AGU PUBLICATIONS

JGR

Journal of Geophysical Research: Atmospheres

RESEARCH ARTICLE
10.1002/2016JD026174

Key Points:
- An IVT-based AR detection algorithm is applied to 20 year weather and climate simulations from 24 global models
- Among the 17 metrics considered, seven have consistently large errors across all models relative to observational uncertainty
- The importance of model horizontal resolution to the overall quality of AR simulation is suggested

Supporting Information:
- Supporting Information S1

Correspondence to:
B. Guan,
bin.guan@jpl.nasa.gov

Citation:
Guan, B., and D. E. Waliser (2017), Atmospheric rivers in 20 year weather and climate simulations: A multimodel, global evaluation, *J. Geophys. Res. Atmos.*, *122*, 5556–5581, doi:10.1002/2016JD026174.

Received 31 OCT 2016
Accepted 29 APR 2017
Accepted article online 2 MAY 2017
Published online 1 JUN 2017

Atmospheric rivers in 20 year weather and climate simulations: A multimodel, global evaluation

Bin Guan[1,2] **and Duane E. Waliser**[1,2]

[1]Joint Institute for Regional Earth System Science and Engineering, University of California, Los Angeles, California, USA, [2]Jet Propulsion Laboratory, California Institute of Technology, Pasadena, California, USA

Abstract Atmospheric rivers (ARs) are narrow, elongated, synoptic jets of water vapor that play important roles in the global water cycle and meteorological/hydrological extremes. Increasing evidence shows that ARs have signatures and impacts in many regions across different continents. However, global-scale characterizations of AR representations in weather and climate models have been very limited. Using a recently developed AR detection algorithm oriented for global applications, the representation of AR activities in multidecade weather/climate simulations is evaluated. The algorithm is applied to 6-hourly (daily) integrated water vapor transport (IVT) from 22 (2) global weather/climate models that participated in the Global Energy and Water Cycle Experiment Atmospheric System Study–Year of Tropical Convection Multimodel Experiment, including four models with ocean-atmosphere coupling and two models with superparameterization. Multiple reanalysis products are used as references to help quantify model errors in the context of reanalysis uncertainty. Model performance is examined for key aspects of ARs (frequency, intensity, geometry, and seasonality), with the focus on identifying and understanding systematic errors in simulated ARs. The results highlight the range of model performances relative to reanalysis uncertainty in representing the most basic features of ARs. Among the 17 metrics considered, AR frequency, zonal IVT, fractional zonal circumference, fractional total meridional IVT, and three seasonality metrics have consistently large errors across all models. Possible connections between AR simulation qualities and aspects of model configurations are discussed. Despite the lack of a monotonic relationship, the importance of model horizontal resolution to the overall quality of AR simulation is suggested by the evaluation results.

1. Introduction

Atmospheric rivers (ARs) are narrow, elongated, synoptic corridors of enhanced water vapor transport that play important roles in the global water cycle and meteorological/hydrological extremes. While on average covering ~10% of the Earth's circumference at midlatitudes at any given time, ARs account for over 90% of the total poleward water vapor transport at these latitudes [*Zhu and Newell*, 1998]. The amount of water vapor transported by an average AR is roughly equivalent, in terms of mass flux, to the discharge of 17 Mississippi Rivers into the Gulf of Mexico [*Ralph et al.*, 2012], and there are on average 11 ARs globally at any given time based on a multidecade climatology [*Guan and Waliser*, 2015]. These features indicate a critical role for ARs in shaping the global water vapor distribution and imply an importance to global climate variability and change. Meanwhile, the meteorological and hydrological impacts of ARs, particularly in terms of extremes, are becoming better understood over the globe. In particular, the role of ARs as precipitation and flood producers in east coasts of Asia and North America [*Hirota et al.*, 2016; *Mahoney et al.*, 2016] and polar areas [*Gorodetskaya et al.*, 2014; *Neff et al.*, 2014] is becoming recognized and complements the volume of studies that have largely focused on the west coasts of the midlatitude continents (as reviewed in *Gimeno et al.* [2014, 2016]). Compared to their impact on flood risks, the less well-known connection between ARs and hazardous winds across the globe has also recently been highlighted [*Waliser and Guan*, 2017].

The importance of ARs in weather and climate has prompted increasing interests on the behavior of ARs in future climate [*Dettinger*, 2011; *Lavers et al.*, 2013; *Payne and Magnusdottir*, 2015; *Warner et al.*, 2015; *Radić et al.*, 2015; *Gao et al.*, 2015, 2016; *Hagos et al.*, 2016; *Ramos et al.*, 2016; *Shields and Kiehl*, 2016a, 2016b]. A number of these studies involved, to different extents of comprehensiveness, the evaluation of model performance in simulating the behavior of ARs in the current climate, in order to establish credibility for the model projections. Focusing on central California and using a simple AR detection method based on integrated water vapor (IWV) and 925 hPa wind at a single model grid cell, *Dettinger* [2011] showed varied

representation of winter AR frequency and probability distributions of AR IWV and upslope wind by seven Coupled Model Intercomparison Project Phase 3 (CMIP3) models. Using a more complex method that detects ARs intersecting a preselected cross section, *Lavers et al.* [2013] showed varied representation of winter AR frequency in Britain by five CMIP5 models. Using the days with integrated water vapor transport (IVT) exceeding the 99th percentile as proxies for AR landfalls along the west coast of North America, *Warner et al.* [2015] showed considerable spread of the 99th percentile IVT values among 10 CMIP5 models. A more comprehensive AR model evaluation study by *Payne and Magnusdottir* [2015] identified 8 out of a total of 28 CMIP5 models to have generally good representation of ARs in the west coast of North America based on a variety of evaluation metrics.

These previous model evaluation studies provide the motivation and a starting point for more comprehensive and global-scale investigations, which include the development and application of simulation diagnostics and model performance metrics relevant to studies on that scale. Notably, model evaluation studies to date have largely focused on the west coasts of North America and Europe despite increasing evidence of ARs' global signatures and impacts in many regions across different continents. In this regard, a number of key aspects of ARs remain to be systematically evaluated in weather/climate models. Some features of ARs, such as the finding in *Zhu and Newell* [1998] that they account for over 90% of the total meridional IVT while occupying only ~10% of the zonal circumference, are necessarily tied to the global distribution of ARs and, as a result of regional focuses in previous studies, have so far not been evaluated for model simulations. Using a recently developed global AR detection algorithm [*Guan and Waliser*, 2015] and output from a suite of 24 weather/climate models [*Petch et al.*, 2011; *Jiang et al.*, 2015], the main objectives of the current study are to (1) examine key characteristics of ARs (frequency, intensity, geometry, and seasonality) simulated by state-of-the-art weather/climate models; (2) identify and quantify systematic errors in simulated ARs relative to reanalysis products and account for uncertainty between them; and (3) understand possible connections between AR simulation qualities and aspects of model configurations.

2. Data and Methodology

2.1. Multimodel Weather/Climate Simulations

Twenty-year global weather/climate simulations from the Global Energy and Water Cycle Experiment (GEWEX) Atmospheric System Study (GASS)–Year of Tropical Convection (YoTC) Multimodel Experiment [*Petch et al.*, 2011; *Jiang et al.*, 2015] are examined. While the initial goal of the GASS-YoTC project was to understand the processes associated with the Madden-Julian Oscillation represented in weather/climate models [*Klingaman et al.*, 2015], the multimodel, global database produced from the project has favorable features (particularly 6-hourly output of moisture, heat, momentum, and their tendencies) that could benefit investigations of other weather and climate phenomena. A total of 24 models from the project are selected for this study. They provide global, 6-hourly (22 models) or daily (2 models) fields of specific humidity and vector winds at 17 pressure levels between 1000 and 300 hPa, which can be used for the calculation of IVT and AR detection. Zonal and meridional components of IVT are calculated as

$$\mathrm{IVT}_x = -\frac{1}{g}\int uqdp \tag{1a}$$

$$\mathrm{IVT}_y = -\frac{1}{g}\int vqdp \tag{1b}$$

where g is gravitational acceleration, u (v) is zonal (meridional) wind, q is specific humidity, and p is pressure. One realization is available from each model. This should be noted, given potentially different error statistics between a single realization and a multimember ensemble mean of a given model [*Shields and Kiehl*, 2016a, 2016b], which may affect the ranking of the model relative to other models. The native horizontal resolution of the models ranges between 2.8° × 2.8° and 0.42° × 0.42°; their output is archived on a common 2.5° × 2.5° grid.

Most of the models are atmosphere-only, forced by observed sea surface temperatures and sea ice concentrations over the period of 1991–2010. Four models have a coupled ocean-atmosphere-land system and

Table 1. Models Participating in the GASS-YoTC Multimodel Experiment[a]

Model Name	Institution	Native Resolution (Longitude × Latitude, # of Vertical Levels)	Remark
BCC-AGCM2.1	Beijing Climate Center, China Meteorological Administration	T42 (2.8°), L26	
ISUGCM	Iowa State University	T42 (2.8°), L18	
SPCAM3	Colorado State University	T42 (2.8°), L30	Superparameterized, daily archive
UCSD-CAM3	Scripps Institution of Oceanography	T42 (2.8°), L26	
GISS-E2	NASA Goddard Institute for Space Studies	2.5° × 2.0°, L40	
TAMU-CAM4	Texas A&M University	2.5° × 1.9°, L26	
FGOALS-s2	Institute of Atmospheric Physics, Chinese Academy of Sciences	R42 (2.8° × 1.6°), L26	
ACCESS1	Centre for Australian Weather and Climate Research	1.875° × 1.25°, L85	
MetUM-GA3	UK Met Office	1.875° × 1.25°, L85	
MIROC5	AORI/NIES/JAMSTEC, Japan	T85 (1.5°), L40	
CNRM-AM	Centre National de Recherches Météorologiques, France	T127 (1.4°), L31	
EC-GEM	Environment Canada	1.4°, L64	
MRI-AGCM3	Meteorological Research Institute, Japan	T159 (1.125°), L48	
CAM5	National Center for Atmospheric Research	1.25° × 0.9°, L30	
CAM5-ZM	Lawrence Livermore National Laboratory	1.25° × 0.9°, L30	
CFS2	NOAA Climate Prediction Center	T126 (1°), L64	
CWB-GFS	Central Weather Bureau, Taiwan	T119 (1°), L40	
ECEarth3	Rossby Centre, Swedish Meteorological and Hydrological Institute	T255 (0.7°), L91	
GEOS5	NASA Global Modeling and Assimilation Office	0.625° × 0.5°, L72	
NavGEM1	Naval Research Laboratory	T359 (0.42°), L42	
CanCM4	Canadian Centre for Climate Modelling and Analysis	2.8°, L35	Coupled
SPCCSM3	George Mason University	T42 (2.8°), L30	Coupled, superparameterized, daily archive
ECHAM5-SIT	Academia Sinica, Taiwan	T63 (2°), L31	Coupled
ECHAM6	Max Planck Institute for Meteorology, Germany	T63 (2°), L47	Coupled

[a]Atmosphere-only models are listed first, followed by four coupled models. Within each group, the models are sorted in descending order of the native grid cell size. Note that the four coupled models have coarser resolutions than most of the atmosphere-only models. This ordering of the models is used in all figures. Two models are superparameterized and have daily archives, as noted in the remarks; the other models have 6-hourly archives.

therefore do not benefit from prescribed, observed sea surface conditions as in the atmosphere-only models. For example, cold biases in simulated surface temperatures can be associated with weakened IVT and too few ARs [*Lavers et al.*, 2013]. Also, the presence/lack of sea ice in a coupled model may affect the location of atmospheric jets [*Deser et al.*, 2015], which in turn affects AR frequencies [*Hagos et al.*, 2015]. Atmosphere-only models are largely used for weather forecasts and short-term predictions, while coupled models are important tools for long-term climate predictions and projections and are not designed to replicate the day-to-day weather. The errors of the two types of models relative to observations are both of interest (the main objective of the current study), but comparison between the two types of models should be interpreted with caution, given their different nature and applications. Two models, one atmosphere-only and one coupled, are run with superparameterization in place of conventional cumulus parameterization—another difference to note when comparing between the models. The coupled and/or superparameterized models are among those with coarse horizontal resolutions. Table 1 lists the name, institution, and spatial resolution for each of the 24 models. Atmosphere-only models are listed first, followed by the four coupled models. Within each group, the models are sorted in descending order of the native grid cell size. Note that the four coupled models have coarser resolutions than most of the atmosphere-only models. The ordering of the models in Table 1 is used in all figures to be presented.

2.2. Reference Data

The primary reference data set used is the European Centre for Medium-Range Weather Forecasts (ECMWF) Re-Analysis Interim (ERA-Interim [*Dee et al.*, 2011]). Key AR characteristics are well represented in ERA-Interim based on comparisons to aircraft observation of six ARs over the northeastern Pacific [*Ralph et al.*, 2012] and satellite observation of AR landfalls along the western U.S. over 15 years [*Jackson et al.*, 2016]. Also used are two other reanalysis products, namely, the National Aeronautics and Space Administration (NASA) Modern-Era Retrospective Analysis for Research and Applications, Version 2 (MERRA2 [*Bosilovich et al.*, 2015]) and the

National Centers for Environmental Prediction (NCEP)/National Center for Atmospheric Research (NCAR) reanalysis [*Kalnay et al.*, 1996]. Both ERA-Interim and MERRA2 are considered "third generation" reanalysis products, which contain improvements in the representation of the global water cycle compared to the NCEP/NCAR reanalysis. MERRA2 is an update of the original version that takes into account advances in the data assimilation system and a decrease in the amount of observations that can be assimilated by the system. The difference between ERA-Interim and MERRA2 serves as our primary measure of the reanalysis uncertainty, to which model errors relative to ERA-Interim will be compared. Comparisons to NCEP/NCAR provide additional assessment. IVT fields are derived from 6-hourly specific humidity and vector wind at all pressure levels between 1000 and 300 hPa and regridded to have the same spatial resolution as the model output (2.5° × 2.5°). The period of the reference data (1991–2010) matches the period of the weather/climate simulations described above.

2.3. AR Detection

A variety of AR detection techniques, indicative of complementing views on the definition of ARs, can be found in the literature. A broad definition is adopted in the current study, which, as in *Zhu and Newell* [1998], does not impose any predetermined geographical requirements to the existence of ARs. For example, the three flood-producing "tropical plume" events in subtropical northwestern and tropical West Africa reported in *Knippertz and Martin* [2005] were all detected as ARs by *Brands et al.* [2017] (http://www.meteo.unican.es/es/atmospheric-rivers), as also in the current study. While two of the three events are outside of active monsoon season, the third event occurred when the African monsoon was active and therefore could include contribution from the monsoon flow. The ability to encompass transient, filamentary moisture transport in the Asian monsoon region was one of the key considerations made by *Zhu and Newell* [1998] in developing their original AR formulation (see their Figure 5c).

