BAMS Article

Evaluating Climate Models with the CLIVAR 2020 ENSO Metrics Package

Yann Y. Planton, Eric Guilyardi, Andrew T. Wittenberg, Jiwoo Lee,
Peter J. Gleckler, Tobias Bayr, Shayne McGregor, Michael J. McPhaden,
Scott Power, Romain Roehrig, Jérôme Vialard, and Aurore Voldoire

ABSTRACT: El Niño–Southern Oscillation (ENSO) is the dominant mode of interannual climate variability on the planet, with far-reaching global impacts. It is therefore key to evaluate ENSO simulations in state-of-the-art numerical models used to study past, present, and future climate. Recently, the Pacific Region Panel of the International Climate and Ocean: Variability, Predictability and Change (CLIVAR) Project, as a part of the World Climate Research Programme (WCRP), led a community-wide effort to evaluate the simulation of ENSO variability, teleconnections, and processes in climate models. The new CLIVAR 2020 ENSO metrics package enables model diagnosis, comparison, and evaluation to 1) highlight aspects that need improvement; 2) monitor progress across model generations; 3) help in selecting models that are well suited for particular analyses; 4) reveal links between various model biases, illuminating the impacts of those biases on ENSO and its sensitivity to climate change; and to 5) advance ENSO literacy. By interfacing with existing model evaluation tools, the ENSO metrics package enables rapid analysis of multipetabyte databases of simulations, such as those generated by the Coupled Model Intercomparison Project phases 5 (CMIP5) and 6 (CMIP6). The CMIP6 models are found to significantly outperform those from CMIP5 for 8 out of 24 ENSO-relevant metrics, with most CMIP6 models showing improved tropical Pacific seasonality and ENSO teleconnections. Only one ENSO metric is significantly degraded in CMIP6, namely, the coupling between the ocean surface and subsurface temperature anomalies, while the majority of metrics remain unchanged.

https://doi.org/10.1175/BAMS-D-19-0337.1
Corresponding author: Yann Y. Planton, yann.planton@noaa.gov
Supplemental material: https://doi.org/10.1175/BAMS-D-19-0337.2
In final form 21 July 2020

AFFILIATIONS: Planton* and Vialard—LOCEAN-IPSL, CNRS-IRD-MNHN-Sorbonne Université, Paris, France; **Guilyardi**—LOCEAN-IPSL, CNRS-IRD-MNHN-Sorbonne Université, Paris, France, and National Centre for Atmospheric Science—Climate, University of Reading, Reading, United Kingdom; **Wittenberg**—NOAA/Geophysical Fluid Dynamics Laboratory, Princeton, New Jersey; **Lee and Gleckler**—Lawrence Livermore National Laboratory, Livermore, California; **Bayr**—GEOMAR Helmholtz Centre for Ocean Research, Kiel, Germany; **McGregor**—School of Earth, Atmosphere and Environment, Monash University, Clayton, Victoria, Australia; **McPhaden**—NOAA/Pacific Marine Environmental Laboratory, Seattle, Washington; **Power**—School of Earth, Atmosphere and Environment, Monash University, Clayton, and Australian Bureau of Meteorology, Melbourne, Victoria, Australia; **Roehrig and Voldoire**—CNRM, Université de Toulouse, Météo-France, CNRS, Toulouse, France

*** CURRENT AFFILIATION:** NOAA/Pacific Marine Environmental Laboratory, Seattle, Washington

El Niño–Southern Oscillation (ENSO), originating in the tropical Pacific, is a dominant mode of interannual climate variability, which has worldwide impacts through atmospheric teleconnections (McPhaden et al. 2006). Weather and climate extremes are strongly modulated by ENSO, leading to severe impacts on human activities and ecosystems. While the fate of ENSO in a warming world remains uncertain (Cai et al. 2014; Santoso et al. 2017; Timmermann et al. 2018), it is most likely that the impacts of climate change will be mediated by how ENSO evolves in the future. Climate models are essential tools for forecasting and projecting future ENSO risks, therefore a careful and regular assessment of how they represent ENSO properties, teleconnections, and mechanisms is of paramount importance. Recognizing this, the International Climate and Ocean: Variability, Predictability and Change (CLIVAR), a component of the World Climate Research Programme (WCRP), has recently established an effort to develop an ENSO metrics software package to facilitate the identification and reduction of ENSO biases in climate models, promoting ENSO literacy among model development teams and the growing audience of stakeholders using climate model outputs (Guilyardi et al. 2016).

