive surface downwelling radiation. ERA-I's albedo, which is also generally too low, helps to make up for the low bias in downwelling shortwave radiative flux density.

In the Arctic environment, longwave radiation is a crucial contributor to the surface energy budget due to the reduced influence of solar radiation during all but summer months. Figure 4[d-f] illustrates the models' performance in simulating surface longwave radiation. Upwelling longwave radiation
505 is governed primarily by surface temperature, and errors in this quantity are generally the result of problems with the lower boundary condition. Conversely, surface downwelling longwave radiation is governed by atmospheric temperature structure, and atmospheric optical depth. The CAM5 and CAM5-PF simulations are demonstrated to have outgoing surface longwave radiation values that are slightly higher (less negative) than observed. This would imply a colder surface, possibly a re-
510 sult of having elevated sea ice concentrations. Improvement in the representation of clouds results in a warming of the near-surface environment in CAM5-PF for the second half of the observation period, which in turn results in more upwelling long wave radiation (as well as more downwelling long wave radiation) at the surface when compared to the CAM5 simulation. The insufficient surface downwelling longwave radiation in CAM5 is in part responsible for temperature biases for the
515 same time period demonstrated in Figure 2. The other models generally produce outgoing surface longwave values that are closer to observations. ModelE2, while performing respectably with regard to outgoing long wave radiation, performs rather poorly in it's prediction of downwelling radiation, implying that while surface temperatures are close to what they should be, atmospheric temperature structure, clouds and/or aerosols are poorly represented. Both R-1 and R-2 trend towards under pre-
520 diction (not negative enough) of incoming and outgoing long wave radiation during the second half of the observation period. This, along with under estimation of downwelling long wave radiation in

CAM5 and ModelE2 is at least partially a result of errors in the cloud fields, as illustrated in Figure 8. ERA-I, with the best estimates of LWP behind GISS ModelE2, performs best in simulating net surface longwave radiation, but does not capture the reductions in downwelling long wave on 21-24 August and after 30 August. Both versions of CAM5 feature lower LWP, with more values falling below the 30 g m^{-2} boundary between black- and greybody clouds (Shupe and Intrieri, 2004). The reduced emissivity associated with these thinner clouds results in a significant reduction in surface LW_{NET}, as shown in Figure 8[a].

Looking at the net surface longwave radiation, ERA-I easily outperforms other models with a median bias of less than -1.43 W m^{-2}. All other models are shown to radiate excessively to the atmosphere, with median biases of -54.58 (R-1), -31.07 (R-2), -34.12 (CAM5), -15.50 (CAM5-PF) and -19.30 (ModelE2) W m^{-2}. As discussed above, these biases are in large part due to problems with simulated cloud cover. CAM5-PF shows significant improvement in the simulation of long wave radiation when compared with CAM5, mainly as a result of improved simulation of liquid-containing stratiform clouds during the second half of the observation period. One of the more confusing results comes from GISS ModelE2, which appears to closely match the observations in both LWP and CF (Figure 8[b]), under represents the surface net long wave radiation substantially. This is likely a result of clouds occurring at the wrong altitude, and therefore emitting at the wrong (colder) temperature.

Sensible heat fluxes (Figure 6[a]) are generally small compared to the radiative terms. These fluxes represent the turbulent transfer of heat across temperature gradients near the surface, and are directed upward when temperature decreases with height at a rate exceeding the adiabatic lapse rate. ERA-I outperforms the other models with a small median bias of -0.85 W m^{-2}. GISS ModelE2 is the next best performer, with a median bias of 2.60 W m^{-2}. R-1, CAM5 and CAM5-PF all have median biases between 5.8-6.4 W m^{-2}, while R-2 has the largest median error (17.84 W m^{-2}). Given the limited amount of information available, it is challenging to say why exactly R-2 features large biases compared to other models. R-2 near-surface air temperature biases (Figure 2[a]) are often smaller than those of other models, and based on upwelling long wave radiation biases, R-2 surface temperatures are not necessarily worse than those for CAM5. R-2 surface winds are perhaps a little bit too high, but the biases here are again not out of line with those detected for CAM5 and GISS ModelE2, for example. This may leave the parameterization of bulk transfer coefficient as a potential culprit, and unfortunately the models do not provide the detailed output required to correctly diagnose where the issues lie beyond what is discussed above.

Latent heat fluxes, representing heat transfer through turbulent transfer of moisture across specific humidity gradients, are illustrated in Figure 6[d]. Again, as with sensible heat, values are generally much smaller than with the radiative terms. The CAM5 simulation generally agrees most closely with observed values, with a median bias of -0.37 W m^{-2}. R-2 also performs well, with a median bias of -0.47 W m^{-2}. These are followed by CAM5-PF (median bias -1.61 W m^{-2}), ERA-I (2.21 W

m^{-2}), ModelE2 (6.28 W m^{-2}) and R-1 (-7.47 W m^{-2}). ERA-I and ModelE2 are the only models that demonstrate a positive median biases, likely indicating a smaller decrease of specific humidity with height in the model's near surface atmosphere than was observed. The relative change in performance between the sensible and latent heat components appears to indicate that the biases result from factors beyond those in common between the two terms, such as near surface winds.

Somewhat incredibly, R-1, which was illustrated to be among the worst performing models for many of the individual terms discussed above, seems to most closely portray the net Q_{SFC} term with a bias of only 2.97 W m^{-2}. ERA-I and GISS ModelE2 are the only other products that demonstrate a positive bias in Q_{SFC} (7.24 and 22.38 W m^{-2}, respectively). The other models all excessively lose heat at the surface, with R-2 featuring a median bias of -11.68 W m^{-2}, CAM5 a bias of -18.62 W m^{-2} and CAM5-PF a bias of -7.57 W m^{-2}. A large portion of these biases is demonstrated to come from long wave radiative biases, with most of those being the result of the downwelling long wave. This illustrates the important role that cloud physics play in regulating this vital component of the climate system. Additionally, this result demonstrates clearly the influence of compensating errors on overall evaluations, and that in general, care must be taken to incorporate all of the budget terms in model evaluations, rather than simply looking at the net product term. If we instead look at the distribution of the absolute values of biases of the budget terms, only including the individual components (SW_{up}, SW_{down}, LW_{up}, LW_{down}, F_{SH}, F_{LH}), (Figure 7) we see a very different representation of model performance. Here, ERA-I outperforms the other models with significantly smaller biases in the budget terms, with GISS ModelE2 a close second. CAM5PF demonstrates significant improvement over CAM5, with R-1 and R-2 falling in between these two. This figure can be thought of as a step towards diagnosing the models' abilities to get the right answer for the right reasons, rather than simply getting the right answer.

5.4 Lower Tropospheric Temperature Structure

In order to evaluate the impact of the surface radiative balance terms on atmospheric state, here we assess the models' ability to simulate lower atmospheric temperature structure. Figure 9 illustrates temperature biases relative to radiosonde observations for the lowest 3000 m of the atmosphere for all five model time series. On the right of each figure is a profile demonstrating the mean (bold line) temperature bias profile, along with the interquartile range of these biases for each model grid box level. Biases detected for the lowest model levels follow biases in the surface energy budget, with CAM5, CAM5-PF and R-2 all featuring cold biases close to the Earth's surface. CAM5-PF has a reduced cold bias when compared to CAM5, but also features a 1-2 degree warm bias higher in the atmosphere. R-1 and ERA-I have slight warm biases low in the atmosphere, which is not surprising given the neutral or positive biases they demonstrate in Q_{SFC}. Both R-1 and R-2 feature cold biases in the upper portion of the evaluated domain, while ERA-I and CAM5 have very small net biases at higher altitudes and CAM5-PF has a warm bias. All models except for ERA-I and GISS ModelE2

595 have warm biases between 500-1000 m, resulting from insufficient low-level cloud cover in those simulations.