AR detection here is based on the algorithm introduced in *Guan and Waliser* [2015], which considers a combination of geometry and intensity thresholds in identifying ARs (i.e., time steps that satisfy the AR thresholds at any given location). For any given spatial field of IVT, the procedure first extracts a set of "objects" (i.e., connected pixels forming a contiguous area) based on the IVT magnitude threshold specific to each grid cell and season, namely, the greater of the 85th percentile (Figure 1, top and middle) and 100 kg m^{-1} s^{-1}. Additional requirements based on IVT direction (within 45° of the AR shape orientation and with an appreciable poleward component), length (>2000 km), and length/width ratio (>2) are then applied to these objects, resulting in a defined set of ARs. A comparison of this technique to those in three previous studies independently conducted for three regions (western North America, Britain, and East Antarctica) showed over ~90% agreement in detected AR landfall dates, i.e., the ratio between the total number of landfall dates reported for the given region from the previous study and the subset of those dates that has a corresponding landfall detected by the current technique within ±1 day in the same region [*Guan and Waliser*, 2015]. The method is applied to 6-hourly (or daily in two of the models) IVT from each of the reanalysis products and weather/climate simulations described above. The ERA-Interim-derived IVT threshold values are used across all data sets. Using a fixed set of reanalysis-based, as opposed to model-dependent, IVT threshold values for AR detection will facilitate the interpretation of systematic model biases in the detection result. Figure 1 (bottom) shows an example of the AR detection result for an arbitrary 6 h time step of ERA-Interim, where each AR is labeled by a unique color. Note that the AR artificially segmented at 0/360°E is properly considered by the detection algorithm as one AR and so labeled in Figure 1 (bottom).

Although not used for AR detection in our main analysis, we do note that the simulated IVT 85th percentile has an overall positive bias in a number of models compared to ERA-Interim (Figures S1 and S2 in the supporting information). Such positive bias in IVT magnitude will lead to positive bias in AR frequency (or offset negative bias in AR frequency caused by other types of errors) where the observationally derived AR threshold is used, as in our main analysis. A model-dependent IVT threshold, if used, would, in general, weaken the bias in AR frequency (comparing Figure 3 to Figure S3) since simulated ARs in that case would no longer be required to have the same IVT magnitude as observed ARs and, as a result, could strengthen the bias in AR IVT (comparing Figure 4 to Figure S4 and Figure 5 to Figure S5). This sensitivity to the choice of observationally based versus model-based IVT thresholds is an important factor to consider when evaluating AR simulations.

The sensitivity of AR detection to horizontal resolution was also considered. This is relevant to the current study as the 2.5° × 2.5° resolution, at which the model simulations were archived, is relatively coarse

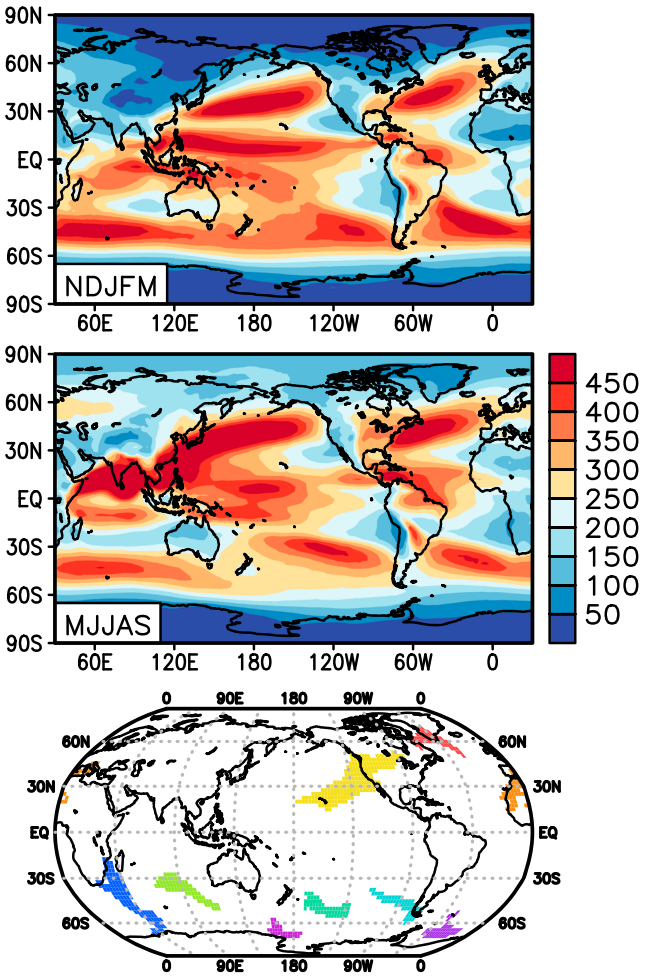

Figure 1. (top and middle) The 85th percentile of IVT magnitude (kg m^{-1} s^{-1}) at each grid cell for the period of 1991–2010 based on ERA-Interim. Shown are November–March (NDJFM) and May–September (MJJAS), for illustration. A total of 12 maps, for 12 overlapping 5 month seasons, are used to threshold 6-hourly (or daily in two of the models) IVT during the center month of the season in the detection of ARs. For example, the map shown in the top (middle) panel is used for detecting ARs in January (July). The same set of maps is used for all data sets. Grid cells with IVT magnitude above the greater of the 85th percentile and 100 kg m^{-1} s^{-1} are retained for AR detection; see text for a brief description of the detection procedures and full details in *Guan and Waliser* [2015]. (bottom) Example output of the AR detection algorithm, showing all ARs detected at an arbitrary 6 h time step of ERA-Interim. Each color indicates a unique AR.

compared to the typical width of ARs (a few hundred km). For this, we ran AR detection for two versions of ERA-Interim IVT that differ only in their horizontal resolution, with one resolution matching the model archive (2.5° × 2.5°) and the other higher resolution (1.5° × 1.5°) representative of the median among the native resolution of the 24 models. A comparison between ARs detected in the two versions of data suggests that the difference in AR frequency and IVT caused by a reasonable coarsening of the horizontal grid is consistently smaller than the difference between different reanalysis products and smaller than the difference between reanalysis and the multimodel ensemble mean (Figure S6). The analysis suggests that the errors introduced by the arbitrary coarsening of the horizontal grid (for the purpose of archiving model output, not to be confused with the resolution at which the models are run) are not an important consideration compared to the reanalysis uncertainty and to the magnitude of model errors and that the 2.5° × 2.5° horizontal resolution used by the model archive, although not optimal, is suitable for evaluating AR simulations.

2.4. Evaluation Methods

To evaluate AR simulations, we first calculate 20 year climatologies for selected AR characteristics (such as AR frequency at each grid cell). For a given simulated climatological field F_s and the corresponding field in the reference (in this case, ERA-Interim) F_r, three error characteristics are related by

$$B^2 + \mathrm{CRMSE}^2 = \mathrm{RMSE}^2 \tag{2}$$

where B is the bias (mean difference between F_s and F_r), RMSE is the root-mean-square error, and CRMSE is the centered RMSE (i.e., RMSE between F_s and F_r after they are respectively centered by taking out the means). We use B to characterize the sign and magnitude of any systematic bias in a simulated climatology and RMSE to characterize the mean magnitude of errors in a simulated field. Equation (2) suggests that the total RMSE is contributed by the systematic bias B, as well as pattern dissimilarities represented by CRMSE.

To help appreciate the magnitude of the terms in equation (2) relative to the variability of the reference field, we normalize each term by the standard deviation of F_r, giving

$$B_{\text{norm}}^2 + \text{CRMSE}_{\text{norm}}^2 = \text{RMSE}_{\text{norm}}^2 \tag{3}$$

In Taylor diagrams [*Taylor*, 2001], $\text{CRMSE}_{\text{norm}}$ can be shown succinctly with correlation and the standard deviation ratio between F_s and F_r (e.g., see Figure 6). Denoting $\text{RMSE}_{\text{norm}}$ by E, we define

$$E^* = \frac{E - E_{\text{median}}}{E_{\text{median}}} \tag{4}$$

where E_{median} is the median E across all models for a given F_r. The relative error, E^*, characterizes the performance of an individual model relative to the "typical" (i.e., median) model within the ensemble (in this case, 24 models) for a given F_r and can be compactly shown in a "portrait diagram" [*Gleckler et al.*, 2008] for a collection of models and evaluation metrics. The "error" E is also calculated for MERRA2 relative to ERA-Interim, which, as previously mentioned, serves as our primary measure of reanalysis uncertainty. The same is also calculated for NCEP/NCAR relative to ERA-Interim, which, given the former is from a much earlier generation of reanalysis hence larger differences from ERA-Interim, provides a secondary, less stringent uncertainty measure.

Besides the evaluation of climatological fields as described above, we also evaluate metrics of relevance to individual ARs. For each model, let a certain AR characteristic (e.g., widths of individual ARs) be represented in a set $\{y_s: s = 1, 2, \ldots, S\}$, where S is the total number of ARs identified in a 20 year simulation. The corresponding AR characteristic in the reference is represented by the set $\{x_r: r = 1, 2, \ldots, R\}$, where R is the total number of reference ARs. For the given AR characteristic, the model bias, B, is simply the difference between the mean of y_s and the mean of x_r. B_{norm} is based on normalization by the standard deviation of x_r. In addition, histograms H_s and H_r (using the same set of bins) are calculated for y_s and x_r, from which E and E^* are calculated as in equations (2)–(4) but in the histogram space. Errors based on 2-D histograms (such as the joint distribution of AR length versus width) are calculated similarly in the 2-D histogram space.

3. Spatial Distributions of AR Frequency and IVT

The spatial distributions of AR frequency and IVT are shown in Figure 2. As in all other figures, the models are grouped into atmosphere-only (rows 1–5) and coupled (row 6) and within each group sorted in descending order of the native grid cell size. Note that the four coupled models have coarser resolutions than most of the atmosphere-only models. As mentioned earlier, direct comparison between the different types of models (atmosphere-only versus coupled and conventional versus super parameterization) should be made with caution, given their different nature and applications, particularly their different bias sources. At each grid cell, AR frequency is calculated as the fraction of time steps the grid cell falls within detected AR shapes, and AR IVT is based on averaging over those time steps with ARs. Note that given the sampling and length of records being considered, a frequency of 10% would indicate ~3000 ARs for a 20 year period with 6 h sampling. The ERA-Interim result resembles that in *Guan and Waliser* [2015], except AR detection here was based on a slightly different period and a 2.5° × 2.5° grid to match model simulations. Results based on MERRA2 compare well with those on ERA-Interim, while AR frequencies are a bit higher in the case of NCEP/NCAR. Larger differences in the case of NCEP/NCAR are somewhat expected, given that the other two products are from a more current generation. The broad-scale patterns of AR frequency and IVT common in the reanalyses are well represented in the ensemble mean of the models and are reasonably captured in most of the individual simulations. These include enhanced AR frequencies and IVT magnitudes in midlatitude ocean basins compared to land areas and other latitudes and the eastward and poleward prevailing direction of AR IVT. In interior Antarctica, no ARs are detected over the data set period (white shading in Figure 2). Along the west coast of North America, the north-south gradient in AR frequency and the reduction of AR

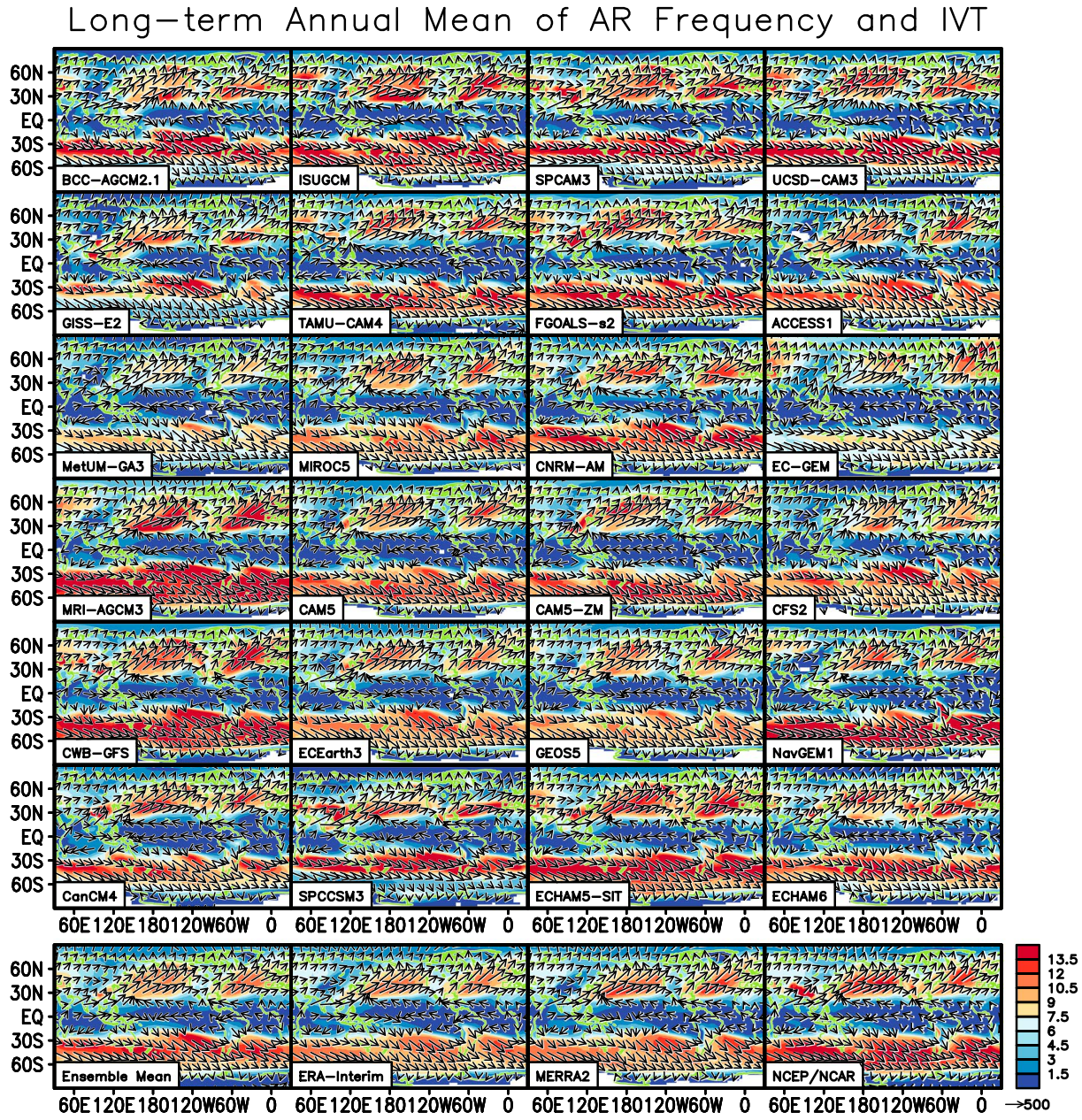

Figure 2. AR frequency (percent of time steps; shading) and mean AR IVT (kg m^{-1} s^{-1}; arrows) at each grid cell. The white shading in limited areas indicates no AR detected over the analysis period. Rows 1–5 (row 6) are for atmosphere-only (coupled) models, sorted in descending order of the native grid cell size as in Table 1. Note that the four coupled models have coarser resolutions than most of the atmosphere-only models. The same ordering of the models is used in subsequent figures. The isolated bottom row shows the ensemble mean over the 24 models and three reanalyses as references. ERA-Interim and MERRA2 data were regridded to the 2.5° × 2.5° model resolution before AR detection, to facilitate model evaluation. NCEP/NCAR has the same resolution as the models.

frequency toward the continental interiors [*Rutz et al.*, 2014; *Guan and Waliser*, 2015] are reasonably captured in approximately one half of the models.