The fact that ENSO emerges from the coupling of independently developed atmosphere and ocean component models is in itself quite an achievement. Thanks to the constant development and improvement of these component models, all latest generations of coupled climate models now exhibit ENSO-like variability in the tropical Pacific. There are still biases in the detailed properties of these simulated ENSO variations, which may be an issue for some science questions. Hence there is a need to develop “metrics” that measure how well models can simulate specific ENSO properties compared to those that are observed.

Previous metrics efforts (e.g., Guilyardi 2006; Bellenger et al. 2014; Chen et al. 2017; Ray et al. 2018a,b) have focused on better understanding how ENSO processes are represented in climate models. However, many stakeholders have more modest needs when evaluating ENSO simulations. For example, developers of atmospheric convective parameterizations may want to evaluate how certain parameters affect the overall ENSO performance in their model; or climate information users assessing future El Niño–related drought risks in East Africa may want to know which models simulate El Niño teleconnections best.

The need for a community approach to consensus model evaluation is driven by several factors (Guilyardi et al. 2020). The first factor is the complexity of climate models and the multiscale interactions that they simulate—which means that a single person, or even a single institution, may not have sufficient expertise to know which metrics are relevant to their specific question. This motivates tight coordination between model development and model evaluation. A second factor is the growing technical challenge of evaluating large and high-resolution multimodel ensembles, requiring significant and often distributed computing resources (Eyring et al. 2016b). A third factor is the growing skill set required to perform scientifically relevant and technically efficient model evaluation in order to satisfy the needs of an ever more diverse group of model users and model data consumers. For example, it is not always obvious which observational datasets should serve as the target for model evaluation, as tropical Pacific observing networks and model-based reanalyses continue to evolve (Kessler et al. 2019). Given the broad spatiotemporal diversity and interdecadal modulation of ENSO events in nature and models (Wittenberg 2004, 2009; Lee et al. 2014; Wittenberg et al. 2014; Capotondi et al. 2015, 2020; Fedorov et al. 2020), it is also not always obvious how long observational records and/or model simulations should be to robustly gauge key ENSO properties in climate models. Pioneering efforts have made progress toward shared model evaluation approaches (e.g., Stoner et al. 2009; Flato et al. 2013; Sperber et al. 2013; Lee et al. 2019; Maloney et al. 2019) or frameworks (e.g., Eyring et al. 2016c; Gleckler et al. 2016) on various climate aspects such as the climatology, monsoons, and modes of natural variability, but not much has been done specifically for ENSO up to date.

The flexible approach described in the sidebar "The three pillars of community model evaluation" provides context for the development of the CLIVAR 2020 ENSO metrics packaged presented here. For climate information users, assessing "how well the model performs" and for "what use" is essential for defining the fitness of a model for a given task. For ENSO, the International CLIVAR Pacific Region Panel (PRP; www.clivar.org/clivar-panels/pacific) started by defining three initial science questions that climate information users frequently ask: 1) How well are background climatology and basic ENSO characteristics simulated in historical simulations? 2) How well are ENSO's worldwide teleconnections represented in historical simulations? 3) How well are ENSO's internal processes and feedbacks represented in historical simulations? These three questions may be asked by different types of users: model developers, vulnerability–impacts–adaptation researchers, or ENSO experts. The PRP established "metrics collections" for individual questions—ENSO performance, ENSO teleconnections, and ENSO processes; together with collection-specific requirements (length of simulation, for observational reference, etc.—see Table 1), noting that a metric might belong to several collections. Beyond presenting this pilot effort, our goal here is to motivate end users to engage with ENSO experts to tailor metrics to their particular question.