Looking at this in a different way, we can evaluate the models' ability to simulate the potential temperature difference between the surface and 850 mb (Hereafter LTS_{850}, Figure 10). For the ASCOS period, 850 mb fell between 1200-1600 m in altitude. LTS_{850} can be thought of as repre-
600 sentative of the mean stability of the lower atmosphere, with stability increasing with the difference between these temperatures. All of the models accurately represent LTS_{850} during the first days of the observational period (12 August - 15 August). After this, the models begin to diverge somewhat, with R-1 and R-2 generally being less stable than observations until 24 August, when R-2 suddenly shifts to generally being more stable than the observations. Looking back at Figure 9, we can see
605 that the instability (relative to observations) between 15-24 August in R-1 and R-2 is the result of a combination of near-surface warm bias and slight 850 mb cold bias. After 24 August, R-2 features a cold near-surface temperature, which results in its shift to a more stable lower atmosphere. When looking at the distributions of LTS_{850} (Figure 10[b]), the general picture is confirmed, with R-1 having an atmosphere that is generally less stable than observations, and R-2 more closely matching
610 observations. ERA-I generally has lower-atmospheric stabilities that are similar to those observed, with two exceptions. The largest differences between ERA-I and observations occur from 21-26 August and 1-2 September, where ERA-I is less stable than observations. During the 21-26 August time period, ERA-I exhibits a near-surface warm bias, resulting mainly from a high bias in the net surface long wave radiation (Figure 4[f]). It appears as though a similar bias in surface radiation
615 may have occurred between 1-2 September, though we do not have surface radiation measurements for much of that time period. In addition to the surface radiation-induced near-surface temperature bias, ERA-I also has a cold bias around 850 mb for the 1-2 September time period, adding to the reduced stability. Overall, ERA-I is shown to be less stable than observations, and in particular does not capture more stable environments well (right hand tail of distributions in Figure 10[b]). CAM5
620 and CAM5-PF simulations are generally too stable after 18 August. The exceptions to this are the periods from 23-24 August where both agree more closely with observed stability, and then the 24-27 August period where CAM5-PF agrees closely with observed stability. The excessive stability is largely the result of near-surface temperatures that are too cold in both versions of this model (Figure 9). The biases in lower atmospheric stability again closely track the net long wave radiation at the
625 surface (Figure 4[f]), which for both versions of CAM5 is generally under predicted. Time periods mentioned above with improved performance match the time periods in which net long wave radiation is in better agreement with observations. Distributions of lower atmospheric stability for the entire ASCOS period demonstrate the slightly improved simulation of this quantity with CAM5-PF, with more of the values moving towards the less stable (left) side of Figure 10[b]. Finally, evalua-
630 tion of ModelE2 provides us with a somewhat puzzling result. Despite large biases in both the net surface energy flux, an evaluation of ModelE2 lower atmospheric stability results in a very favorable

comparison with observations. The main exceptions to this occur between 27-28 August (too unstable) and 28-31 August (too stable). The GISS ModelE2 distribution of LTS_{850} matches observations more closely than any of the other models. Figure 10[c,d] demonstrate the relative contributions of near surface air temperature and 850 mb temperature on LTS_{850}. These values were obtained by using one of the observed values (either θ_{SFC} or θ_{850}) together with the model value for the other in order to calculate LTS_{850}, and then calculating the bias in LTS_{850} relative to values calculated directly from the observations. As discussed above, most of the discrepancies result from differences in the near surface potential temperature, where variability is somewhat larger. The two main exceptions to this are the event on 27 August that all of the models appeared to misrepresent at 850 mb, and the general contribution of CAM5-PF's warm bias at 850 mb. CAM5's excessive stability is clearly illustrated to result mainly from the near-surface potential temperature, with LTS_{850} biases much larger when using the model θ_{SFC} than when using the model θ_{850}.

6 Summary and Discussion

The different models had variable success in simulating the evaluated properties. A brief summary of the results is included here:

- *Surface Meteorology*: With some minor exceptions, all of the models had small biases in wind speed and direction. Surface pressure was generally well simulated, with the exception of R-1 (-3.01 mb bias) and R-2 (4.93 mb bias). Variables more closely tied to clouds and radiation such as near surface temperature and humidity were shown to be more poorly simulated. Median near-surface temperature biases ranged between -3.39 K (CAM5) to 1.15 K (R-1), while median relative humidity biases ranged between -7.80% (GISS ModelE2) to 2.39% (ERA-I) with all models demonstrating substantial variability. ERA-I generally featured superior correlation with observations over the other models for these variables.

- *Clouds and Precipitation*: Cloud-related processes continue to represent a major stumbling block for accurate simulation of the Arctic environment. With the exception of ERA-I and GISS ModelE2, models struggled to produce the amount of cloud cover observed. Cloud liquid water path was underestimated by three of the four models that provided this quantity, with GISS ModelE2 producing the closest amount of liquid water (median bias of 2.87 g m^{-2}), though care needs to be taken since the observations represent a lower limit, and likely an underestimate of the true liquid water present. CAM5-PF did demonstrate improvement over the standard version in simulating liquid water amount. While overall precipitation amounts were low during ASCOS, for what was recorded, median model errors were generally between 1-5%, which is likely well within the errors associated with measuring precipitation. However, variability in those numbers was large, with models not reproducing three of the four more substantial precipitation events observed during ASCOS.

– *Surface Energy Budget*: Model biases in cloud-related processes discussed above result in large biases in the radiative components of the surface energy budget. Shortwave radiation biases are generally as may be expected, with models that have too few clouds also featuring too much downwelling shortwave radiation at the surface, and GISS ModelE2 agreeing most directly with measurements (median bias of 3.97 W m^{-2}). Biases in upwelling shortwave radiation generally mirror those in downwelling shortwave, although the relative magnitude is modulated by the model surface albedos, which were demonstrated to vary widely with ModelE2 having the lowest (0.29-0.55) and R-2 having the highest (0.85-0.88). Net shortwave median biases were generally less than 10 W m^{-2} with R-1 and GISS ModelE2 being exceptions (median biases of 48.08 and 19.57 W m^{-2}, respectively). Longwave radiation biases are more variable. Except for ERA-I all of the models have negative median biases in downwelling long wave radiation resulting from a lack of optically thick clouds in the lower atmosphere. All of the models demonstrate a reduction in upwelling long wave radiation with a reduction in surface temperature, as is present in the observations, but the magnitude of upwelling long wave values varies from model to model and includes a feedback from the downwelling long wave caused by clouds (or a lack thereof). Net long wave radiation biases generally follow those in the downwelling long wave, with ERA-I comparing most favorably (median bias of -1.43 W m^{-2}) and other models featuring excessive radiation to the atmosphere. Sensible and latent turbulent heat fluxes are observed and modeled to be small compared to radiative fluxes, with R-2 featuring the largest biases in sensible heat fluxes and R-1 featuring the largest latent heat flux biases. Overall, ERA-I outperforms other models in representing surface energy budget terms, as indicated by the absolute bias illustrated in Figure 7. Somewhat incredibly, despite relatively large absolute biases, compensating errors result in R-1 having the smallest net surface flux bias and best correlation. The ModelE2 performs the worst, with a net surface energy median bias of 22.38 W m^{-2}.

– *Lower Atmospheric Temperature Structure*: The models vary widely on their representation of lower atmospheric temperature. R-2 and both versions of CAM5 both feature near-surface cold biases for much of the ASCOS period, and all models except ERA-I feature a slight warm bias between 500-1200 m. R-1 and R-2 both have significant (around 2 K) cold biases near 3000 m. From a bulk stability perspective, R-1 and ERA-I tend to have lower atmospheres that is slightly less stable than observations, while CAM5 (both versions) tend to be too stable. ModelE2 and R-1 most closely match the observed lower tropospheric stability, as defined by the temperature difference between the surface and 850 mb.

The demonstrated model biases can have wide ranging impacts. For one, reanalysis output is sometimes used to force large-scale or local sea ice and ocean models to evaluate their performance against available observations (e.g., Brodeau et al., 2010; Miller et al., 2007). Curry et al. (2002) evaluated several datasets used to force sea ice models during the SHEBA time period and noted that

substantial differences can be found between the different datasets. Our evaluation results in a similar conclusion, with notable differences between products in quantities relevant to ocean processes and sea ice growth and decay, such as winds, precipitation and radiation. Additionally, reanalyses are often used to evaluate the performance of climate models in simulation of the present-day atmosphere (e.g., de Boer et al., 2012; Schmidt et al., 2006). For variables with well-characterized, small biases, this may be acceptable, but for other variables (generally those related to clouds, precipitation and radiation), this sort of evaluation is generally inappropriate given the relatively comparable magnitudes of errors in the reanalyses and differences between reanalyses and models.

Biases in GCMs can cause serious issues with determining equilibrated climatic states. Since small biases in a fully-coupled climate model can run away into new climatic regimes, biases in the model themselves can, over several year periods, result in large errors in predicted climate. This is particularly true in a sensitive area such as the Arctic where small shifts in climatic equilibrium can result in stark transitions, such as those from ice-covered to melted surfaces. One challenging aspect of a model evaluation such as this one is distinguishing between the errors resulting from the model itself and those resulting from the dataset from which forecasts were initialized. In the present study, there are some GCM biases that appear to result from the model itself, such as the lack of low clouds in CAM5 and the associated surface energy balance errors and biases in lower atmospheric stability.