Model biases in AR frequency and IVT are shown in Figures 3–5. About half of the models show notable positive biases in AR frequency beyond reanalysis uncertainties in the extratropical Pacific/Atlantic and extending to the west coasts of North/South America and Europe where AR landfalls play important roles in the regional weather and hydrology (Figure 3). Among these are four coarsest resolution atmosphere-only models and three of the four coupled models. Two models (Met Office Unified Model Global Atmosphere 3

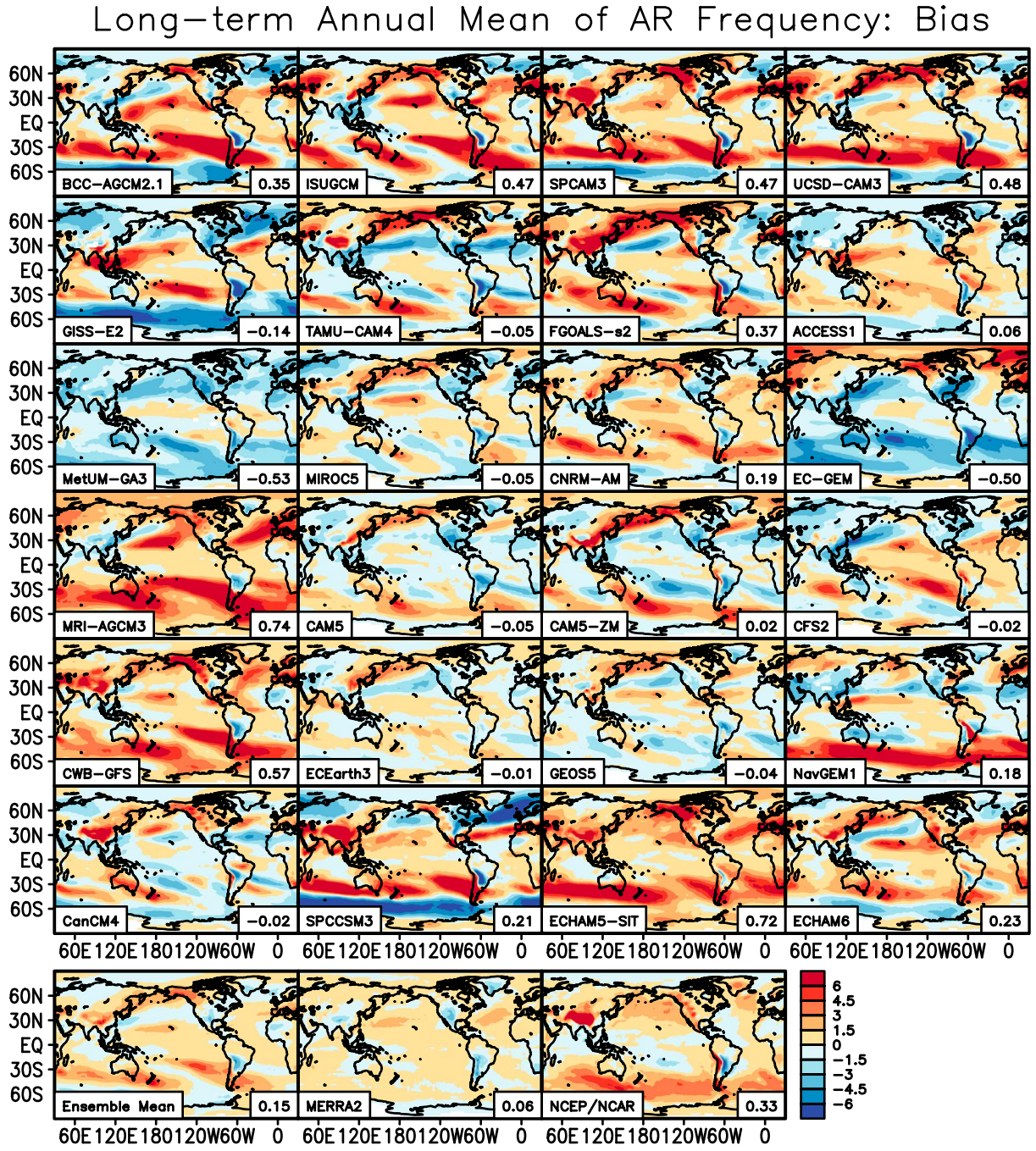

Figure 3. Mean model biases in AR frequency (percent of time steps) relative to ERA-Interim based on Figure 2. Also shown is the mean difference of MERRA2 (NCEP/NCAR) relative to ERA-Interim, serving as the primary (secondary) measure of reanalysis uncertainty. In each panel, the number at the lower right corner indicates the normalized global mean biases (B_{norm} in equation (3)), calculated only over the regions where the ERA-Interim AR frequency is over 1% (i.e., regions with few observed ARs, mostly the tropics and Antarctica, are excluded from the calculation of globally averaged error statistics, both here and in subsequent figures, although these regions are retained in the display of spatial maps).

(MetUM-GA3) and Environment Canada Global Environmental Multiscale Model (EC-GEM)) show a dominance of negative biases in AR frequency across the majority of the globe. Negative biases are also notable in two other models (Goddard Institute for Space Studies (GISS) E2 and Superparameterized Community Climate System Model 3 (SPCCSM3)) in the Southern Ocean and northern North Atlantic. The considerable

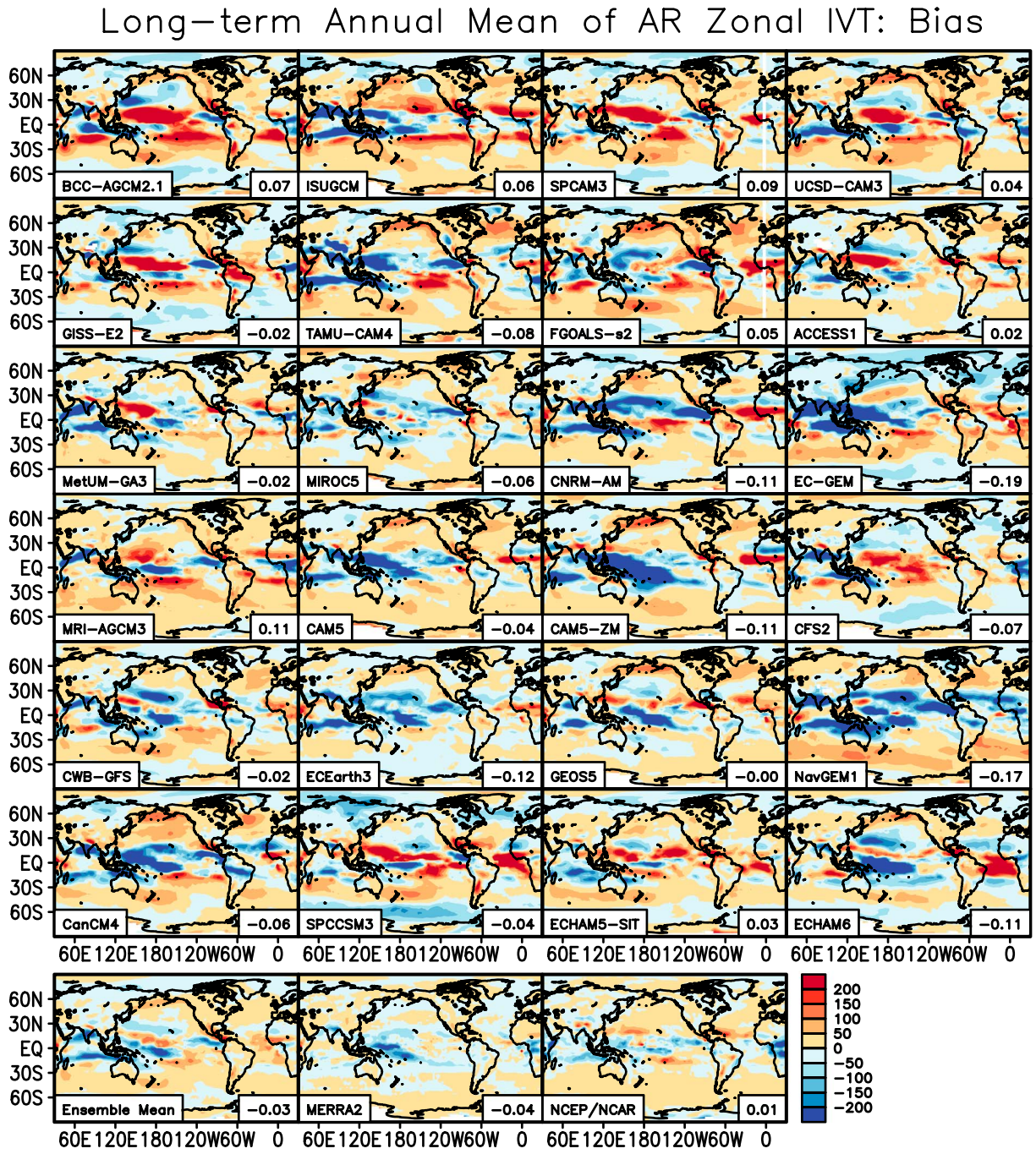

Figure 4. As in Figure 3 but for AR zonal IVT (kg m^{-1} s^{-1}).

but nonmonotonic relationship between AR frequency biases and native model resolutions is similar to what *Payne and Magnusdottir* [2015] found in the northeastern Pacific for 28 CMIP5 models. The relationship is somewhat consistent with the sensitivity experiments in *Hagos et al.* [2015], although they found that the sensitivity of AR frequency to model resolution was notable only in the southeastern Pacific. Biases in landfalling AR frequencies in the west coasts of North America and Europe have been largely attributed to biases in the near-surface subtropical jet locations [*Gao et al.*, 2015, 2016]. Biases in AR zonal IVT are, in general, stronger than those in meridional IVT (Figures 4 and 5; note different scales between the two figures). For

zonal IVT, the tropics/subtropics tend to have stronger biases than higher latitudes. Such a consistent latitudinal contrast is not found for biases in meridional IVT. Given that the two components of IVT share the same specific humidity in their calculations, stronger biases in AR zonal IVT compared to meridional IVT in the tropics/subtropics suggest the dominance of biases in zonal winds over the biases in specific humidity in these regions. In extratropical regions, both specific humidity and winds may contribute comparably to biases in IVT, as suggested in *Payne and Magnusdottir* [2015]. To prevent locations with few ARs from disproportionally affecting model evaluation, locations where AR frequency is below 1% (mostly the tropics and Antarctica) are excluded from calculations of globally averaged error statistics here and in subsequent analysis. Averaged globally and normalized by the standard deviation of the reference field, biases (B_{norm} in equation (3)) are stronger in AR frequency than in both components of AR IVT in 18 of the 24 models (see numbers in the lower right corner of each panel in Figures 3–5; the signs in the case of meridional IVT are such that positive/negative values correspond to poleward/equatorward biases).

Pattern similarities between simulated and ERA-Interim AR frequency and IVT are evaluated using Taylor diagrams shown in Figure 6, where the left column is based on directly comparing maps shown in Figure 2 and the right column is based on the same set of maps but after the zonal mean is removed from each map to highlight contributions apart from the climatological north-south gradient. As in other figures, the models are grouped into atmosphere-only (red) and coupled (green) and within each group sorted in descending order of the native grid cell size, here indicated by the sizes of the colored dots. Larger pattern dissimilarities exist in the case of AR frequency, indicated by the larger scattering of the points compared to the cases of AR zonal and meridional IVT. The scattering is even more pronounced when the zonal mean AR frequency is removed. As in the case of biases, the coarsest resolution atmosphere-only models (red dots with largest sizes) and the coupled models (green dots) tend to exhibit the largest pattern errors in AR frequency. Errors in all (most) individual models and the ensemble mean are beyond the primary (secondary) measure of reanalysis uncertainty. Model errors in AR IVT are smaller compared to those in AR frequency but are still beyond reanalysis uncertainties. Better performance of the models in AR IVT versus frequency, both in terms of overall biases (Figures 3–5) and pattern similarities (Figure 6), suggests that realistic simulation of AR occurrence is more challenging than AR intensity in terms of IVT. We note that AR intensity can also be measured in terms of precipitation and surface wind from an impact perspective, in which case the duration of individual ARs will also be relevant [*Ralph et al.*, 2013].

The overall importance of ARs in the global water cycle is represented by the finding that they account for the vast majority of the total meridional IVT in the midlatitudes while occupying only a small fraction of the zonal circumference. This was initially found in *Zhu and Newell* [1998] and reaffirmed in *Guan and Waliser* [2015] based on more contemporary data and a more sophisticated algorithm for AR detection. The performance of the models in simulating this fundamental aspect of ARs is evaluated in Figures 7 and 8. The fractional total meridional IVT accounted for by ARs at a given latitude is calculated as

$$f_{\text{IVT}_y} = \frac{\sum_{k=1}^{K} \text{IVT}_y(\lambda_k)}{\sum_{i=1}^{I} \text{IVT}_y(\lambda_i)} \tag{5}$$

where λ is longitude and K (I) is the number of AR grid cells (total number of grid cells) at the given latitude. The fractional zonal circumference accounted for by ARs is simply

$$f_{\text{circ}} = \frac{K}{I} \tag{6}$$

The AR fractional total meridional IVT has notable biases in the majority of the models (Figure 7). The AR fraction is evaluated only for 30–60° latitude (as shown) where the fraction values are relatively stable and below 100% (ARs occur most frequently around this region; see Figure 2). The fraction values are less coherent equatorward of this band and become overwhelmingly large poleward of this band (not shown). Three of the four coarsest resolution atmosphere-only models have large positive biases over most of the latitudes between 30 and 60°, which are more pronounced in the Southern Hemisphere. Large biases in both signs are also seen with some of the higher resolution models and three of the coupled models. One outlying