Table 1. Information needed for each metric.

Documentation	Motivation and relevance of the metric
	Published reference(s)
	Provenance and traceability of the metric (version of the metric package, versions of all the software used, etc.)
Algorithm	Domain, land–ocean masking, averaging, or filtering
	Composite methods
	Baseline statistics (e.g., bias, correlation, RMSE)
Reference datasets	Several if possible
	Epoch(s) to use
Input requirements	Frequency: daily, monthly, yearly data
	Minimum duration of the simulation
	Minimum number of members in an ensemble
	Spatial grid/scales
Dive-down diagnostics	Curves, distributions, maps, Hovmöller diagrams
	Summary statistics (spatial mean, STD, extremes, etc.)
Plotting details	Normalization method for the portrait plot
	Color-bar type, scale, min/max range, etc.

We present here the "CLIVAR 2020 ENSO metrics package," freely available in Python (see sidebar "Using and accessing the CLIVAR 2020 ENSO metrics Python package"), and its application to the historical simulations of the two most recent phases of the Coupled Model Intercomparison Project (CMIP5 and CMIP6; see appendix A for the list of datasets used in the paper). Throughout the paper, only the first ensemble member of each model is used, except in the section on metric robustness, in which 32 historical simulations of the IPSL-CM6A-LR model are used.

The CLIVAR 2020 ENSO metrics package

To answer the three initial questions listed above, the PRP designed the CLIVAR 2020 ENSO metrics package with an eye toward the trade-offs between comprehensiveness, conciseness, and simplicity of the calculations. The initial discussions identified a large number of potential metrics (more than twice the number finally retained), which were then culled based on expert judgment as well as identification of redundant information indicated by high correlations between some metrics (discussed below). All of the selected metrics have also been established in previously published works. The current metrics collections (appendix B for method details) are labeled with an identifier "CLIVAR ENSO metrics v1.0–2020," recognizing that these metrics may be updated as knowledge, expertise, or reference datasets evolve. To follow this evolution, we have created an online resource to navigate from the summary *metrics* to the underlying *diagnostics* from which they were derived (see sidebar "Using and accessing the CLIVAR 2020 ENSO metrics Python package").

The three pillars of community model evaluation

Three types of actors can be distinguished in the workflow of model evaluation: climate information users, climate experts, and software and data engineers (Fig. SB1). Each has different requirements and expectations, expertise, and community organization. *Climate information users* can be climate modelers, climate impacts experts, or any other climate model output user. They define and own the science question. *Climate experts* have detailed knowledge of a particular area of climate science (cloud feedbacks, monsoons, ocean dynamics, El Niño, sea ice, agriculture, fisheries, etc.) and know which model evaluation to perform for which science question, the algorithms to compute relevant diagnostics and metrics, and which reference observations to use. Climate experts also know when to revise the science of model evaluation, as new knowledge emerges in their field of expertise. Finally, *software and data engineers* understand how to efficiently realize the evaluation (store and access the distributed data, program and run the calculations, visualize and archive the results, etc.) and perform the necessary software, data, and server maintenance over time. Realizing this clear separation of concerns enhances reusability, builds trust, helps to ensure scientific accuracy and relevance, and facilitates future updates for continued reliability and efficiency. Such flexibility requires that clear interfaces be devised for three types of exchanges (Fig. SB1): the science question interface, the science/IT interface, and the user interface. Embedded in this workflow is a mechanism to document each step, ensuring the provenance of each metric is kept alongside the metric value itself. The present paper explores the science/IT interface for ENSO, in support of efforts within the Infrastructure for the European Network for Earth System Modelling Phase 3 (IS-ENES3) project (https://is.enes.org/) to propose standard science/IT interfaces for a broad set of science questions.