The ASCOS campaign helps to illustrate some of the issues faced by models in the Arctic environment. For the reanalyses, there may be significant advances in product accuracy with the integration of additional observations into the GTS. This would help to improve the analyses used for forecast initialization, giving the model less room to stray from the measured atmospheric state. In terms of model performance, current parameterizations continue to be challenged in correctly simulating Arctic processes. While previous campaigns (e.g., SHEBA, ASCOS) have greatly improved our understanding of some of these processes, others could use additional improved observations to enhance our understanding. In particular, cloud processes are demonstrated here to cause problems for all of the models involved, with errors in the representation of clouds translating to surface energy and near-surface temperature errors. This can have significant impacts on climate simulation as these imbalances will likely drive the climate into an altered state through various feedback processes. The short duration of ASCOS ultimately results in an incomplete evaluation of these models and similar datasets need to be obtained for other times of year and other parts of the Arctic to complete such evaluations. Ultimately, increased observations are necessary to better understand and consequently simulate this environment. This is particularly true over minimally-studied areas such as the marginal ice zone, which continues to expand with the shrinking sea ice pack.

Acknowledgements. The authors wish to thank ECMWF for making YOTC analysis data available for research purposes. ASCOS was made possible by funding from the Knut and Alice Wallenberg Foundation, the DAMOCLES European Union 6th Framework Program, the Swedish National Research Council (VR), the US National Science Foundation (NSF), the National Atmospheric and Oceanic Administration (NOAA) and the UK Natural

Environment Research Council (NERC). CAM5 and the CESM project are supported by the National Science Foundation and the Office of Science (BER) of the U.S. Department of Energy. NCEP Reanalysis data provided by the NOAA/OAR/ESRL PSD, Boulder, Colorado, USA, from their Web site at http://www.esrl.noaa.gov/psd/. This work was prepared in part at the Cooperative Institute for Research in Environmental Sciences (CIRES) with support in part from the National Oceanic and Atmospheric Administration, U.S. Department of Commerce, under cooperative agreement NA17RJ1229 and other grants. The statements, findings, conclusions, and recommendations are those of the author and do not necessarily reflect the views of the National Oceanic and Atmospheric Administration or the Department of Commerce. This research was supported in part by the Director, Office of Science, Office of Biological and Environmental Research of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231. Additionally, this work was supported by the National Science Foundation under grant numbers ARC-1023366 and ARC1203902 as well as the US Department of Energy under grant DE-SC0008794. Resources supporting this work were additionally provided by the NASA High-End Computing (HEC) Program through the NASA Center for Climate Simulation (NCCS) at Goddard Space Flight Center. The efforts of PMC, JSB, and SAK were supported by the Earth System Modeling program of the United States Department of Energy's Office of Science and were performed under the auspices of the United States Department of Energy by Lawrence Livermore National Laboratory under contract DE-AC52-07NA27344.

References

Bauer, S. and Menon, S.: Aerosol direct, indirect, semi-direct and surface albedo effects from sector contributions based on the IPCC AR5 emissions for pre-industrial and present day conditions, J. Geophys. Res., 117, D01 206, doi:10.1029/2011JD016 816, 2012.

Bauer, S., Wright, D., Koch, D., Lewis, E., McGraw, R., Chang, L.-S., Schwartz, S., and Ruedy, R.: MATRIX (Multiconfiguration Aerosol TRacker of mIXing state): an aerosol microphysical module for global atmospheric models, Atmos. Chem. Phys., 8, 6003–6035, 2008.

Birch, C., Brook, I., Tjernström, M., Shupe, M., Mauritsen, T., Sedlar, J., Lock, A., Earnshaw, P., Persson, P., Milton, S., and Leck, C.: Modelling atmospheric structure, cloud and their response to CCN in the central Arctic: ASCOS case studies, Atmos. Chem. Phys., 12, 3419–3435, 2012.

Boé, J., Hall, A., and Qu, X.: Current GCMs' unrealistic negative feedback in the Arctic, J. Clim., 22, 4682–4695, 2009.

Boyle, J., Williamson, D., Cederwall, R., Fiorino, M., Hnilo, J., Olson, J., Phillips, T., Potter, G., and Xie, S.: Diagnosis of Community Atmosphere Model 2 (CAM2) in numerical weather forecast configuration at atmospheric radiation measurement sites, J. Geophys. Res., 110, D15S15, 2005.

Brodeau, L., Barnier, B., Treguier, A.-M., Penduff, T., and Gulev, S.: An ERA40-based atmospheric forcing for global ocean circulation models, Ocean Model., 31, 88–104, doi: 10.1016/j.ocemod.2009.10.005, 2010.

Chylek, P., Folland, C., Lesins, G., Dubey, M., and Wang, M.: Arctic air temperature change amplification and the Atlantic Multidecadal Oscillation, Geophys. Res. Lett., 36, L14 801, 2009.

Compo, G., Whitaker, J., Sardeshmukh, P., Matsui, N., Allan, R., Yin, X., Gleason, B., Vose, R., Rutledge, G., Bessemoulin, P., Brönnimann, S., Brunet, M., Crouthamel, R., Grant, A., Groisman, P., Jones, P., Kruk, M., Kruger, A., Marshall, G., Maugeri, M., Mok, H., Nordli, O., Ross, T., Trigo, R., Wang, X., Woodruff, S., and Worley, S.: The twentieth century reanalysis project, Quart. J. Roy. Meteor. Soc., 137, 1–28, 2011.

Curry, J., Schramm, J., and Ebert, E.: Sea ice-albedo climate feedback mechanism, J. Clim., 8, 240–247, 1995.

Curry, J., Schramm, J., Alam, A., Reeder, R., Arbetter, T., and Guest, P.: Evaluation of data sets used to force sea ice models in the Arctic Ocean, J. Geophys. Res., 107, 10.1029/2000JC000 466, 2002.

de Boer, G., Chapman, W., Kay, J., Medeiros, B., Shupe, M., Vavrus, S., and Walsh, J.: A characterization of the present-day Arctic atmosphere in CCSM4, J. Clim., 25, 2676–2695, 2012.

de Boer, G., Bauer, S., Toto, T., Menon, S., and Vogelman, A.: Evaluation of aerosol-cloud interaction in the GISS ModelE using ARM observations, J. Geophys. Res., Accepted, 2013.

Dee, D., Uppala, S., Simmons, A., Berrisford, P., Poli, P., Kobayashi, S., Andreae, U., Balmaseda, M., Balsamo, G., Bauer, P., Bechtold, P., Beljaars, A., Van de Berg, L., Bidlot, J., Bormann, N., Delsol, C., Dragani, R., Fuentes, M., Geer, A., Haimberger, L., Healy, S., Hersbach, H., Hólm, E., Isaksen, L., Kållberg, P., Köhler, M., Matricardi, M., McNally, A., Monge-Sanz, B., Morcrette, J.-J., Park, B.-K., Peubey, C., de Rosnay, P., Tavolato, C., Thépaut, J.-N., and Vitart, F.: The ERA-Interim reanalysis: configurationand performance of the data assimilation system, Quart. J. Roy. Meteor. Soc., 137, 553–597, 2011.

Ebert, E. and Curry, J.: An intermediate one-dimensional thermodynamic sea ice model for investigating ice-atmosphere interactions, J. Geophys. Res., 98C, 10 085–10 109, 1993.

Gascard, J.-C., Festy, J., LeGoff, H., Weber, M., Bruemmer, B., Offermann, M., Doble, M., Wadhams, P., Forsberg, R., Hanson, S., Skourup, H., Gerland, S., Nicolaus, M., Metaxian, J.-P., Grangeon, J., Haapala, J.,

Rinne, E., Haas, C., Heygster, G., Jakobson, E., Palo, T., Wilkinson, J., Kaleschke, L., Claffey, K., Elder, B., and Bottenheim, J.: Exploring Arctic transpolar drift during dramatic sea ice retreat, EOS, 89, 21–28, 2008.

Graversen, R., Mauritsen, T., Tjernström, M., Källén, E., and Svensson, G.: Vertical structure of recent Arctic warming, Nature, 541, 53–56, 2008.

Hansen, J., Sato, M., Ruedy, R., Kharecha, P., Lacis, A., Miller, R., Nazarenko, L., Lo, K., Schmidt, G., Russel, G., Aleinov, I., Bauer, S., Baum, E., Cairns, B., Canuto, V., Chandler, M., Cheng, Y., Cohen, A., Del Genio, A., Faluvegi, G., Fleming, E., Friend, A., Hall, T., Jackman, C., Jonas, J., Kelley, M., Kiang, N., Koch, D., Labow, G., Lerner, J., Menon, S., Novakov, T., Oinas, V., Perlwitz, J., Perlwitz, J., Rind, D., Romanou, A., Schmunk, R., Shindell, D., Stone, P., Sun, S., Streets, D., Tausnev, N., Thresher, D., Unger, N., Yao, M., and Zhang, S.: Climate simulations for 1880-2003 with GISS modelE, Clim. Dyn., 29, 661–696, 2007.

Inoue, J., Liu, J., Pinto, J., and Curry, J.: Intercomparison of Arctic regional climate models: modeling clouds and radiation for SHEBA in May 1998, J. Climate, 19, 4167–4178, 2006.