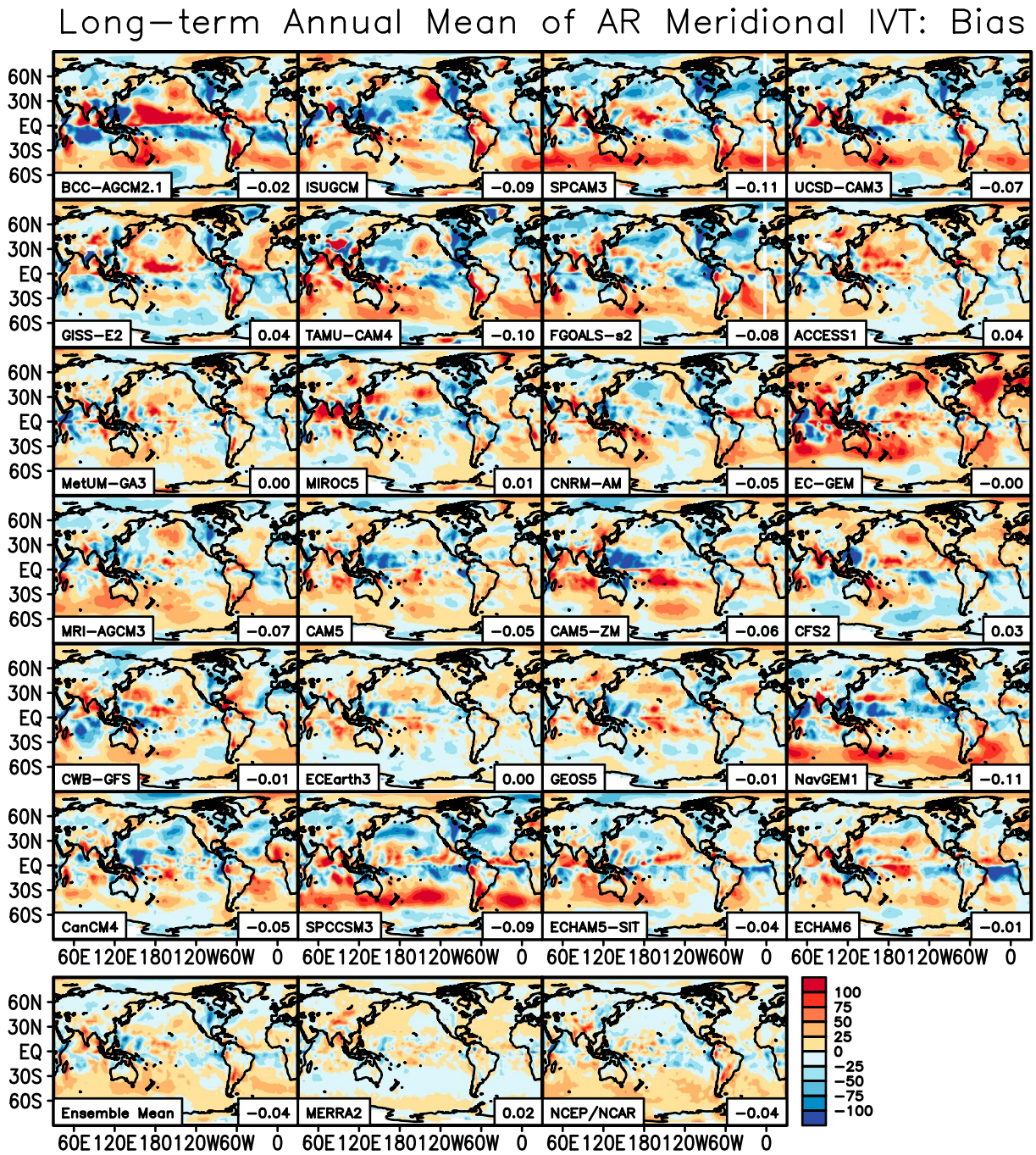

Figure 5. As in Figure 3 but for AR meridional IVT (kg m^{-1} s^{-1}). The signs for the normalized biases indicated at the lower right corner of each panel are such that positive/negative values indicate poleward/equatorward biases. Note the different scales between Figures 4 and 5.

model is EC-GEM, in which case the total IVT has exceptionally strong northward biases in both hemispheres. Interestingly, the AR IVT itself is reasonably represented in this model. Due to the abnormal behavior in its total IVT (thus overly small AR fraction), this model is excluded from the calculation of the ensemble mean in Figure 7; the latter represents the AR fraction remarkably well at most of the latitudes between 30 and 60°.

As in the cases above, biases in AR fractional zonal circumference tend to be largest with the coarsest resolution atmosphere-only models and the coupled models (Figure 8). Also noteworthy is the bias of NCEP/NCAR relative to ERA-Interim. In these cases, the strongest biases, positively signed, tend to occur in the southern

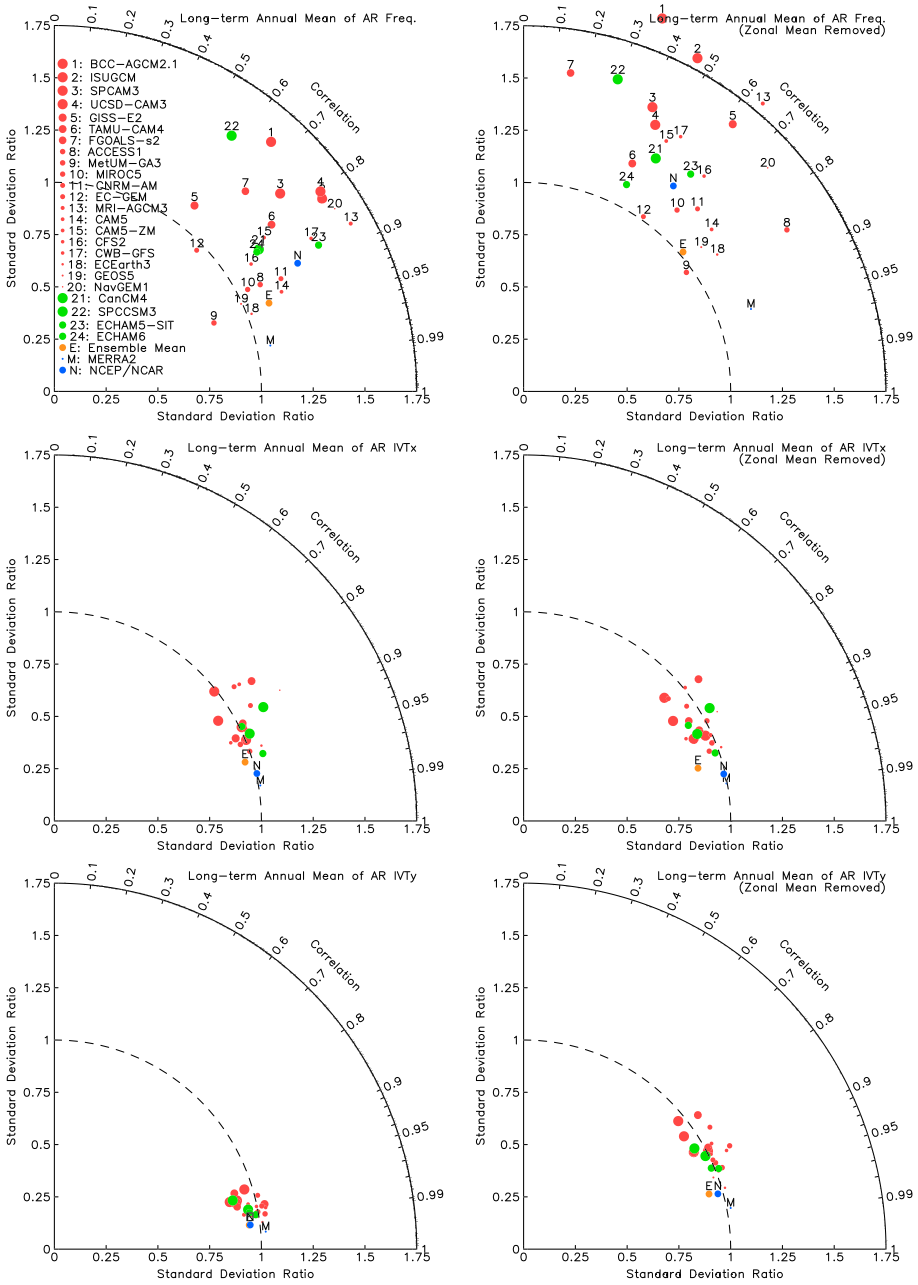

Figure 6. Taylor diagrams comparing the spatial distribution of simulated (top) AR frequency and (middle and bottom) AR zonal and meridional IVT to ERA-Interim. Also shown is the comparison of MERRA2 (NCEP/NCAR) to ERA-Interim, serving as the primary (secondary) measure of reanalysis uncertainty. The left column is directly based on the maps shown in Figure 2, while the right column is based on removing the zonal mean from each map. In these diagrams, the radial distance of a colored dot to the origin indicates the standard deviation ratio between the simulation and the reference, the azimuthal angle indicates the correlation, and the distance between a dot and the reference point (where the dashed arc crosses the horizontal axis) indicates the normalized centered RMSE ($CRMSE_{norm}$ in equation (3)). The dots are sized proportional to the native grid cell sizes. In the bottom four panels, the dots for individual models are not numbered for better visualization of the dots themselves.

extratropics, which roughly corresponds to large positive biases in AR fractional total meridional IVT. To understand this relationship, equation (5) can be rewritten as

$$f_{\mathrm{IVTy}} = \frac{K \cdot \overline{\mathrm{IVT}_y(\lambda_k)}}{I \cdot \overline{\mathrm{IVT}_y(\lambda_i)}} \tag{7}$$

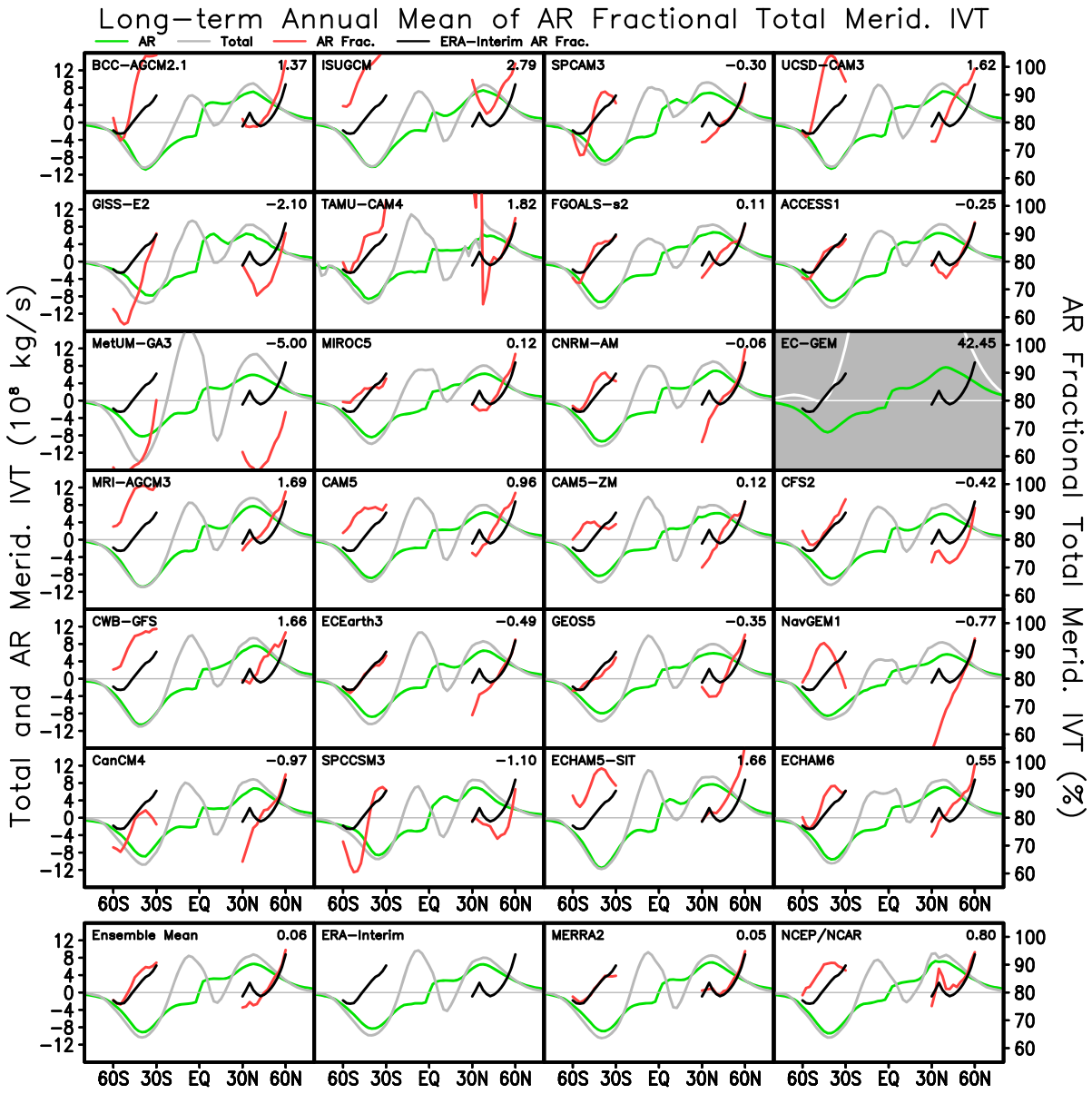

Figure 7. Total zonally integrated meridional IVT (kg s^{-1}; gray), the part associated with ARs (green), and the fraction of the total zonally integrated meridional IVT accounted for by ARs (%; red). Note for EC-GEM (shaded panel) the total IVT is shown in white, and the AR fraction values are completely off the scale due to exceptionally large biases in total IVT; this model is not included in the calculation of the ensemble mean in this figure. The AR fraction based on ERA-Interim, the reference, is reproduced in each panel (black). The AR fraction is calculated only for 30–60° latitude where the fraction values are relatively stable and below 100%. In each panel, the number at the upper right corner indicates the normalized mean bias of the AR fractional total meridional IVT.

where the over bars represent zonal averaging over AR grid cells (numerator) or all grid cells (denominator) at a given latitude. The number of AR grid cells, K, is determined by

$$K = \sum_{n=1}^{N} w_n = N \cdot \overline{w_n} \tag{8}$$

where N is the number of ARs at the given latitude (e.g., 4 ARs at 45°S in the bottom panel of Figure 1) and w_n is the number of grid cells spanned by each AR in the zonal direction. A positive bias in K, meaning a positive bias in zonal mean AR frequency when summarized over many time steps, would contribute to