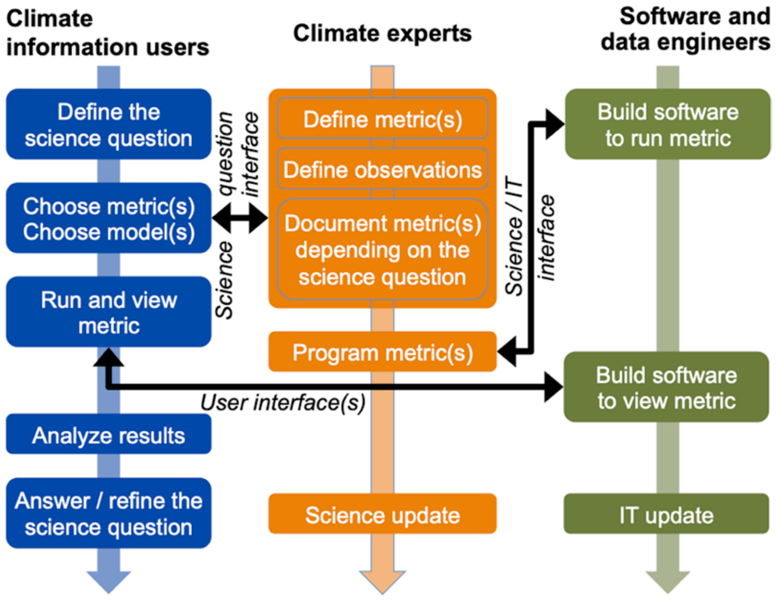
Climate information users
Climate experts
Software and data engineers
Define the science question
Choose metric(s)
Choose model(s)
Run and view metric
Analyze results
Answer / refine the science question
Science question interface
Define metric(s)
Define observations
Document metric(s) depending on the science question
Program metric(s)
Science update
Science / IT interface
Build software to run metric
Build software to view metric
IT update
User interface(s)

Fig. SB1. Building trust and ensuring efficiency: a possible framework for community climate model evaluation. The separation of concerns is organized around three types of actors, recognizing their different scopes, expectations, skills, community organizations, and workflows.

We define here a *diagnostic* as any quantity derived from a given dataset. For example, a diagnostic can be a map of the ENSO pattern (diagnostic 1; Figs. 1a–c). This *diagnostic* can be simplified into a single curve (diagnostic 2; Fig. 1d; here the fields are averaged across latitudes in the box displayed in Figs. 1a–c) to condense the information and ease model intercomparison. In contrast, we define a *metric* as a positive scalar measure of the agreement with a reference observation, with zero indicating a perfect simulation of the reference, and increasing values indicating increasing error (see appendix B). In the case of

Using and accessing the CLIVAR 2020 ENSO metrics Python package

The Python package developed to compute the CLIVAR 2020 ENSO metrics and the associated documentation are freely available at https://github.com/CLIVAR-PRP/ENSO_metrics. The package can be used for a few simulations by experienced Python users. Its application to large databases, and the associated update as new simulations are available, requires integration into dedicated software infrastructures (software and data engineers in Fig. SB1). This is first provided for the ENSO package by the PCMDI metrics package (PMP; Gleckler et al. 2016; Lee et al. 2019), with results made available via an online resource (https://cmec.llnl.gov/results/enso/), which also offers access to underlying diagnostics via interactive portrait plots. The goals of this online resource are 1) to make the results of the CLIVAR 2020 ENSO metrics package readily available so that users do not need to compute them, and 2) to routinely update the results as new CMIP6 simulations are published or as ENSO knowledge or references evolve. In time, other model evaluation infrastructures will make use of the package (integration in ESMValTool is underway) and propose additional tools to explore these metrics and their building blocks.