IPCC: Climate Change 2007: The Physical Science Basis, Cambridge University Press, 2007.

Jakobson, E., Vihma, T., Palo, T., Jakobson, L., Keernik, H., and Jaagus, J.: Validation of atmospheric reanalyses over the central Arctic Ocean, Geophys. Res. Lett., 39, L10 802, 2012.

Kalnay, E., Kanamitsu, M., Kistler, R., Collins, W., Deaven, D., Gandin, L., Iredell, M., Saha, S., White, G., Woollen, J., Zhu, Y., Chelliah, M., Ebisuzaki, W., Higgins, W., Janowiak, J., Mo, K., Ropelewski, C., Wang, J., Leetmaa, A., Reynolds, R., Jenne, R., and Joseph, D.: The NCEP/NCAR 40-year reanalysis project, Bull. Am. Meteorol. Soc., 77, 437–471, 1996.

Kanamitsu, M., Ebisuzaki, W., Woollen, J., S.-K., Y., Hnilo, J., Fiorino, M., and Potter, G.: NCEP-DOE AMIP-II Reanalysis (R-2), Bull. Am. Meteorol. Soc., 83, 1631–1643, 2002.

Kay, J. and Gettelman, A.: Cloud influence on and response to seasonal Arctic sea ice loss, J. Geophys. Res., 114, D18 204, doi:10.1029/2009JD011 773, 2009.

Kazutoshi, K., Tsutsui, J., Koide, H., Sakamoto, M., Kobayashi, S., Hatsushika, H., Matsumoto, T., Yamazaki, N., Kamahori, H., Takahashi, K., Kadokura, S., Wada, K., Kato, K., Oyama, R., Ose, T., Mannoji, N., and Ryusuke, T.: The JRA-25 reanalysis, J. Meteor. Soc. Japan, 85, 369–432, 2007.

Lamarque, J.-F., Bond, T., Eyring, V., Granier, C., Heil, A., Klimont, Z., Lee, D., Liousse, C., Mieville, A., Owen, B., Schultz, M., Shindell, D., Smith, S., Stehfest, E., Van Aardenne, J., Cooper, O., Kainuma, M., Mahowald, N., McConnel, J., Naik, V., Riahi, K., and van Vuuren, D.: Historical (1850-2000) gridded anthropognic and biomass burning emissions of reactive gases and aerosols: Methodology and application, Atmos. Chem. Phys., 10, 7017–7039, 2010.

Liu, J., Zhang, Z., Hu, Y., Chen, L., Dai, Y., and Ren, X.: Assessment of surface air temperature over the Arctic Ocean in reanalysis and IPCC AR4 model simulations with IABP/POLES observations, J. Geophys. Res., 113, D10 105, 2007.

Liu, X., Easter, R., Ghan, S., Zaveri, R., Rasch, P., Shi, X., Lamarque, J.-F., Gettelman, A., Morrison, H., Vitt, F., Conley, A., Park, S., Neale, R., Hannay, C., Ekman, A., Hess, P., Mahowald, N., Collins, W., Iacono, M., Bretherton, C., Flanner, M., and Mitchell, D.: Toward a minimal representation of aerosoldirect and indirect effects: model description and evaluation, Geosci. Model Dev. Discuss., 4, 3485–3598, 2011.

McComiskey, A. and Feingold, G.: The scale problem in quantifying aerosol indirect effects, Atmos. Chem. Phys., 12, 1031–1049, 2012.

Mesinger, F., DiMego, F., Kalnay, E., Mitchell, K., Shafran, P., Ebisuzaki, W., Jović, D., Woollen, J., Rogers, E., Berbery, E., Ek, M., Fan, Y., Grumbine, R., Higgins, W., Li, H., Lin, Y., Manikin, G., Parrish, D., and Shi, W.: North American regional reanalysis, Bull. Am. Meteorol. Soc., 87, 343–360, 2006.

Miller, P., Laxon, S., and Feltham, D.: Consistent and contrasting decadal Arctic sea ice thickness predictions from a highly optimized sea ice model, J. Geophys. Res., 112, C07 020, doi:10.1029/2006JC003 855, 2007.

Moran, K., Martner, B., Post, M., Kropfli, R., Welsch, D., and Widener, K.: An unattended cloud-profiling radar for use in climate research, Bull. Am. Meteorol. Soc., 79, 443–455, 1998.

Neale, R., Gettelman, A., Park, S., Chen, C.-C., Lauritzen, P., Williamson, D., Conley, A., Garcia, R., Kinnison, D., Lamarque, J.-F., Marsh, D., Mills, M., Smith, A., Tilmes, S., Vitt, F., Morrison, H., Cameron-Smith, P., Collins, W., Iacono, M., Easter, R., Liu, X., Ghan, S., Rasch, P., and Taylor, M.: Description of the NCAR Community Atmosphere Model (CAM 5.0), Tech. rep., NCAR, 2010.

Phillips, T., Potter, G., Williamson, D., Cederwall, R., Boyle, J., Fiorino, M., Hnilo, J., Olson, J., Xie, S., and Yio, J.: Evaluating parameterizations in general circulation models: Climate simulation meets weather prediction, Bull. Am. Meteorol. Soc., 85, 1903–1915, 2004.

Polyakov, I., Alekseev, G., Bekryaev, R., Bhatt, U., Colony, R., Johnson, M., Karklin, V., Makshtas, A., Walsh, D., and Yulin, A.: Observationally based assessment of polar amplification of global warming, Geophys. Res. Lett., 29, 1878, 2002.

Rienecker, M., Suarez, M., Gelaro, R., Todling, R., Bacmeister, J., Liu, E., Bosilovich, M., Schubert, S., Takacs, L., Kim, G.-K., Bloom, S., Chen, J., Collins, D., Conaty, A., da Silva, A., Gu, W., Joiner, J., Koster, R., Lucchesi, R., Molod, A., Owens, T., Pawson, S., Pegion, P., Redder, C., Reichle, R., Robertson, F., Ruddick, A., Sienkiewicz, M., and Woollen, J.: MERRA: NASA's Modern-Era Retrospective Analysis for Research and Applications, J. Clim., 24, 3624–3648, 2011.

Rigor, I., Colony, R., and Martin, S.: Variations in surface air temperature observations in the Arctic, 1979-97, J. Climate, 13, 896–914, 2000.

Saha, S., Moorthi, S., Pan, H.-L., Wu, X., Wang, J., Nadiga, S., Tripp, P., Kistler, R., Woollen, J., Behringer, D., Liu, H., Stokes, D., Grumbine, R., Gayno, G., Wang, J., Hou, Y.-T., Chuang, H.-Y., Juang, H.-M. H., Sela, J., Iredell, M., Treadon, R., Kleist, D., Van Delst, P., Keyser, D., Derber, J., Ek, M., Meng, J., Wei, H., Yang, R., Lord, S., Van Den Dool, H., Kumar, A., Wang, W., Long, C., Chelliah, M., Xue, Y., Huang, B., Chemm, J.-K., Ebisuzaki, W., Lin, R., Xie, P., Chen, M., Zhou, S., Higgins, W., Zou, C.-Z., Liu, Q., Chen, Y., Han, Y., Cucurull, L., Reynolds, R., Rutledge, G., and Goldberg, M.: The NCEP climate forecast system reanalysis, Bull. Am. Meteorol. Soc., 91, 1015–1057, 2010.

Schmidt, G., Ruedy, R., Hansen, J., Aleinov, I., Bell, N., Bauer, M., Bauer, S., Cairns, B., Canuto, V., Cheng, Y., Del Genio, A., Faluvegi, G., Friend, A., Hall, T., Hu, Y., Kelley, M., Kiang, N., Koch, D., Lacis, A., Lerner, J., Lo, K.-W., Miller, R., Nazarenko, L., Olinas, V., Perlwitz, J., Perlwitz, J., Rind, D., Romanou, A., Russell, G., Sato, M., Shindell, D., Stone, P., Sun, S., Tausnev, N., Thresher, D., and Yao, M.-S.: Present day atmospheric simulations using GISS ModelE: Comparison to in-situ, satellite and reanalysis data, J. Climate, 19, 153–192, 2006.

Schmidt, G. A., Kelley, M., Nazarenko, L., Ruedy, R., Russell, G. L., Aleinov, I., Bauer, M., Bauer, S., Bhat, M. K., Bleck, R., Canuto, V., Chen, Y., Cheng, Y., Clune, T. L., DelGenio, A., de Fainchtein, R., Faluvegi, G., Hansen, J. E., Healy, R. J., Kiang, N. Y., Koch, D., Lacis, A. A., LeGrande, A. N., Lerner, J., Lo,

K. K., Marshall, J., Mathews, E. E., Menon, S., Miller, R. L., Oinas, V., Oloso, A., Perlwitz, J., Puma, M. J., Putman, W. M., Rind, D., Romanou, A., Sato, M., Shindell, D. T., Sun, S., Syed, R., Tausnev, N., Tsigaridis, K., Unger, N., Voulgarakis, A., Yao, M.-S., and Zhang, J.: Configuration and assessment of the GISS ModelE2 contributions to the CMIP5 archive, J. Climate, in preparation.