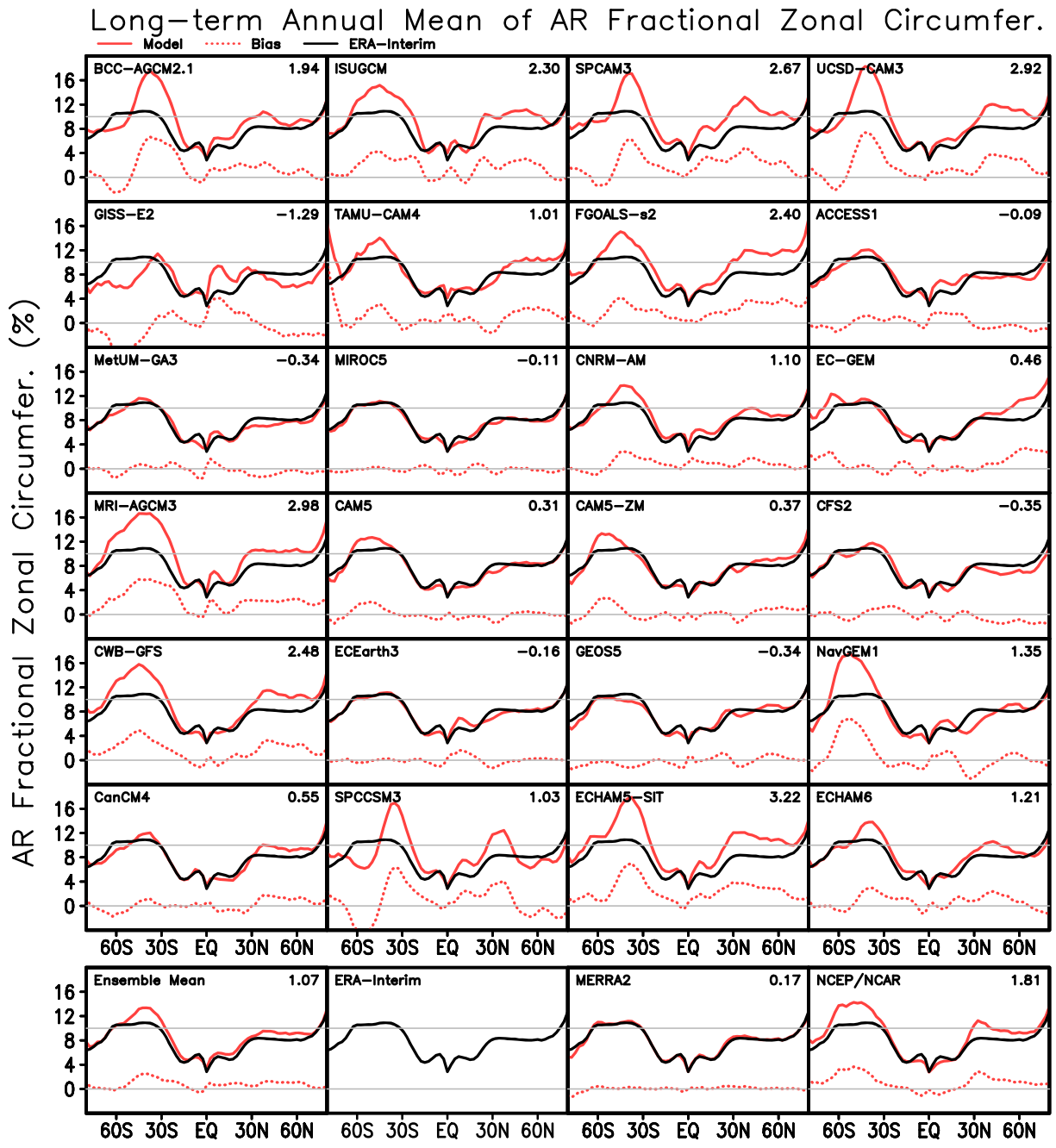

Figure 8. Total zonal distance spanned by all ARs at a given latitude expressed as percent of the zonal circumference at that latitude (red solid). The ERA-Interim result is reproduced in each panel (black), based on which the biases are calculated (red dotted). In each panel, two horizontal lines are drawn at values of 0 and 10 for reference, and the number at the upper right corner indicates the normalized mean bias of the AR fractional zonal circumference for 30–60° latitude (the same band used in Figure 7).

positive biases in both f_{circ} and f_{IVT_y} according to equations (6) and (7). This is consistent with the results shown in Figures 3, 7, and 8. For example, positive biases in AR frequency around 30–45°S in the four coarsest resolution atmosphere-only models (Figure 3) are all associated with positive biases in f_{IVT_y} (Figure 7) and f_{circ} (Figure 8). Zonal mean AR meridional IVT (the upper term after the dot in equation (7)) is likely not a primary contributor to biases in f_{IVT_y}, given relatively small biases in AR meridional IVT compared to those in AR frequency (see normalized biases indicated in Figure 3 versus Figure 5).

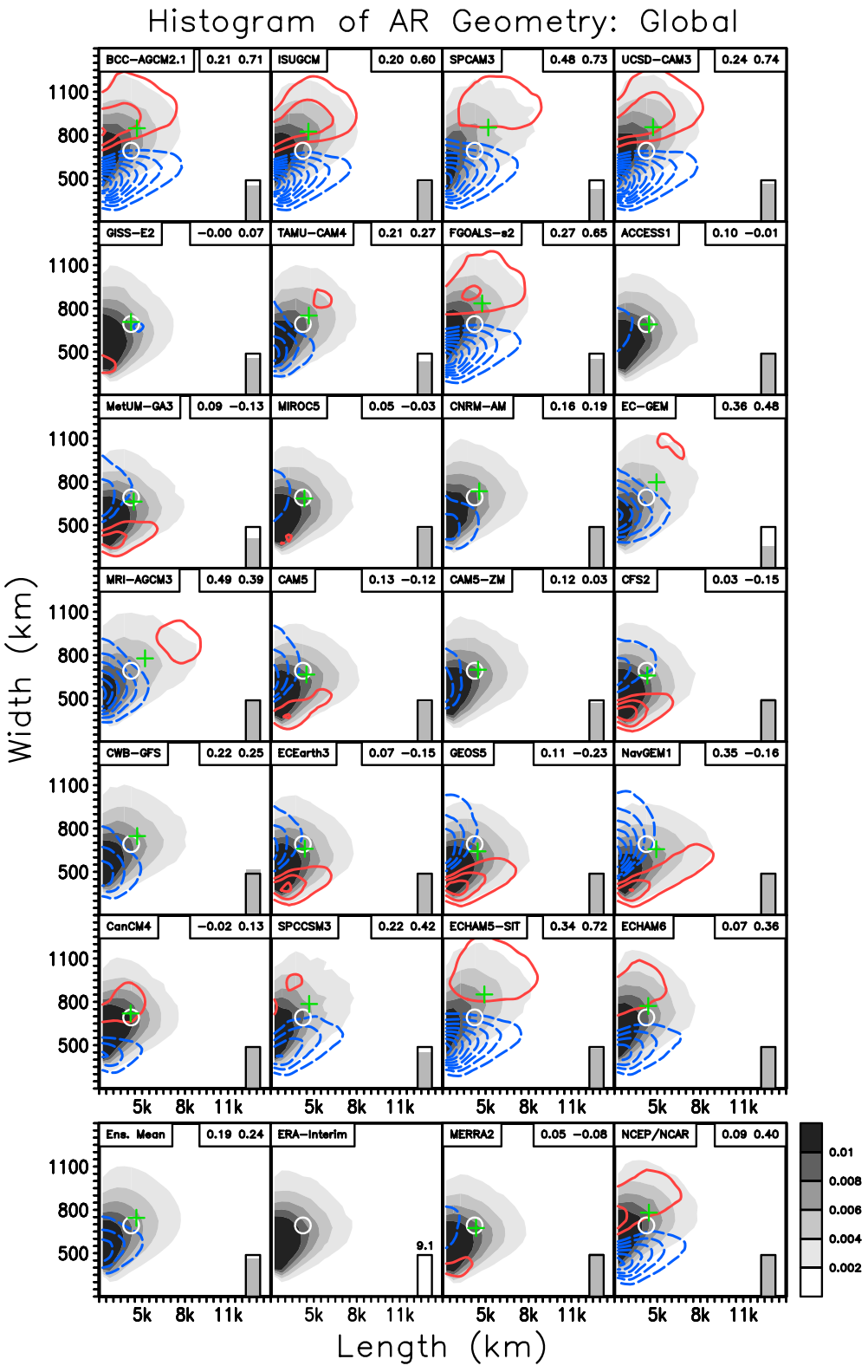

Figure 9. Joint histogram of AR length versus width (shading), and the bias relative to ERA-Interim (negative in blue, positive in red, and contoured at half of the shading interval). In each panel, the white circle (green plus sign) indicates the mean AR length and width based on ERA-Interim (model). Note different scales for AR length and width. Normalized biases in AR length and width are indicated by the numbers in the upper right corner of each panel. The black bar outline repeated in the lower right corner of each panel indicates the number of individual ARs globally per time step averaged over 20 years based on ERA-Interim (9.1), and the gray fill indicates the corresponding value from the simulations or other reanalyses.

Equation (8) suggests that biases in K could come from both N (too many/few ARs) and w_n (too wide/narrow ARs), which are examined in the next section.

4. Probability Distributions of AR Geometry and Intensity

Model performance in simulating the geometry of individual ARs is evaluated by examining the joint probability distribution (i.e., histogram) of AR length versus width. In *Guan and Waliser* [2015] (see their Figure 6), it was shown that AR lengths follow a monotonically decreasing probability distribution (i.e., probability of occurrence decreases as AR length increases), while AR widths follow a positively skewed distribution (probability is larger for medium-sized AR widths, reduced for narrower/wider ARs with a long tail at large widths). As such, the joint AR length-width distribution is characterized by maximum probability of occurrence for ARs with small lengths and medium widths and features a slant pattern where the most probable AR width increases as AR length increases (Figure 9, shading). The shape of the pattern is reasonably represented in the model simulations. However, the pattern is displaced toward larger AR width for the four coarsest resolution atmosphere-only models and the coupled models, indicating too many (few) ARs with large (small) width in these cases (Figure 9, contours). The pattern is displaced toward the opposite direction for the three highest resolution models (Earth system model based on ECMWF's seasonal forecasting system 3 (ECEarth3), Goddard Earth Observing System Model 5 (GEOS5), and Navy Global Environmental Model 1 (NavGEM1)), indicating too many (few) ARs with small (large) width, although the biases are much weaker than in the coarsest resolution models. The biases in the three highest resolution models could be an artifact of the reference data (ERA-Interim) being on a coarser native grid than the three models, given the sign of the relationship suggested by the coarse-resolution models. In 18 of the 24 models, AR width has larger normalized biases than AR length.

The number of samples that contributed to the calculation of the above probability distributions is also examined to understand if the models simulate too few or too many individual ARs. For this, the average number of individual ARs detected per time step of global IVT is calculated for each data set. The result for ERA-Interim, ~9 ARs on average at any given time (Figure 9, black bar outline repeated in each panel), is 2 ARs fewer than the result in *Guan and Waliser* [2015], which is attributable to the coarser horizontal resolution used here for AR detection (i.e., 2.5° versus 1.5°). This number agrees with the other two reanalyses (Figure 9, gray fill inside the black outlines). The quantity is reasonably represented in most models, with a few exceptions where the models simulate considerably fewer ARs. In the most extreme case (EC-GEM), the number of simulated ARs is only about one half of the reference. This is likely a result of the strong northward IVT bias in that model (Figure 7, shaded panel), where the bias leads to the lack of poleward IVT in the Southern Hemisphere (unfavorable for ARs) and overly large IVT in the Northern Hemisphere (meaning a broad region can persistently meet the observationally derived AR IVT threshold and therefore may not satisfy the AR geometry criteria). Biases in individual AR width, together with biases in the number of individual ARs, contribute to biases in AR frequency (Figure 3), AR fractional total meridional IVT (Figure 7), and AR fractional zonal circumference (Figure 8) based on equations (6)–(8). As suggested in *Guan and Waliser* [2015], AR width can also determine the time it takes for an AR to propagate across a given location, therefore contributing to the severity of the event at that location [*Ralph et al.*, 2013]. The result here suggests the importance to consider ARs as individual "objects" when evaluating model performance in addition to other statistics less specific to individual ARs as considered in previous sections.

The two components of AR IVT each have a positively skewed distribution, although AR meridional IVT has a smaller mean and a longer tail (not shown). In that regard, AR IVT is typically stronger in the zonal direction [*Guan and Waliser*, 2015], although in terms of fractional contribution, AR meridional IVT dominates the total poleward IVT in the midlatitudes [*Zhu and Newell*, 1998]. The overall pattern of the joint distribution is reasonably represented in most of the model simulations (Figure 10, shading). The four coarsest resolution atmosphere-only models are all marked by notable positive (negative) biases in AR zonal (meridional) IVT, as are a few models with higher resolutions and one coupled model (SPCCSM3) (Figure 10, contours). Note that for AR meridional IVT, a positive (negative) bias here indicates poleward (equatorward) bias. Further examination indicates that biases in AR IVT are mainly related to biases in its direction, not magnitude (not shown, but see Figure 15 for RMSE in the direction versus magnitude of AR IVT, which includes the contribution from bias). In general, normalized biases for AR zonal and meridional IVT are much weaker than those for AR length and width (comparing Figures 9 and 10), suggesting a more challenging task of simulating AR geometry than intensity in terms of IVT.

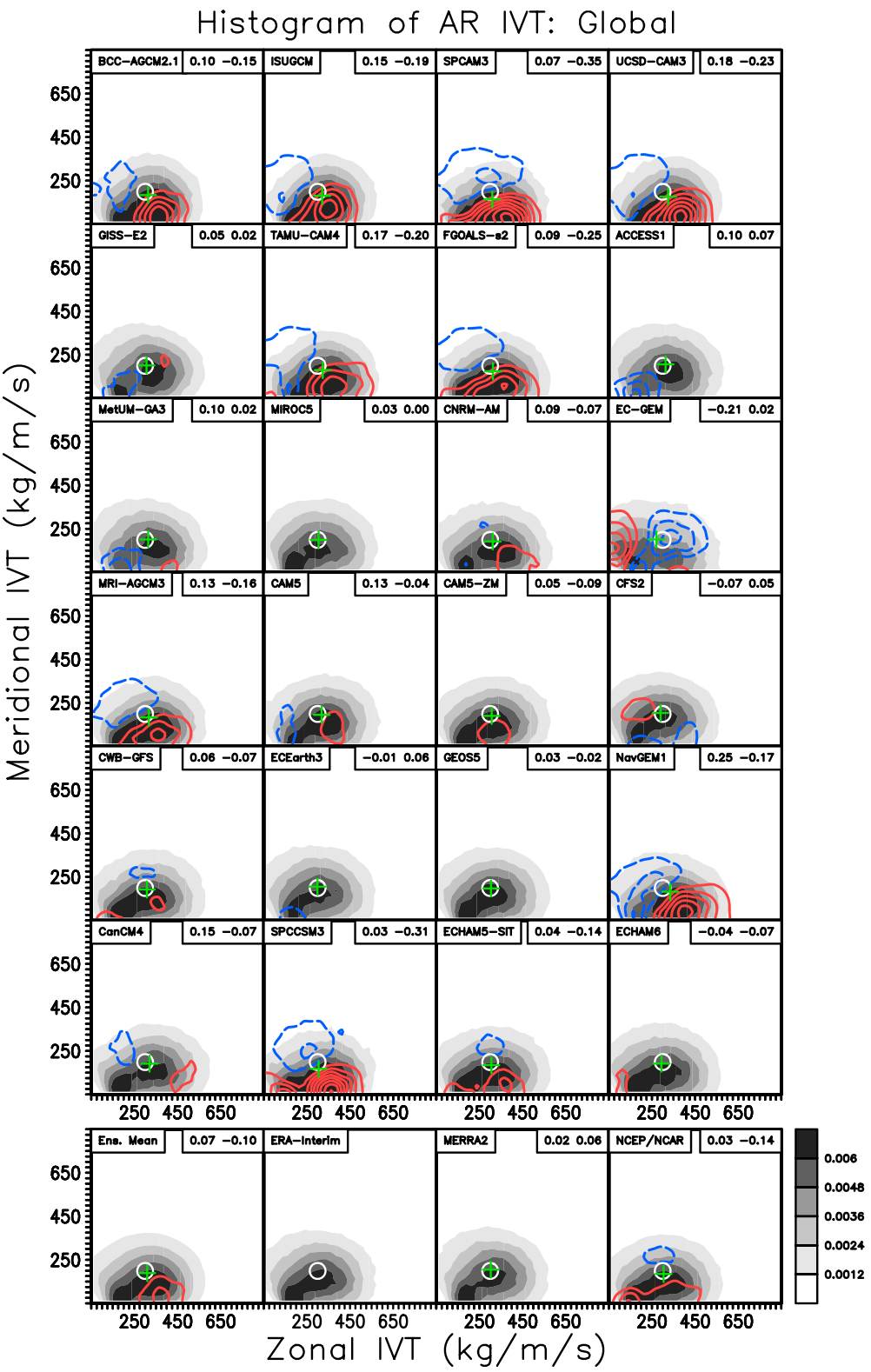

Figure 10. As in Figure 9 but for AR zonal and meridional IVT. Normalized biases in AR zonal and meridional IVT are indicated by the numbers in the upper right corner of each panel. Note for AR meridional IVT a positive (negative) bias here indicates the bias is directed poleward (equatorward).