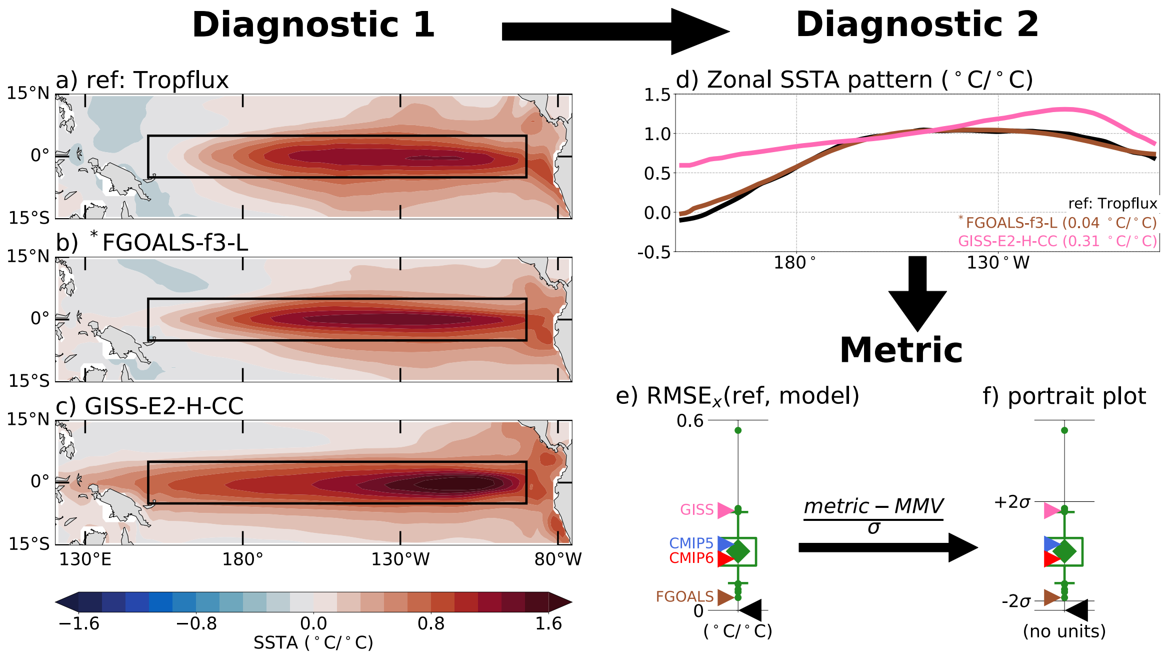

Fig. 1. The steps leading from (a)–(d) a diagnostic to (e),(f) a metric (see also sidebar "From a measured quantity to a portrait plot"). Bias in the zonal structure of sea surface temperature anomalies (SSTA) during boreal winter (ENSO_pattern): December SSTA regressed onto December Niño-3.4 SSTA (5-month triangular-weighted averaged). Map of the SSTA in the equatorial Pacific in (a)–(c) , zonal mean SSTA averaged 5°S–5°N across Pacific longitudes in (d), distribution of metric values in (e), and standardized distribution of metric values in (f) displayed in the portrait plot (Fig. 2). The metric, derived from (d), is the zonal RMSE between the model curve and the reference curve (RMSEx) (values given in the legend for the two example models displayed). In (e) and (f) triangles indicate metric values for the reference observation (black), *FGOALS-f3-L (brown), GISS-E2-H-CC (pink), and mean metric values (MMV) computed with CMIP5 dataset (blue) and CMIP6 dataset (red). Green boxplots represent the distribution of metric values computed with all CMIP models (both CMIP5 and CMIP6): whiskers extend to the 10th and 90th percentiles; boxes encompass the 25th and 75th percentiles; a diamond marks the mean (MMV); and dots indicate models that fall outside the whiskers. A star (*) before a model name indicates that the model is part of the CMIP6 dataset.

the ENSO pattern, the metric is the root-mean-square error (RMSE) between the reference dataset (black curve in Fig. 1d) and models (brown and pink curves in Fig. 1d).