Screen, J. and Simmonds, I.: The central role of diminishing sea ice in recent Arctic temperature amplification, Nature, 464, 1334–1337, 2010.

Sedlar, J., Tjernström, M., Mauritsen, T., Shupe, M., Brook, I., Persson, P., Birch, C., Leck, C., Sirevaag, A., and Nicolaus, M.: A transitioning Arctic surface energy budget: the impacts of solar zenith angle, surface albedo and cloud radiative forcing, Clim. Dyn., 37, 1643–1660, doi:10.1007/s00 382–010–0937–5, 2011.

Serreze, M. and Francis, J.: The Arctic amplification debate, Climatic Change, 76, 241–264, 2006.

Serreze, M., Barrett, A., Stroeve, J., Kindig, D., and Holland, M.: The emergence of a surface-based Arctic amplification, Cryosphere, 3, 11–19, 2009.

Shupe, M.: A ground-based multisensor cloud phase classifier, Geophys. Res. Lett., 34, L22 809, 2007.

Shupe, M. and Intrieri, J.: Cloud radiative forcing of the Arctic surface: The influence of cloud properties, surface albedo, and solar zenith angle, J. Climate, 17, 616–628, 2004.

Svensson, G. and Karlsson, J.: On the Arctic wintertime climate in global climate models, J. Clim., 24, 5757–5771, 2011.

Tjernström, M., Sedlar, J., and Shupe, M.: How well do regional cliamte models reproduce radiation and clouds in the Arctic?, J. Appl. Meteor. Climatol., 47, 2405–2422, 2008.

Tjernström, M., Leck, C., Birch, C., Bottenheim, J., Brooks, B., Brook, I., Bäcklin, L., Chang, R. Y.-W., Granath, E., Graus, M., Hansel, A., Heintzenberg, J., Held, A., Hind, A., de la Rosa, S., Johnston, P., Knulst, J., de Leeuw, G., Di Liberto, L., Martin, M., Matrai, P., Mauritsen, T., Müller, M., Norris, S., Orellana, M., Orsini, D., Paatero, J., Persson, P., Gao, Q., Rauschenberg, C., Ristovski, Z., Sedlar, J., Shupe, M., Sierau, B., Sirevaag, A., Sjogren, S., Stetzer, O., Swietlicki, E., Szczodrak, M., Vaattovaara, P., Wahlberg, N., Westberg, M., and Wheeler, C.: The Arctic Summer Cloud-Ocean Study (ASCOS): Overview and experimental design, Atmos. Chem. Phys., Submitted, 2013.

Uppala, S., Kållberg, P., Simmons, A., Andreae, U., Da Costa Betchold, V., Fiorino, M., Gibson, J., Haseler, J., Hernandez, A., Kelley, G., Li, X., Onogi, K., Saarinen, S., Sokka, N., Allan, R., Andersson, E., Arpe, K., Balmaseda, M., Beljaars, A., Van de Berg, L., Bidlot, J., Bormann, N., Caires, S., Chevallier, F., Dethof, A., Dragosavac, M., Fisher, M., Fuentes, M., Hagemann, S., Hólm, E., Hoskins, B., Isaksen, L., Janssen, P., Jenne, R., McNally, A., Mahfouf, J.-F., Morcrette, J.-J., Rayner, N., Saunders, R., Simon, P., Sterl, A., Trenberth, K., Untch, A., Vasiljevic, D., Viterbo, P., and Woollen, J.: The ERA-40 re-analysis, Quart. J. Roy. Meteor. Soc., 131, 2961–3012, 2005.

Vihma, T., Jaagus, J., Jakobson, E., and Palo, T.: Meteorological conditions in the Arctic Ocean in spring and summer 2007 as recorded on the drifting ice station Tara, Geophys. Res. Lett., 35, L18 706, 2008.

Walsh, J., Kattsov, V., Chapman, W., Govokova, V., and Pavlova, T.: Comparison of Arctic climate simulations by uncoupled and coupled global models, J. Climate, 15, 1429–1446, 2002.

Walsh, J., Chapman, W., and Portis, D.: Arctic cloud fraction and radiative fluxes in atmospheric reanalyses, J. Clim., 22, 2316–2334, 2008.

Westwater, E., Han, Y., Shupe, M., and Matrosov, S.: Analysis of integrated cloud liquid and precipitable water

vapor retrievals from microwave radiometers during the Surface Heat Budget of the Arctic Ocean project, J. Geophys. Res., 106, 32 019–23 030, 2001.

Zib, B., Dong, X., Xi, B., and Kennedy, A.: Evaluation and intercomparison of cloud fraction and radiative fluxes in recent reanalyses over the Arctic using BSRN surface observations, J. Clim., 25, 2291–2305, 2012.

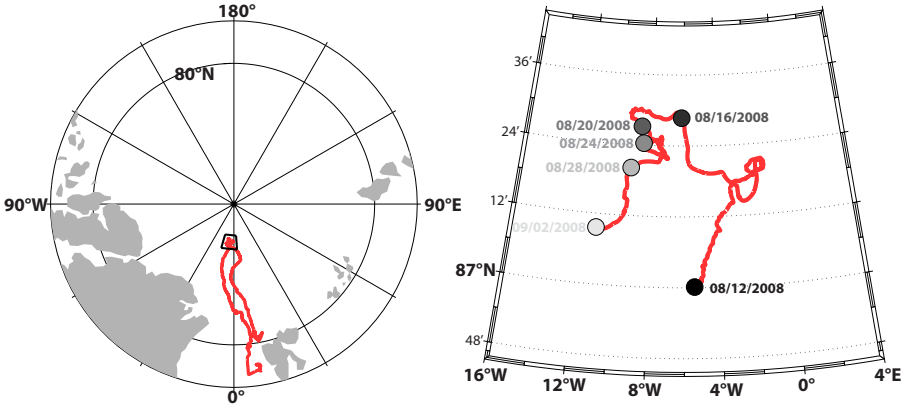

Fig. 1. A map illustrating the path of Oden (red line) during ASCOS. The region of ice drift (black box) is enlarged, and greyscale points provide the location of the ship on given dates.

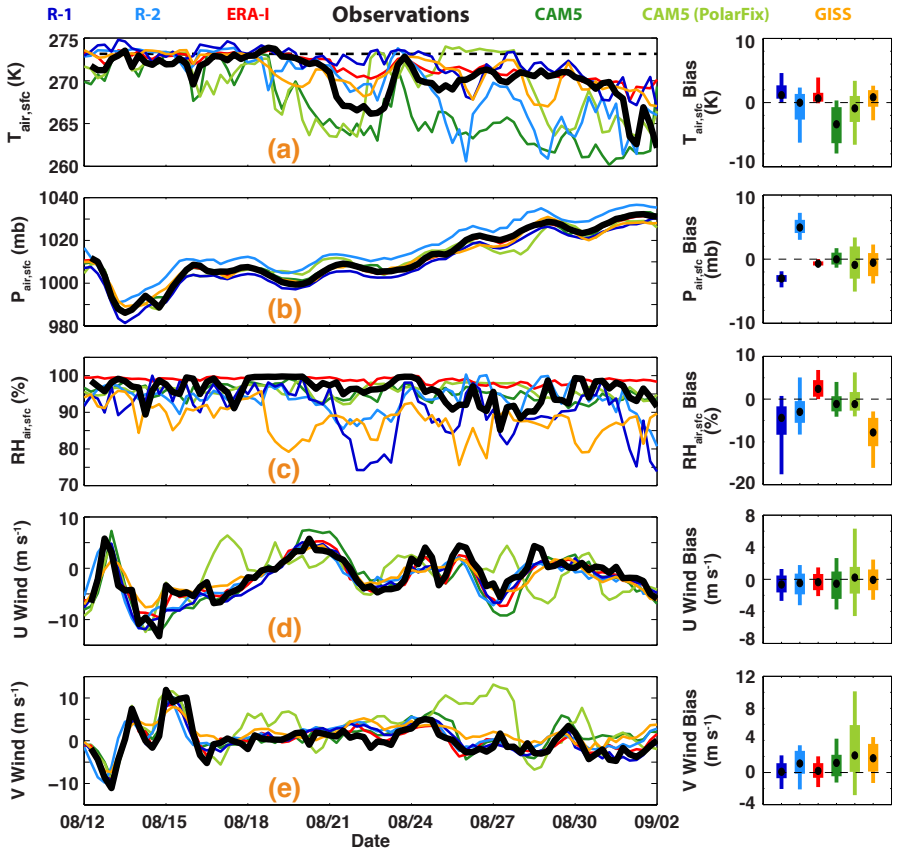

Fig. 2. Time series of basic meteorological quantities (from top to bottom: Surface air temperature, surface air pressure, surface air relative humidity, 10-meter U wind component, and 10m V wind component. Included are lines for the observations from Oden (bold black line), CAM5 (standard version, dark green; CAM5-PF, light green), the GISS-ModelE2 (orange), and R-1 (dark blue), R-2 (light blue) and ERA-I (red) reanalyses. ERA-I, R-1 and R-2 lines represent 6-hourly analysis (0 hr) time instantaneous values, while the CAM5 lines depicts the an interpolated value for the analysis time from the model 1-hourly averages. The right-hand side of the figure includes distributions depicting the median (black dot), IQR (wide bar) and 10th-90th percentiles (thin bar) of the differences between simulated and observed values.