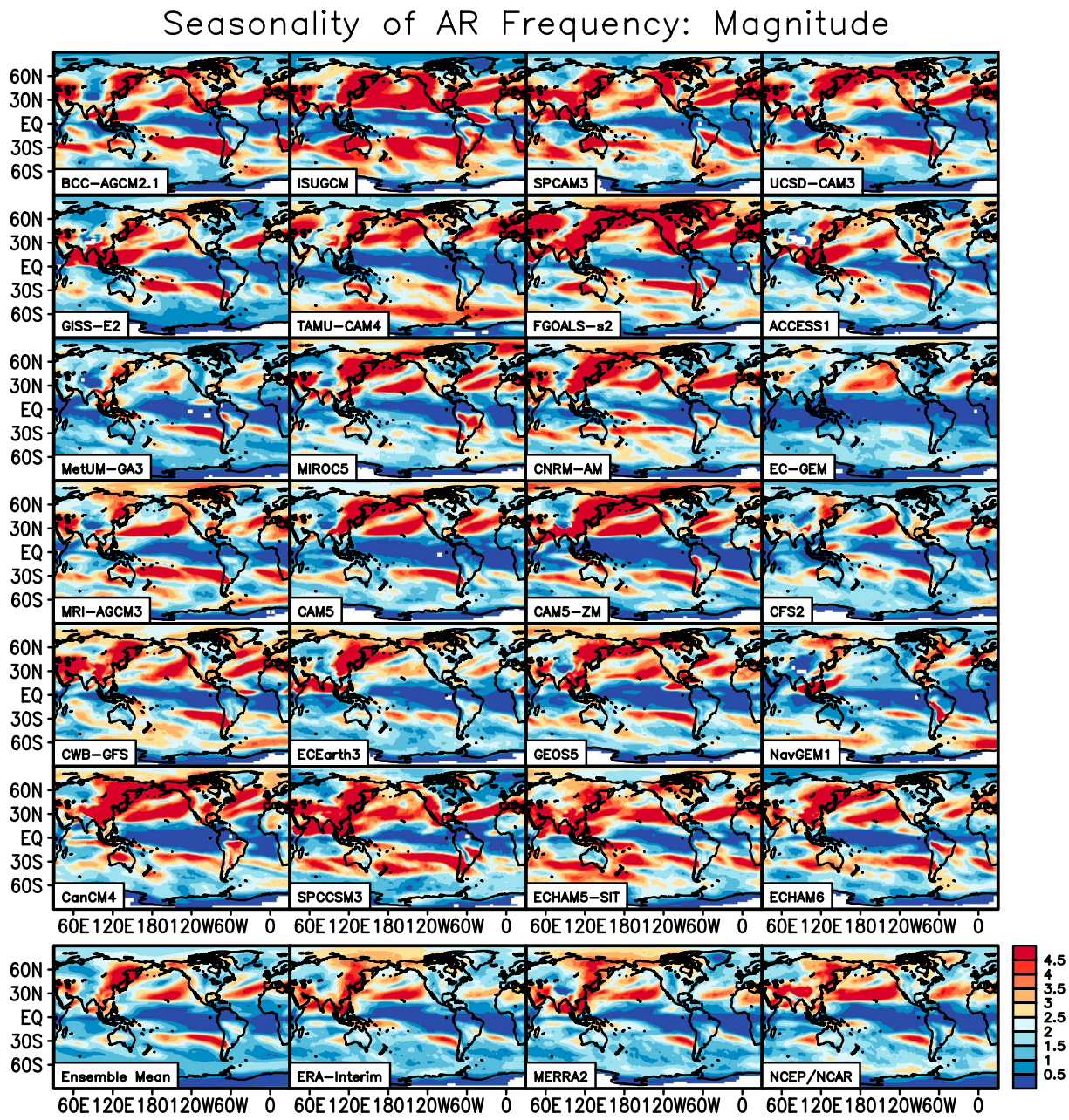

Figure 11. As shading in Figure 2 but for the magnitude (i.e., standard deviation) of the monthly resolved climatological annual cycle of AR frequency (percent of time steps). For example, a shaded value of 5 here at a location where the shaded value is 10 in Figure 2 would indicate that the magnitude of the annual cycle is half the long-term annual mean. The annual cycle is defined here as the summation of the first three Fourier harmonics without the time mean (the time mean was evaluated in Figure 3).

5. Seasonality of AR Occurrences

Seasonality of AR occurrences is examined by the annual cycle of AR frequency, defined here as the summation of the first three Fourier harmonics. Monthly mean AR frequencies are first calculated, based on which a monthly resolved climatological annual cycle is formed. Note that the time mean AR frequency was already evaluated in section 3 and therefore excluded from the calculation of the annual cycle here. Evaluation is first based on the magnitude and phase of the annual cycle separately, then the full spatiotemporal distribution.

Overall, the annual cycle is stronger (in terms of standard deviation over the 12 months) in the Northern Hemisphere, as shown by the three reanalyses (Figure 11). The annual cycle is especially weak (i.e., the

number of ARs tend to be similar in all seasons) across the southern midlatitudes compared to corresponding regions in the Northern Hemisphere. The annual cycle has the largest magnitude in coastal East/South Asia and around subtropical eastern Pacific/Atlantic, consistent with strong seasonal contrasts in AR frequency between November–March and May–September in these regions [*Guan and Waliser*, 2015]. In these regions, the annual cycle can modify the annual mean AR frequency by one third or more (comparing Figures 2 and 11). The overall hemispheric contrast, i.e., stronger annual cycle in the Northern Hemisphere, tends to be represented in the majority of the simulations. The simulated annual cycle tends to be positively biased in magnitude in the extratropics in most cases, which is most notable with the coarsest resolution atmosphere-only models and the coupled models (Figure 12). Averaged globally, only four models have negative biases (stronger in MetUM-GA3 and EC-GEM and weaker in Climate Forecast System 2 (CFS2) and NavGEM1). The bias in the ensemble mean is much weaker than that in individual models but still beyond reanalysis uncertainties in many locations. In some locations, the bias in the ensemble mean is weaker even though the individual models consistently overestimate the magnitude of the annual cycle. This can only be possible when the annual cycle is not synced in time between individual models, which is examined below.

To demonstrate timing errors in simulated annual cycles, the peak month of AR frequency is examined for three selected regions, namely, southern/central California, Britain, and western Greenland, where AR frequencies peak in three different months in two seasons (Figure 13, dots). In southern/central California, AR frequency has a well-defined peak in January, which is consistently the case for the three reanalyses. The model ensemble mean has a smoother peak, which lags the reanalyses by 1 month. Peaks in individual models can be up to 2 months too early or 3 months too late, with only five models simulating the exact peak month. Errors in AR peak month were weaker in *Payne and Magnusdottir* [2015] for the 28 CMIP5 models, likely due to averaging over the entire northeastern Pacific, hence compensating errors over the relatively large analysis domain. The general difficulty in realistically simulating the phase of the annual cycle is also seen in the other two regions, with only two (for Britain) or three (for western Greenland) models simulating the exact peak month. Magnitude errors are also notable for many individual models (Figure 13, bars; see also Figure 12). The magnitude and timing errors in the annual cycle do not appear to be correlated.

Noting the importance of both magnitude and timing errors as illustrated in the above three examples, the full spatiotemporal distribution of the annual cycle (i.e., 12 month climatologies over all grid cells globally) is now evaluated using a Taylor diagram (Figure 14). The magnitude and timing errors discussed above are reflected, respectively, in the standard deviation ratios and correlations. The largest errors are seen with the coarsest resolution models. One coupled model (SPCCSM3) shows the lowest correlation and the largest normalized centered RMSE (i.e., the distance to the reference point). Errors in all (most) individual models are beyond the primary (secondary) measure of reanalysis uncertainty. The ensemble mean performs notably worse than the difference between MERRA2 and ERA-Interim in terms of correlation, although better than the difference between NCEP/NCAR and ERA-Interim in terms of standard deviation ratio.

6. Summary Error Characteristics

The overall performance of the 24 models in terms of 17 metrics is summarized in a portrait diagram (Figure 15), with a few enhancements from the conventional version devised in *Gleckler et al.* [2008]. Ten of these metrics are the exact ones examined in sections 3–5. The other seven metrics, namely, the zonal means of AR frequency, zonal, and meridional IVT, and 1-D histograms of length, width, zonal, and meridional IVT of individual ARs, provide closely related information that complement the other results. The normalized RMSE ($RMSE_{norm}$ in equation (3)) for each model and each metric is first calculated, based on which the relative error (E^* in equation (4)) of each model for a given metric compared to the median error across all models for that metric is constructed and shown in the main section of the diagram. For example, E^* for AR frequency (row A, columns 1–24) is based on scaling the $RMSE_{norm}$ of each model (not shown) by the median $RMSE_{norm}$ (row A, column 25) using equation (4). These relative errors provide a convenient way to visually compare the different models in terms of their performances, including the degree of consistencies/discrepancies between model performances, against observations. Warm (cold) colors indicate cases with positive (negative) values of E^*, i.e., where model errors are larger (smaller) than the "typical" (i.e., median) model error

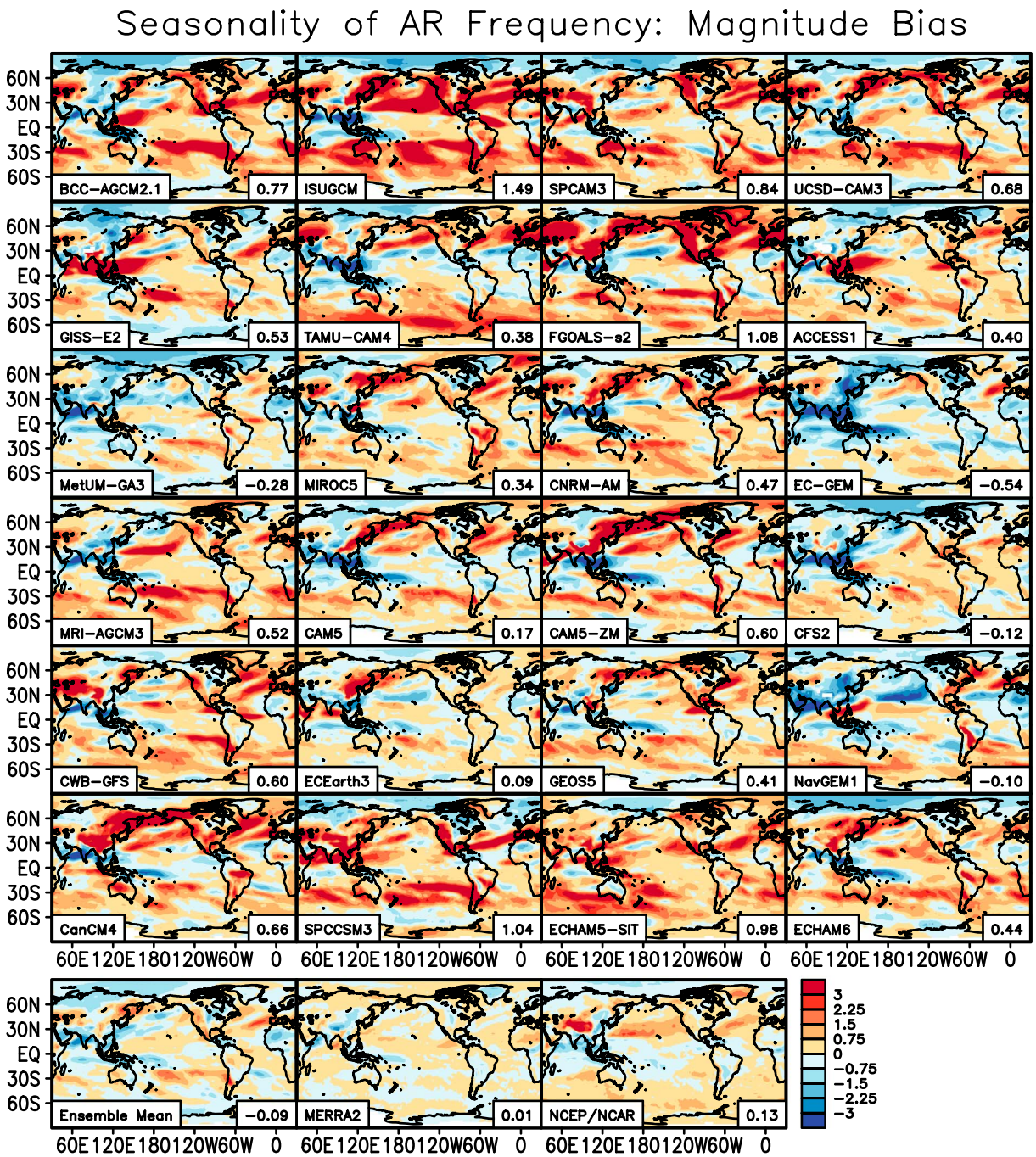

Figure 12. As Figure 3 but for biases in the magnitude (i.e., standard deviation) of the monthly resolved climatological annual cycle of AR frequency (percent of time steps).

for a given metric, and the darkness/lightness of the colors indicates to what extent the model errors deviate from the median error.

For such a construction, the typical/median model, by definition, has a zero relative error. In that regard, the relative error by itself no longer reflects the absolute distance between a model and the observation. Either consistently good or consistently bad performances across the models for a given metric will lead to light colors across the corresponding row in the diagram, and well separated model performances will lead to more layers of colors—in any of these cases the absolute distance between a model and the observation is not

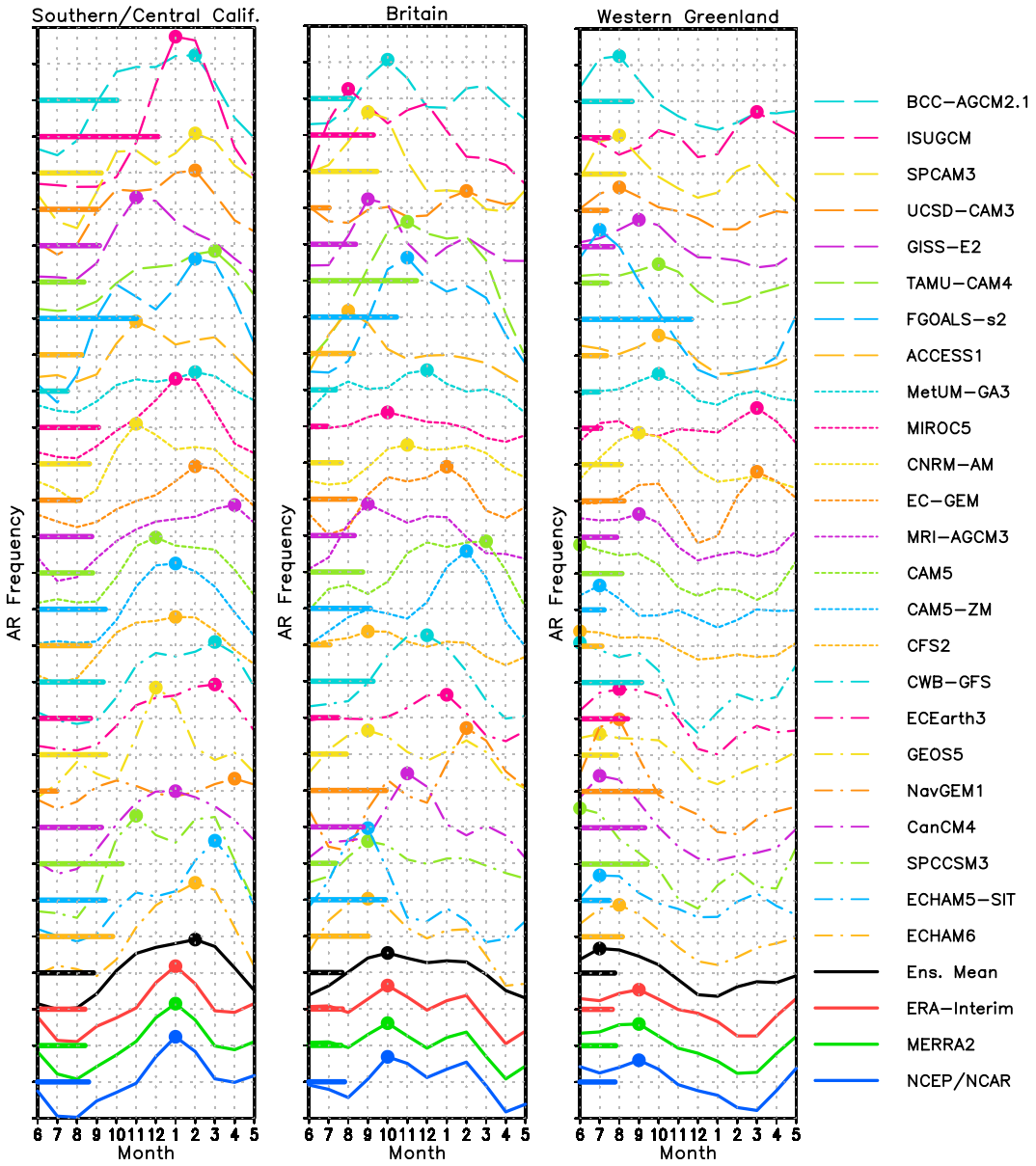

Figure 13. Climatological annual cycle of AR frequency (percent of time steps) in simulations and reanalyses for three selected regions (colored curves). For each colored curve, the standard deviation (horizontal bar) and the peak month (dot) are indicated as measures of the magnitude and phase of the annual cycle. Each horizontal bar is drawn at the zero value of the correspondingly colored curve.

reflected by the colors themselves. To provide information on absolute errors, gray circles are placed on top of the color shading to signify cases where the RMSE of a simulated field is large compared to the standard deviation of the reference field (i.e., when the former normalized by the latter, $RMSE_{norm}$, exceeds 0.25) and meanwhile large compared to the primary measure of reanalysis uncertainty (i.e., $RMSE_{norm}$ between model and ERA-Interim exceeds $RMSE_{norm}$ between MERRA2 and ERA-Interim; the latter is shown in column 27). To additionally show information on bias, an arrow is placed inside a gray circle to signify the sign of the bias (positive/negative if pointing upward/downward), but only if the magnitude of the bias itself is large compared to the standard deviation of the reference field (i.e., when the former normalized by the latter, B_{norm}, exceeds 0.05 in terms of magnitude) and meanwhile large compared to the primary measure of reanalysis uncertainty (i.e., B_{norm} between model and ERA-Interim exceeds B_{norm} between MERRA2 and

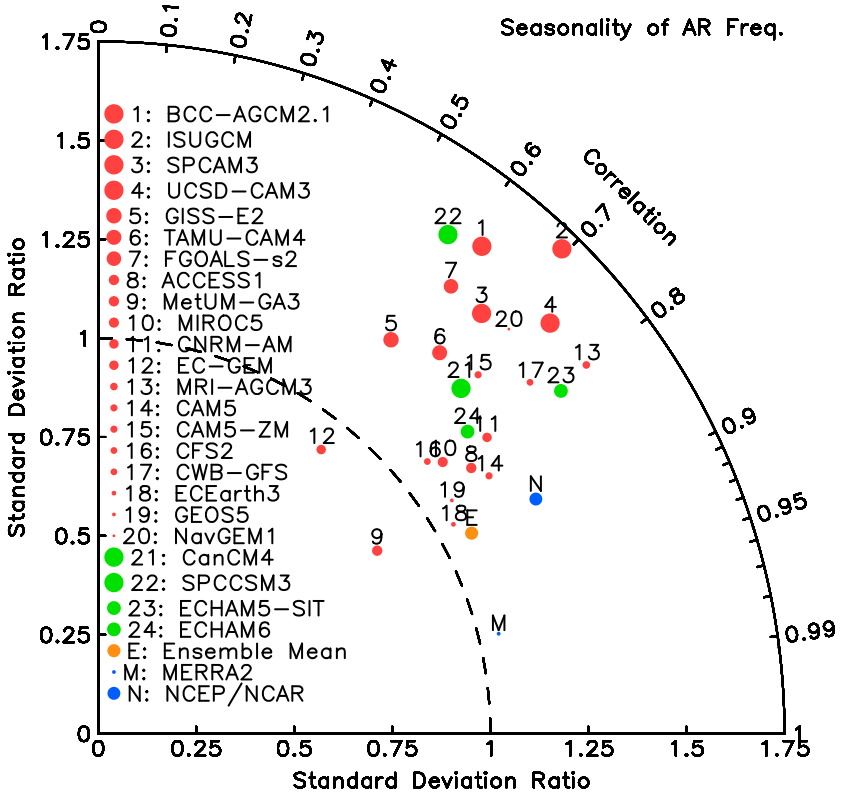

Figure 14. As in Figure 6 but for evaluating the climatological annual cycle of AR frequency.

ERA-Interim in terms of magnitude). As an example, the color/circle/arrow markings described above for AR frequency simulated by GEOS5 (row A, column 19) indicate above-average performance (blue shading) compared to the median model, with notable RMSE (gray circle) but negligible overall bias (no arrow) compared to the reanalyses. Model biases were examined in the previous sections and will not be repeated in the discussion below. Above the main section of the diagram is an isolated row that gives the median E^* across all metrics for each model and a row that shows the native horizontal resolutions of the models. The coupled models are to the right, encased by a green outline. On the far right are four isolated columns that give, respectively, the median $RMSE_{norm}$ across the models, the $RMSE_{norm}$ for the multimodel ensemble mean, for MERRA2, and for NCEP/NCAR, all using ERA-Interim as the reference. Note the distinction between the error of the multimodel ensemble mean (column 26) and the error associated with the typical individual model (column 25).

The absolute errors of individual models are notably large in a number of cases (gray circles in Figure 15). In particular, absolute errors are consistently large among all models in the cases of AR frequency, zonal IVT, fractional total meridional IVT, fractional zonal circumference, and the three seasonality metrics. For AR frequency and zonal IVT, even the zonal means show large absolute errors in the majority of the models. Absolute errors are also large in the majority of the models for AR width. We note again that the darkness/lightness of the colors in Figure 15 does not indicate the magnitude of the absolute errors, and it is possible for a row with largely light colors (small intermodel discrepancies) to be dominated by gray circles (large deviations from observation)—the case for AR frequency, zonal IVT, and seasonality. Overall, model performance is better for AR meridional IVT than zonal IVT and AR frequency, better for the histogram of AR length than width, and better for the histogram of AR IVT magnitude than direction (based on comparing the number of models showing gray circles). The best-simulated metrics are the zonal mean of AR meridional IVT and the histogram of AR IVT magnitude, for which absolute errors are consistently small among the models (lack of gray circles). In 15 of the 17 metrics considered, the error of the typical individual model is larger than the error of the multimodel ensemble mean (comparing shading in columns 25 and 26). For 10 metrics,

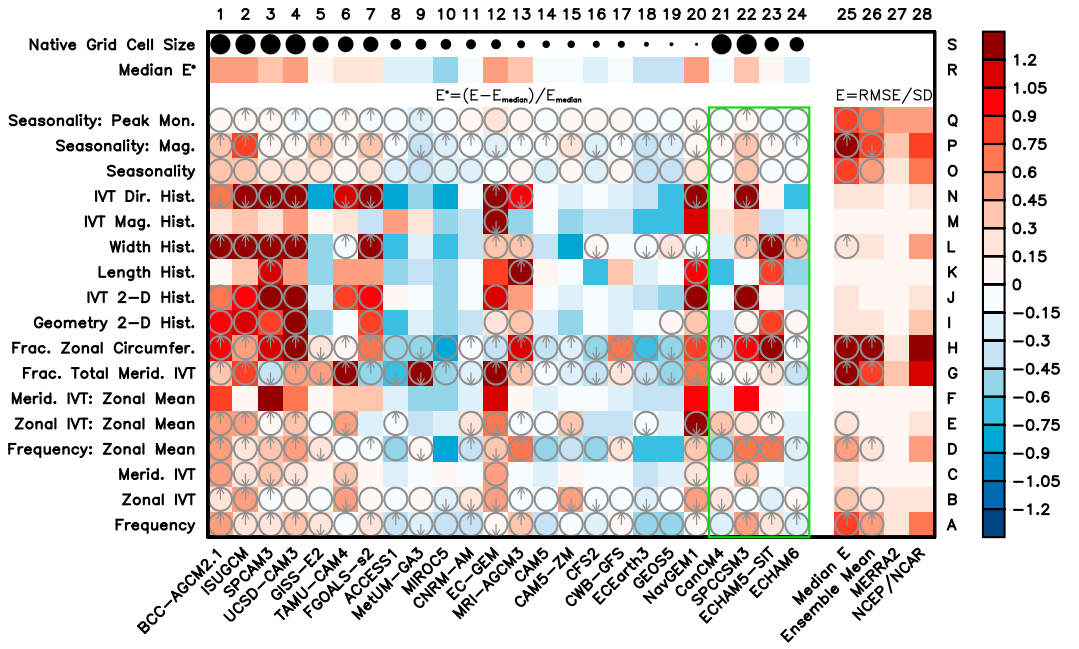

Figure 15. Portrait diagram [*Gleckler et al.*, 2008] showing the relative error E^* of each model for 17 metrics considered in this study (color shading), with a few enhancements described below. E^* is defined as $(E - E_{\text{median}})/E_{\text{median}}$, where E is the normalized RMSE of an individual model relative to ERA-Interim for a given metric (not shown) and E_{median} is the median E across all models for that metric. E_{median} and E for the ensemble mean of the models, MERRA2, and NCEP/NCAR relative to ERA-Interim are shown in the four isolated columns on the far right. For each model, the median E^* across all metrics is shown in the isolated row near the top. The gray circles indicate cases where the E of a given model and metric is greater than 0.25 and meanwhile exceeds the primary measure of reanalysis uncertainty (MERRA2 relative to ERA-Interim). Wherever a gray circle is shown, an arrow additionally shows the sign of the bias (positive/negative if pointing upward/downward), but only if the magnitude of the normalized bias itself is greater than 0.05 and meanwhile exceeds the primary measure of reanalysis uncertainty. Biases are not defined for AR geometry and IVT 2-D histograms and are by definition zero for the first metric of AR seasonality; no arrows are drawn for these three metrics. The top row shows the approximate size of native model grid cells for reference, based on which the models are sorted. The four coupled models are placed together to the right and encased by a green outline. For convenience, rows (columns) are labeled on the right (top) by capital letters (numbers).

the absolute error of the typical model is notably large (gray circles in column 25). This is slightly improved for the ensemble mean, where eight metrics remain to show notably large errors (gray circles in column 26). The result suggests the dependence of model performance on the specific metric of interest and consistently large errors across individual models in simulating some of the basic characteristics of ARs.

As discussed earlier, there is a tendency for the coarsest resolution models to exhibit the largest errors (dark red in Figure 15), although the relationship is not monotonic. The correlation between the median relative errors and the sizes of native grid cells (rows R and S) is 0.55 ($p < 0.01$). For the atmosphere-only models, there appears to be a transition around the median resolution of ~1.5° (between Flexible Global Ocean-Atmosphere-Land System Model, Spectral Version 2 (FGOALS-s2) and Australian Community Climate and Earth-System Simulator 1 (ACCESS1)) where the relative errors go from largely positive to largely negative. Two of the four coupled models exhibit largely positive relative errors, while largely negative for the other two. The two models with superparameterization, one atmosphere-only and one with ocean-atmosphere coupling, are among the worst performing models, which could be partly related to the coarse resolution (T42, 2.8°) of the parent models and the difference between resolved convection and convective parameterization. As mentioned in the introduction, coupled models are subject to biases in simulated sea surface temperatures and sea ice concentrations, which can contribute to biases in IVT and in turn ARs [*Lavers et al.*, 2013] (see also Figures S1 and S2). Among the better performing models, the most noteworthy is ECEarth3 (high resolution), which never performed below average (no red shading in any single metric). Immediately following are Model for Interdisciplinary Research on Climate 5 (MIROC5) and Community Atmosphere Model 5

(CAM5) (medium resolutions), each of which performed below average only once. The result suggests possibly important contribution of model horizontal resolution to AR simulation qualities, with no clear indication for the contribution by ocean-atmosphere coupling or superparameterization due to limited number of samples examined for these cases and possible dominance of the resolution impact in these coarse-resolution cases. The transition in overall model performance around the median resolution of ~1.5° suggests that a resolution on this order is likely a necessary, although not sufficient, condition for a model to outperform the typical model represented in the current ensemble.