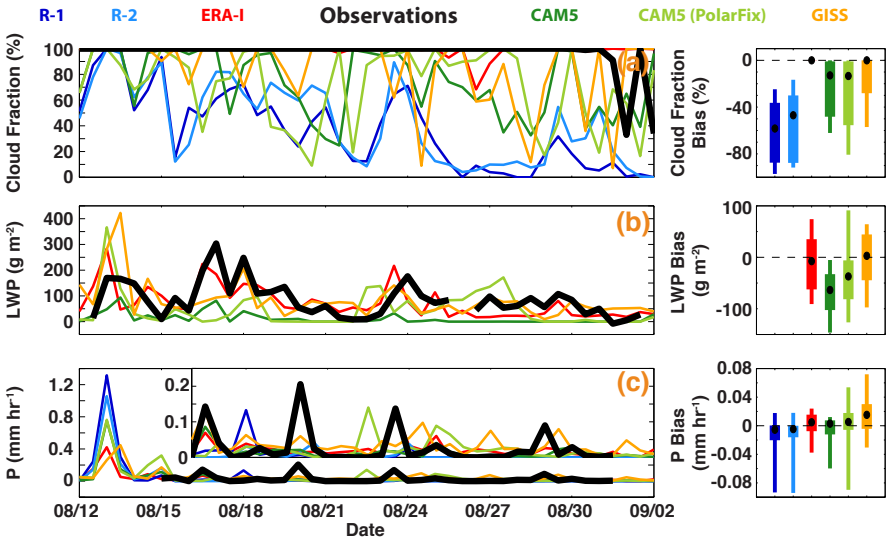

Fig. 3. Time series of common cloud and precipitation quantities (from top to bottom: average cloud fraction, average liquid water path, and average precipitation rate). Included are lines for the observations from Oden (bold black line), CAM5 (standard version, dark green; CAM5-PF, light green), the GISS-ModelE2 (orange), and R-1 (dark blue), R-2 (light blue) and ERA-I (red) reanalyses. The insert in (c) modifies the scale of the vertical axis to better distinguish between models where all have low values. All lines represent six-hour average values. Please note that R-1 and R-2 do not provide liquid water path, and are therefore not included in the center figure. The right-hand side of the figure includes distributions depicting the median (black dot), IQR (wide bar) and 10th-90th percentiles (thin bar) of the differences between simulated and observed values.

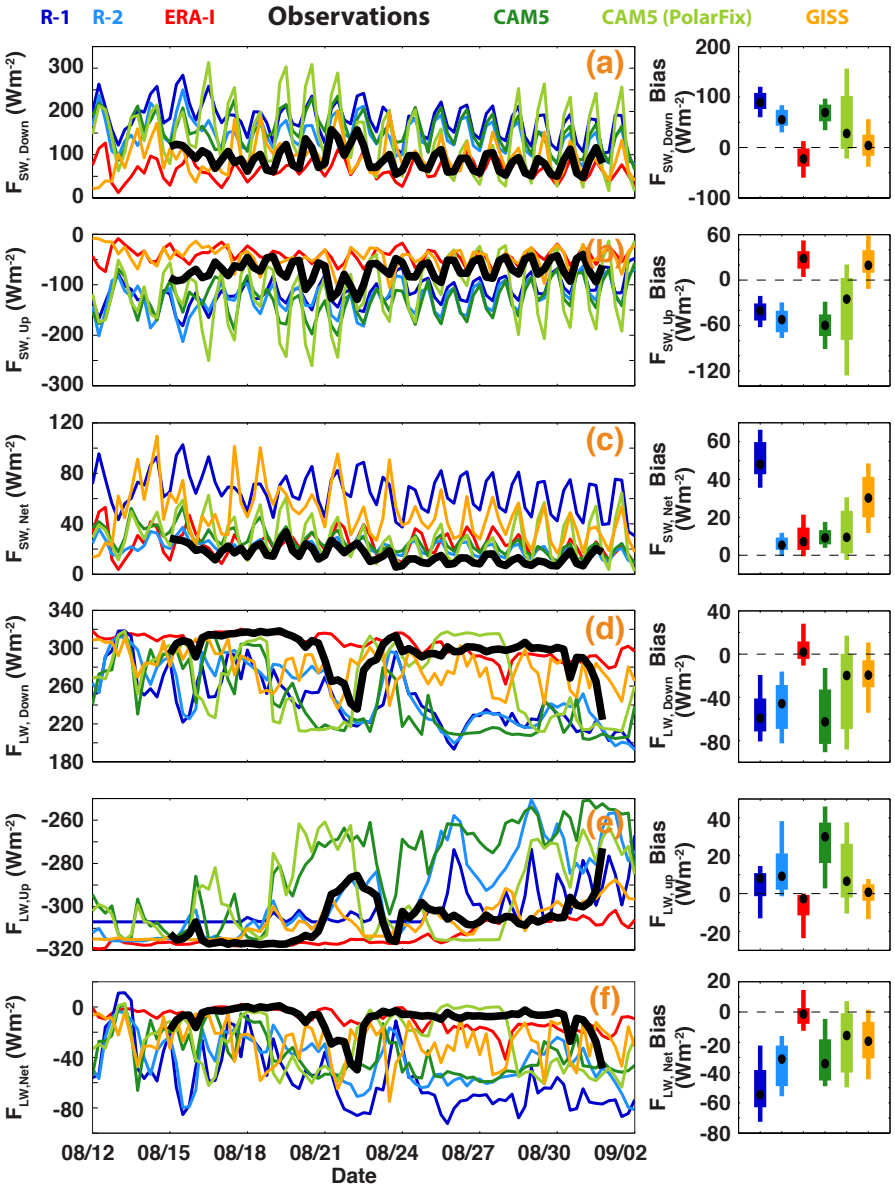

Fig. 4. Time series of surface energy budget terms (from top to bottom: average downwelling shortwave flux density, average upwelling shortwave flux density, average net shortwave flux density, average downwelling longwave flux density, average upwelling longwave flux density, average net longwave flux density). Included are lines for the observations from Oden (bold black line), CAM5 (standard version, dark green; CAM5-PF, light green), the GISS-ModelE2 (orange), and R-1 (dark blue), R-2 (light blue) and ERA-I (red) reanalyses. All lines represent six-hour average values. The right-hand side of the figure includes distributions depicting the median (black dot), IQR (wide bar) and 10th-90th percentiles (thin bar) of the differences between simulated and observed values.

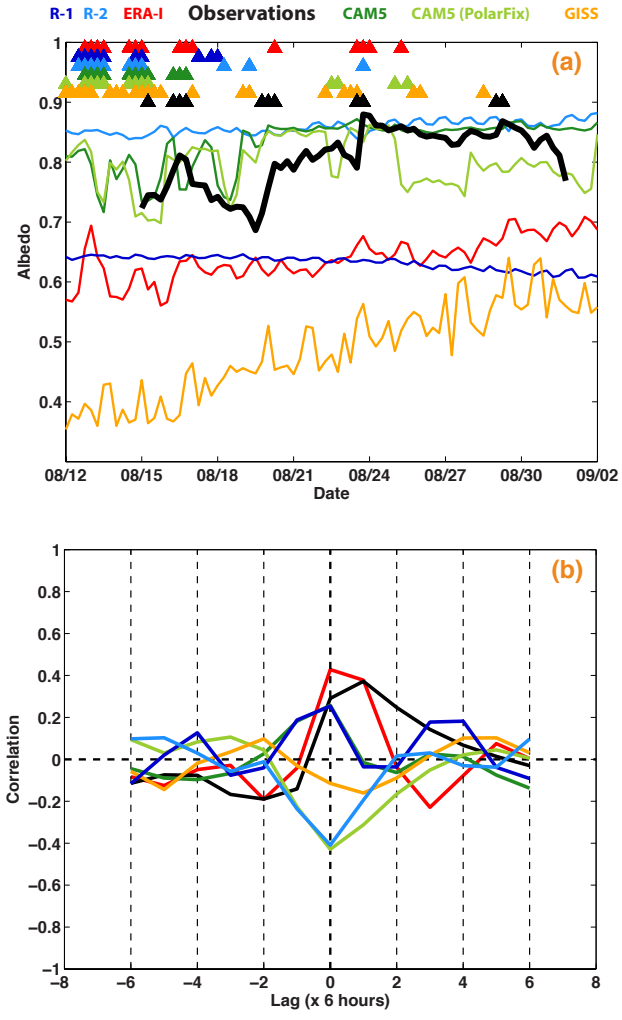

Fig. 5. Time series of surface albedo as calculated from the 6-hour average shortwave radiation terms. Included are lines for the observations from Oden (bold black line), CAM5 (standard version, dark green; CAM5-PF, light green), the GISS-ModelE2 (orange), and R-1 (dark blue), R-2 (light blue) and ERA-I (red) reanalyses. Triangles along the top of the figure represent 6-hourly periods that averaged precipitation of 0.5 mm hr^{-1} or greater.

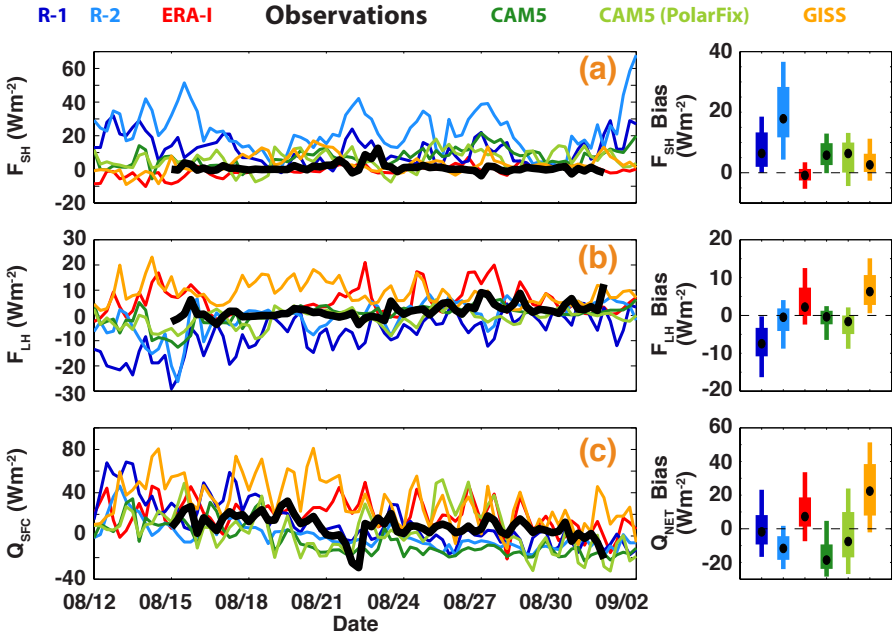

Fig. 6. Time series of surface energy budget terms (from top to bottom: average sensible heat flux density, average latent heat flux density, and net surface energy (radiation, latent heat and sensible heat)). Included are lines for the observations from Oden (bold black line), CAM5 (standard version, dark green; CAM5-PF, light green), the GISS-ModelE2 (orange), and R-1 (dark blue), R-2 (light blue) and ERA-I (red) reanalyses. All lines represent six-hour average values. The right-hand side of the figure includes distributions depicting the median (black dot), IQR (wide bar) and 10th-90th percentiles (thin bar) of the differences between simulated and observed values.

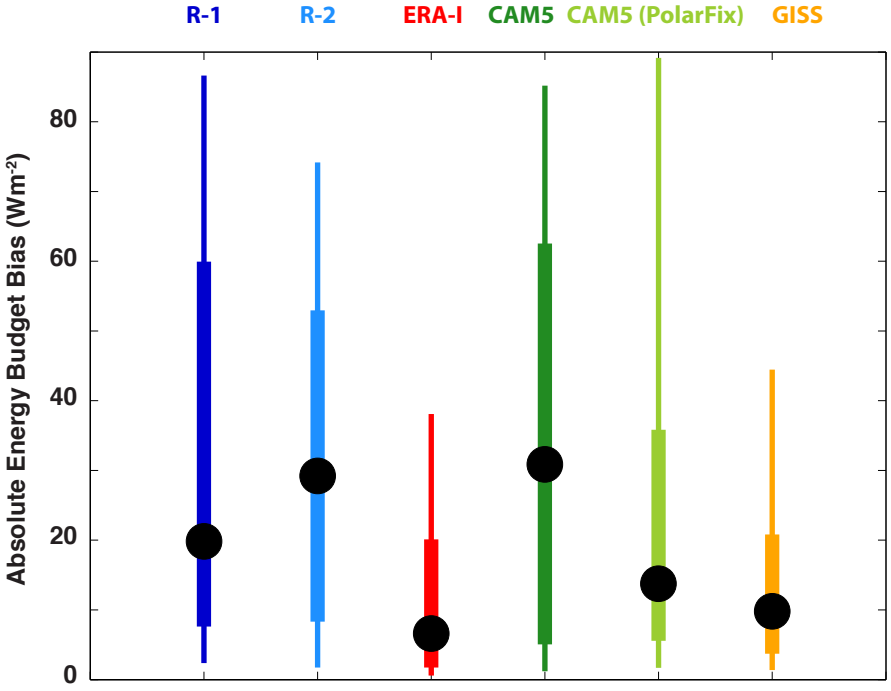

Fig. 7. Combined distributions of the absolute values of biases in LW_{up}, LW_{down}, SW_{up}, SW_{down}, F_{SH} and F_{LH}. Distributions are depicted including the median (black dot), IQR (wide bar) and 10th-90th percentiles (thin bar).

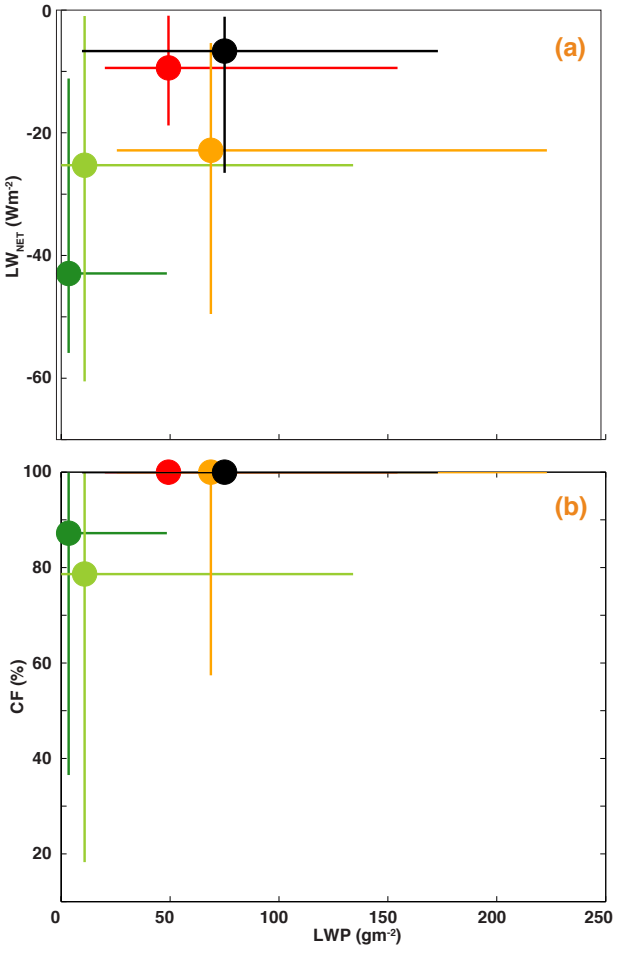

Fig. 8. Figures demonstrating the range of values of liquid water path for observations and models that provide it, as well as it's relationship to surface net longwave radiation (a), and cloud fraction (b). The circles represent the mean value over the ASCOS observational period, and the whiskers demonstrate the extent of the 10th and 90th percentiles of the datasets.