7. Conclusions

Using a recently developed global AR detection algorithm, the current study examined 17 metrics of key AR characteristics in 20 year simulations by 24 state-of-the-art weather/climate models. Among the 17 metrics, 8 are related to spatial distributions of climatological AR frequency and IVT, 6 related to histograms of AR geometries and intensities, and 3 related to seasonality of AR occurrences. Among the 24 models are four with ocean-atmosphere coupling and two with superparameterization. Quantitative evaluation based on an array of error statistics in the context of reanalysis uncertainties suggests the following:

1. *Spatial distributions of AR frequency and IVT*. About half of the models show notable positive biases in AR frequency in the extratropical Pacific/Atlantic and extending to the west coasts of North/South America and Europe, and only a few models exhibit notable negative biases (Figure 3). Biases in AR zonal IVT are, in general, stronger than those in meridional IVT (Figures 4 and 5; note different scales between the two figures). The coarsest resolution atmosphere-only models as well as the coupled models tend to exhibit the largest pattern errors (Figure 6) and overall biases (Figures 3–5) in AR frequency and IVT. Biases in both AR fractional total meridional IVT (Figure 7) and AR fractional zonal circumference (Figure 8) are attributable to biases in AR frequency (Figure 3). Better performance of the models in AR IVT relative to AR frequency, both in terms of overall biases and pattern similarities, suggests that it is more challenging to simulate AR occurrence than AR intensity in terms of IVT.
2. *Probability distributions of AR geometry and intensity*. The joint AR length-width distribution is reasonably represented in the model simulations in terms of the overall pattern but biased toward larger AR width for the four coarsest resolution atmosphere-only models and the coupled models and toward smaller AR width for the three highest resolution models (Figure 9). The biases in the latter three models are much weaker than in the coarsest resolution models and could be an artifact of the reference data (ERA-Interim) being on a coarser native grid than the three highest resolution models. In 18 of the 24 models, AR width has larger normalized biases than AR length. The overall pattern of the joint distribution of AR zonal versus meridional IVT is reasonably represented in most of the model simulations, but with notable eastward and/or equatorward biases in the four coarsest resolution atmosphere-only models, as well as a few higher resolution models and one coupled model (Figure 10). The biases in AR IVT are mainly related to biases in its direction, not magnitude (Figure 15). In general, normalized biases for AR zonal and meridional IVT are much weaker than for AR length and width (comparing Figures 9 and 10), suggesting a more challenging task of simulating AR geometry than AR intensity in terms of IVT.
3. *Seasonality of AR occurrences*. Simulated AR annual cycle tends to be stronger than observed in the extratropics in the majority of the models, which is most notable with the coarsest resolution atmosphere-only models and the coupled models (Figure 12). The peak month, as a measure of the phase of the annual cycle, is accurately simulated only in 5/2/3 individual models based on three example regions where ARs peak in three different months in two seasons (southern/central California, Britain, and western Greenland) (Figure 13). Evaluation of the full spatiotemporal distribution of the annual cycle over the globe suggests the largest errors to be associated with the coarsest resolution models (Figure 14) and that it is even more challenging to simulate the seasonal variations in AR frequency than the annual mean (Figure 15).
4. *Summary error characteristics*. Normalized total RMSE, which consolidates the overall biases and pattern errors, is consistently large among all models in the cases of AR frequency, zonal IVT, fractional total meridional IVT, fractional zonal circumference, and seasonality and is large in the majority of the models for AR width (Figure 15). Overall, model performance is better for AR meridional IVT than zonal IVT and AR frequency, better for the histogram of AR length than width, and better for the histogram of AR IVT magnitude than direction. Model-wise, there is a tendency for the coarsest resolution models to exhibit the largest errors, although the relationship is not monotonic. For the atmosphere-only models, there

appears to be a transition around the median resolution of ~1.5° where the model performance goes from largely above average to largely below average. Two of the four coupled models exhibit largely above-average performance, while largely below-average for the other two. The two models with super-parameterization, one atmosphere-only and one with ocean-atmosphere coupling, are among the worst performing models, which could be partly related to the coarse resolution (T42, 2.8°) of the parent models, the difference between conventional and super parameterization, and, in the case of the coupled model, possible biases in sea surface temperatures and sea ice concentrations.

The results presented herein suggest the dependence of model performance on the specific metric of interest and consistently large errors across individual models in simulating some of the basic characteristics of ARs. Moreover, the results suggest possibly important contribution of model horizontal resolution to AR simulation qualities and that a resolution of ~1.5° is likely a necessary, although not sufficient, condition for a model to outperform the typical model represented in the ensemble of models examined. The results suggest that the coupled models and/or superparameterized models have no contribution over conventional, atmosphere-only models for validation purposes, likely due to limited number of samples examined for these cases and possible dominance of the resolution impact in these coarse-resolution cases.

These findings indicate considerable challenges for the state-of-the-art weather/climate models in simulating the fundamental characteristics of ARs. Difficulties of the models in simulating the fractional total meridional IVT accounted for by ARs have implications to the fidelity of the global water and energy cycles represented in models, which may affect the ability of these models in projecting future changes related to these basic processes in the Earth system. Notable errors in simulated AR seasonality, in terms of both magnitude and timing, may affect the accuracy of subseasonal to seasonal forecasts of AR activities of relevance to water resource and flood/drought management. Further investigations are needed to understand the sources of these systematic errors across the models. It is expected that the current work will contribute to the development of a suite of AR simulation diagnostics and model performance metrics that will facilitate AR-oriented model evaluation and development efforts.

7.1. Code Availability

The code for the AR detection algorithm [*Guan and Waliser*, 2015] in the form of a MATLAB function is available from the corresponding author upon request.

Acknowledgments

The GASS-YoTC model archive is freely available at https://www.earthsystem-cog.org/projects/gass-yotc-mip/ and was supported by NSF AGS-1221013. The ERA-Interim, MERRA2, and NCEP/NCAR reanalyses are freely available, respectively, at http://apps.ecmwf.int/datasets/data/interim-full-daily/, https://gmao.gsfc.nasa.gov/reanalysis/MERRA-2/, and https://www.esrl.noaa.gov/psd/data/gridded/data.ncep.reanalysis.html. This research was supported by the NASA Energy and Water cycle Study (NEWS) program and the California Department of Water Resources. Encouraging comment by P. Gleckler at Lawrence Livermore National Laboratory is appreciated. DEW's contribution to this study was carried out on behalf of the Jet Propulsion Laboratory, California Institute of Technology, under a contract with the National Aeronautics and Space Administration.

References

Bosilovich, M. G., et al. (2015), MERRA-2: Initial evaluation of the climate, technical report series on global modeling and data assimilation, Volume 43, NASA/TM–2015-104606/Vol. 43, Greenbelt, Md, 139 pp. [Available online at https://gmao.gsfc.nasa.gov/pubs/docs/Bosilovich803.pdf.]

Brands, S., J. M. Gutiérrez, and D. San-Martín (2017), Twentieth-century atmospheric river activity along the west coasts of Europe and North America: Algorithm formulation, reanalysis uncertainty and links to atmospheric circulation patterns, *Clim. Dyn.*, *48*, 2771–2795, doi:10.1007/s00382-016-3095-6.

Dee, D. P., et al. (2011), The ERA-Interim reanalysis: Configuration and performance of the data assimilation system, *Q. J. R. Meteorol. Soc.*, *137*, 553–597, doi:10.1002/qj.828.

Deser, C., R. A. Tomas, and L. Sun (2015), The role of ocean-atmosphere coupling in the zonal mean atmospheric response to Arctic sea ice loss, *J. Clim.*, *28*, 2168–2186, doi:10.1175/JCLI-D-14-00325.1.

Dettinger, M. (2011), Climate change, atmospheric rivers, and floods in California—A multimodel analysis of storm frequency and magnitude changes, *J. Am. Water Resour. Assoc.*, *47*, 514–523, doi:10.1111/j.1752-1688.2011.00546.x.

Gao, Y., J. Lu, and L. R. Leung (2016), Uncertainties in projecting future changes in atmospheric rivers and their impacts on heavy precipitation over Europe, *J. Clim.*, *29*, 6711–6726, doi:10.1175/JCLI-D-16-0088.1.

Gao, Y., J. Lu, L. R. Leung, Q. Yang, S. Hagos, and Y. Qian (2015), Dynamical and thermodynamical modulations on future changes of landfalling atmospheric rivers over western North America, *Geophys. Res. Lett.*, *42*, 7179–7186, doi:10.1002/2015GL065435.

Gimeno, L., R. Nieto, M. Vázquez, and D. A. Lavers (2014), Atmospheric rivers: A mini-review, *Front. Earth Sci.*, *2*, 2.1–2.6, doi:10.3389/feart.2014.00002.

Gimeno, L., F. Dominguez, R. Nieto, R. Trigo, A. Drumond, C. J. C. Reason, A. S. Taschetto, A. M. Ramos, R. Kumar, and J. Marengo (2016), Major mechanisms of atmospheric moisture transport and their role in extreme precipitation events, *Annu. Rev. Env. Resour.*, *41*, 117–141, doi:10.1146/annurev-environ-110615-085558.

Gleckler, P. J., K. E. Taylor, and C. Doutriaux (2008), Performance metrics for climate models, *J. Geophys. Res.*, *113*, D06104, doi:10.1029/2007JD008972.

Gorodetskaya, I. V., M. Tsukernik, K. Claes, M. F. Ralph, W. D. Neff, and N. P. M. Van Lipzig (2014), The role of atmospheric rivers in anomalous snow accumulation in East Antarctica, *Geophys. Res. Lett.*, *41*, 6199–6206, doi:10.1002/2014GL060881.

Guan, B., and D. E. Waliser (2015), Detection of atmospheric rivers: Evaluation and application of an algorithm for global studies, *J. Geophys. Res. Atmos.*, *120*, 12,514–12,535, doi:10.1002/2015JD024257.

Hagos, S., L. R. Leung, Q. Yang, C. Zhao, and J. Lu (2015), Resolution and dynamical core dependence of atmospheric river frequency in global model simulations, *J. Clim.*, *28*, 2764–2776, doi:10.1175/JCLI-D-14-00567.1.

Hagos, S. M., L. R. Leung, J.-H. Yoon, J. Lu, and Y. Gao (2016), A projection of changes in landfalling atmospheric river frequency and extreme precipitation over western North America from the Large Ensemble CESM simulations, *Geophys. Res. Lett.*, *43*, 1357–1363, doi:10.1002/2015GL067392.

Hirota, N., Y. N. Takayabu, M. Kato, and S. Arakane (2016), Roles of an atmospheric river and a cutoff low in the extreme precipitation event in Hiroshima on 19 August 2014, *Mon. Weather Rev.*, *144*, 1145–1160, doi:10.1175/MWR-D-15-0299.1.

Jackson, D. L., M. Hughes, and G. A. Wick (2016), Evaluation of landfalling atmospheric rivers along the U.S. west coast in reanalysis data sets, *J. Geophys. Res. Atmos.*, *121*, 2705–2718, doi:10.1002/2015JD024412.

Jiang, X., et al. (2015), Vertical structure and physical processes of the Madden-Julian Oscillation: Exploring key model physics in climate simulations, *J. Geophys. Res. Atmos.*, *120*, 4718–4748, doi:10.1002/2014JD022375.

Kalnay, E., et al. (1996), The NMC/NCAR 40-year reanalysis project, *Bull. Am. Meteorol. Soc.*, *77*, 437–471, doi:10.1175/1520-0477(1996)077<0437:TNYRP>2.0.CO;2.

Klingaman, N. P., et al. (2015), Vertical structure and physical processes of the Madden-Julian Oscillation: Linking hindcast fidelity to simulated diabatic heating and moistening, *J. Geophys. Res. Atmos.*, *120*, 4690–4717, doi:10.1002/2014JD022374.

Knippertz, P., and J. E. Martin (2005), Tropical plumes and extreme precipitation in subtropical and tropical West Africa, *Q. J. R. Meteorol. Soc.*, *131*, 2337–2365, doi:10.1256/qj.04.148.

Lavers, D. A., R. P. Allan, G. Villarini, B. Lloyd-Hughes, D. J. Brayshaw, and A. J. Wade (2013), Future changes in atmospheric rivers and their implications for winter flooding in Britain, *Environ. Res. Lett.*, *8*, 034010, doi:10.1088/1748-9326/8/3/034010.

Mahoney, K., D. L. Jackson, P. Neiman, M. Hughes, L. Darby, G. Wick, A. White, E. Sukovich, and R. Cifelli (2016), Understanding the role of atmospheric rivers in heavy precipitation in the Southeast United States, *Mon. Weather Rev.*, *144*, 1617–1632, doi:10.1175/MWR-D-15-0279.1.

Neff, W., G. P. Compo, F. M. Ralph, and M. D. Shupe (2014), Continental heat anomalies and the extreme melting of the Greenland ice surface in 2012 and 1889, *J. Geophys. Res. Atmos.*, *119*, 6520–6536, doi:10.1002/2014JD021470.

Payne, A. E., and G. Magnusdottir (2015), An evaluation of atmospheric rivers over the North Pacific in CMIP5 and their response to warming under RCP 8.5, *J. Geophys. Res. Atmos.*, *120*, 11,173–11,190, doi:10.1002/2015JD023586.

Petch, J., D. Waliser, X. Jiang, P. Xavier, and S. Woolnough (2011), A global model inter-comparison of the physical processes associated with the MJO, GEWEX News, August.

Radić, V., A. J. Cannon, B. Menounos, and N. Gi (2015), Future changes in autumn atmospheric river events in British Columbia, Canada, as projected by CMIP5 global climate models, *J. Geophys. Res. Atmos.*, *120*, 9279–9302, doi:10.1002/2015JD023279.

Ralph, F. M., G. A. Wick, P. J. Neiman, B. J. Moore, J. R. Spackman, M. Hughes, F. Yong, and T. Hock (2012), Atmospheric rivers in reanalysis products: A six-event comparison with aircraft observations of water vapor transport, WCRP Reanalysis Conf., 1 pp., Silver Spring, Md. [Available online at https://www.wcrp-climate.org/ICR4/posters/Hughes_AT-20.pdf.]

Ralph, F. M., T. Coleman, P. J. Neiman, R. J. Zamora, and M. D. Dettinger (2013), Observed impacts of duration and seasonality of atmospheric-river landfalls on soil moisture and runoff in coastal northern California, *J. Hydrometeorol.*, *14*, 443–459, doi:10.1175/JHM-D-12-076.1.

Ramos, A. M., R. Tomé, R. M. Trigo, M. L. R. Liberato, and J. G. Pinto (2016), Projected changes in atmospheric rivers affecting Europe in CMIP5 models, *Geophys. Res. Lett.*, *43*, 9315–9323, doi:10.1002/2016GL070634.

Rutz, J. J., W. J. Steenburgh, and F. M. Ralph (2014), Climatological characteristics of atmospheric rivers and their inland penetration over the western United States, *Mon. Weather Rev.*, *142*, 905–921, doi:10.1175/MWR-D-13-00168.1.

Shields, C. A., and J. T. Kiehl (2016a), Atmospheric river landfall-latitude changes in future climate simulations, *Geophys. Res. Lett.*, *43*, 8775–8782, doi:10.1002/2016GL070470.

Shields, C. A., and J. T. Kiehl (2016b), Simulating the pineapple express in the half degree Community Climate System Model, CCSM4, *Geophys. Res. Lett.*, *43*, 7767–7773, doi:10.1002/2016GL069476.

Taylor, K. E. (2001), Summarizing multiple aspects of model performance in a single diagram, *J. Geophys. Res.*, *106*, 7183–7192, doi:10.1029/2000JD900719.

Waliser, D. E., and B. Guan (2017), Extreme winds and precipitation during landfall of atmospheric rivers, *Nat. Geosci.*, *10*, 179–183, doi:10.1038/ngeo2894.

Warner, M. D., C. F. Mass, and E. P. Salathé Jr. (2015), Changes in winter atmospheric rivers along the North American west coast in CMIP5 climate models, *J. Hydrometeorol.*, *16*, 118–128, doi:10.1175/JHM-D-14-0080.1.

Zhu, Y., and R. E. Newell (1998), A proposed algorithm for moisture fluxes from atmospheric rivers, *Mon. Weather Rev.*, *126*, 725–735, doi:10.1175/1520-0493(1998)126<0725:APAFMF>2.0.CO;2.