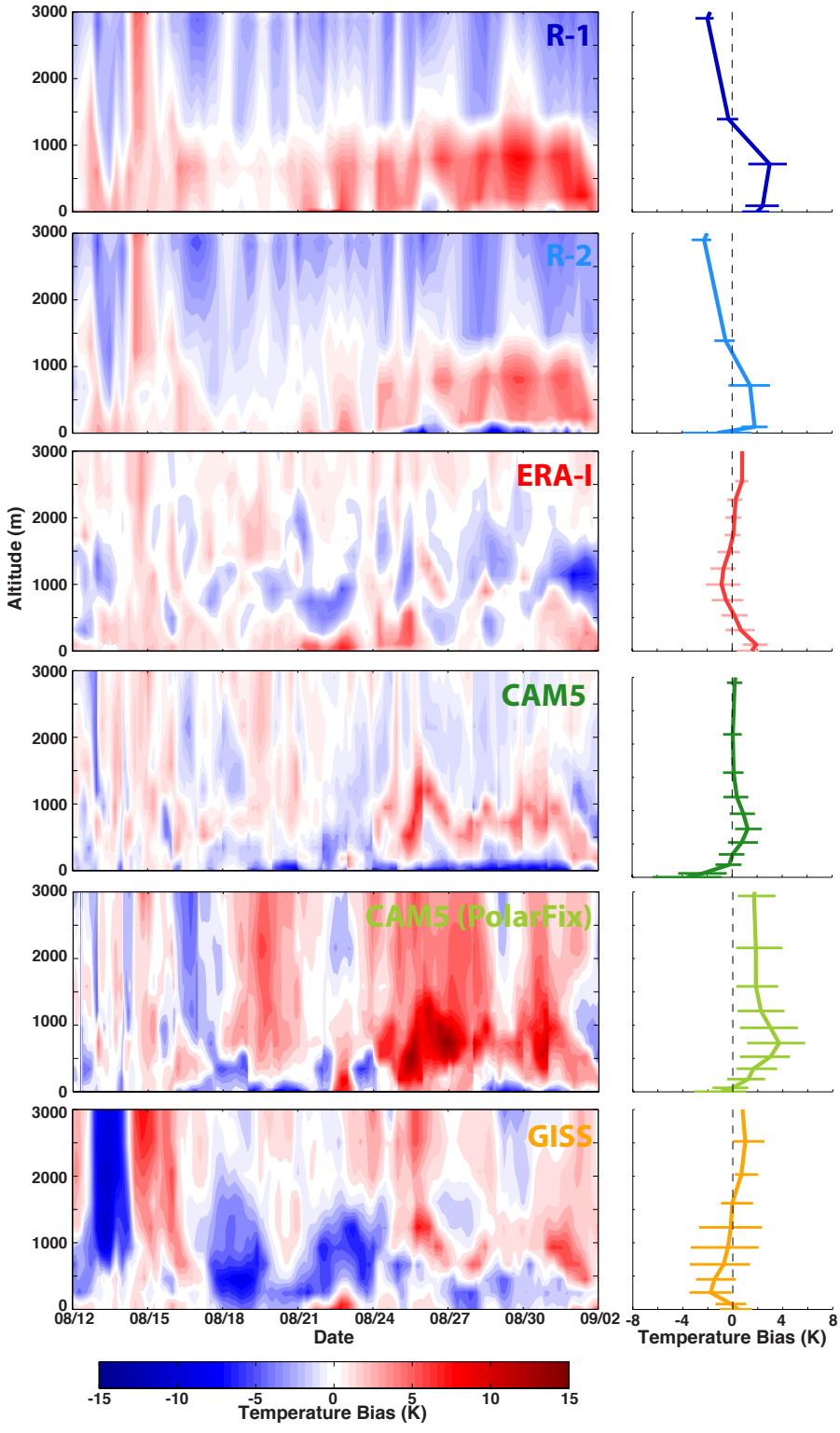

Fig. 9. Time-height cross sections of air temperature bias (model-observations, left). Profiles on the right side illustrate the mean temperature bias with height (bold line) and the interquartile range of the bias (lighter, horizontal lines).

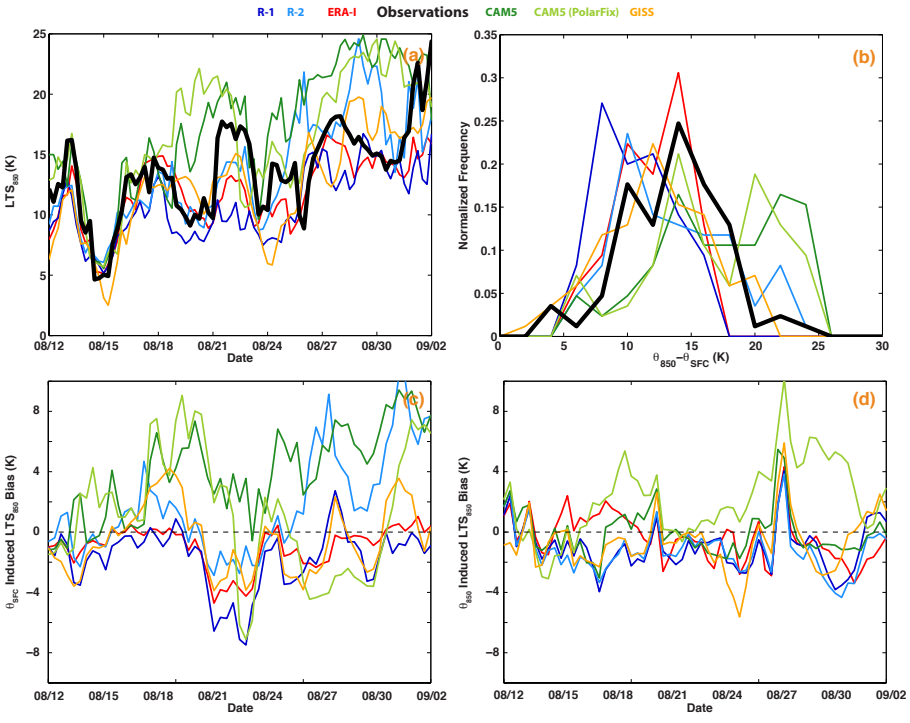

Fig. 10. Timeseries (a) and histograms (b) of lower tropospheric stability (LTS_{850}), as defined by the potential temperature difference between 850 mb and the surface. The lower panels (c,d) illustrate the relative contributions of near-surface potential temperature and the potential temperature at 850 mb to biases in LTS_{850}.

Table 1. Median biases and correlation (in parentheses) to observations for each of the evaluated variables. *Italicized* correlations are not statistically significant at the 95% level. The best performing statistics for each variable are included in **bold**.

Variable	ERA-I	R-1	R-2	CAM5	CAM5PF	GISS ModelE2
T_{sfc}	0.65 (**.82**)	1.15 (.62)	**-0.02** (.54)	-3.39 (.59)	-0.90 (.37)	0.82 (.58)
P_{sfc}	-0.71 (.99)	-3.01 (**1.00**)	4.93 (.99)	**-0.04** (**1.00**)	-0.90 (.96)	-0.55 (.99)
RH_{sfc}	2.39 (**.65**)	-4.39 (.41)	-2.99 (*.06*)	**-1.15** (.27)	-1.17 (*-.15*)	-7.80 (*.10*)
U_{sfc}	-0.36 (**.92**)	-0.65 (.91)	-0.48 (.83)	-0.55 (.84)	0.22 (.41)	**-0.07** (.83)
V_{sfc}	0.20 (**.91**)	**0.09** (.88)	1.12 (.77)	1.20 (.82)	2.13 (.50)	1.78 (.76)
CF	**0.00** (*-.04*)	-58.30 (***.28***)	-47.00 (***.28***)	-12.88 (*-.04*)	-13.38 (*.01*)	**0.00** (*-.08*)
LWP	-7.50 (.55)	N/A	N/A	-63.56 (**.69**)	-37.18 (*.21*)	**2.87** (.42)
Precip	0.005 (**.52**)	-0.006 (*-.01*)	-0.005 (*.05*)	**0.003** (.35)	0.005 (*-.05*)	0.015 (*.28*)
$F_{sw,down}$	-22.73 (.39)	89.10 (.77)	55.14 (**.78**)	69.20 (.71)	27.75 (.70)	**3.97** (.37)
$F_{sw,up}$	28.53 (.41)	-40.66 (.74)	-52.24 (**.77**)	-59.81 (.69)	-25.05 (.67)	**19.57** (.27))
$F_{sw,net}$	7.11 (.31)	48.08 (.62)	**5.33** (**.70**)	9.21 (.74)	9.47 (.62)	30.26 (.52)
$F_{lw,down}$	**1.80** (.35)	-59.32 (.48)	-45.79 (**.52**)	-62.60 (.51)	-19.76 (.25)	-19.41 (*.15*)
$F_{lw,up}$	-2.69 (.35)	8.01 (.27)	9.17 (.51)	30.06 (**.69**)	6.52 (.24)	**0.70** (.49)
$F_{lw,net}$	**-1.43** (.25)	-54.58 (.38)	-31.07 (**.39**)	-34.12 (.26)	-15.50 (*.23*)	-19.30 (*.00*)
F_{SH}	**-0.85** (*-.01*)	6.38 (*-.08*)	17.84 (***.11***)	5.78 (*-.06*)	6.37 (*-.10*)	2.60 (*-.03*)
F_{LH}	2.21 (.25)	-7.47 (.35)	-0.47 (.28)	**-0.37** (**.37**)	-1.61 (*.09*)	6.28 (*-.13*)
Q_{sfc}	7.24 (*.22*)	**2.97** (**.48**)	-11.68 (.42)	-18.62 (**.48**)	-7.57 (.36)	22.38 (.42)