ENSO overview in CMIP models

Applying the CLIVAR 2020 ENSO metrics collections to the CMIP archives (both CMIP5 and CMIP6) results in an objective overview "portrait plot" (Fig. 2) that is similar to those shown in Gleckler et al. (2008), Flato et al. (2013), Sillmann et al. (2013), and Lee et al. (2019). For the purpose of displaying all metrics in a single figure and readily distinguishing differences across models, the metric values have been standardized (see sidebar "From a measured

From a measured quantity to a portrait plot

For model evaluation or model selection applications it is often helpful to have positive-defined metrics with zero indicating a perfect simulation of the reference, and increasing values indicating increasing error. In the CLIVAR ENSO 2020 metrics package, two types of metrics are used: RMSE and scalar. RMSEs are a measurement of the distance between modeled and observed fields; they are always positive and therefore can be used directly as metrics (Fig. SB2a, first to second column). To create a metric from a scalar diagnostic (e.g., the standard deviation of a distribution), the absolute value of the relative difference [|(model − ref) × 100/ref|] is computed (Fig. SB2b, first to second column), i.e., the fractional absolute error, for which a metric value of 50 indicates that the simulated statistic differs by 50% from the observed statistic. Other adjustments of metrics are possible (e.g., Lee et al. 2019; Ahn et al. 2017), and it is important to document them properly.

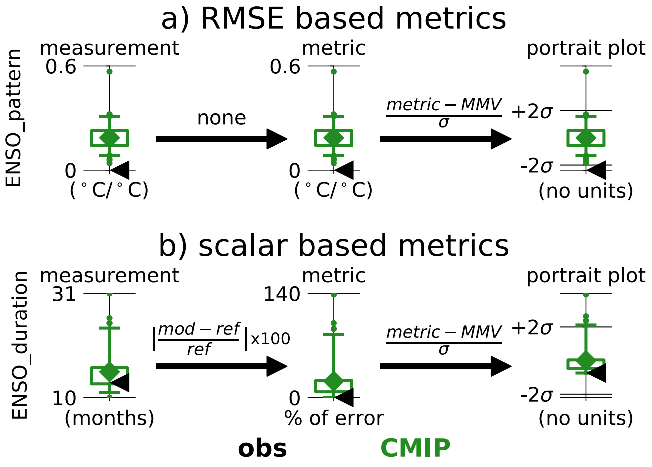

Fig. SB2. Several steps are needed to transform a diagnostic into a metric and to display it in a portrait plot using a common scale. Diagnostic, metric, and standardization definitions for (a) metrics based on root-mean-square error (RMSE; here ENSO_pattern), (b) metrics based on scalar diagnostics (e.g., linear regression slope, skewness, standard deviation; here ENSO_duration). In this context, the MMV statistic is computed by averaging the metrics across all 88 CMIP models for a given metric and *σ* is the standard deviation of the distribution of metrics values across all 88 CMIP models. Black triangles indicate metric values for the observations. Formulas above each arrow indicate the operation applied to the distribution to create a metric or standardized values to be displayed in a portrait plot (Fig. 2).

Then comes the challenge of displaying on a common scale the metric distributions that have different units and ranges (Fig. SB2, second column). A portrait plot, that uses a single color bar, created directly with the metrics distributions, does not provide useful information on the difference between the models (Figs. ES1, ES2 in the online supplement; https://doi.org/10.1175/BAMS-D-19-0337.2) as some distributions range from 0 to 1 (Fig. SB2a, second column) and others from 0 to 100 (Fig. SB2b, second column). With this in mind, it is necessary to standardize each metric distribution. To compare climate variables, it is usual to use anomalies (mean value removed) and normalize them by their standard deviation. The same kind of standardization is applied to create our portrait plot (Fig. 2): for each column, the MMV (mean value of the distribution) is removed, and values are normalized by the standard deviation of the distribution (*σ*) [(metric − MMV)/*σ*]. This step is illustrated in Figs. 1e, 1f, and SB2 (second to third columns). After this standardization, if the distributions were normal, 95% of the values would lie between −2*σ* and 2*σ*. Usually, the reference value (fourth line from the bottom in Fig. 2) falls below −2*σ* and is therefore displayed in the darkest blue. However, if the distribution is positively skewed (e.g., Fig. SB2b), the reference value can lie within the interval and appear in a light blue color.

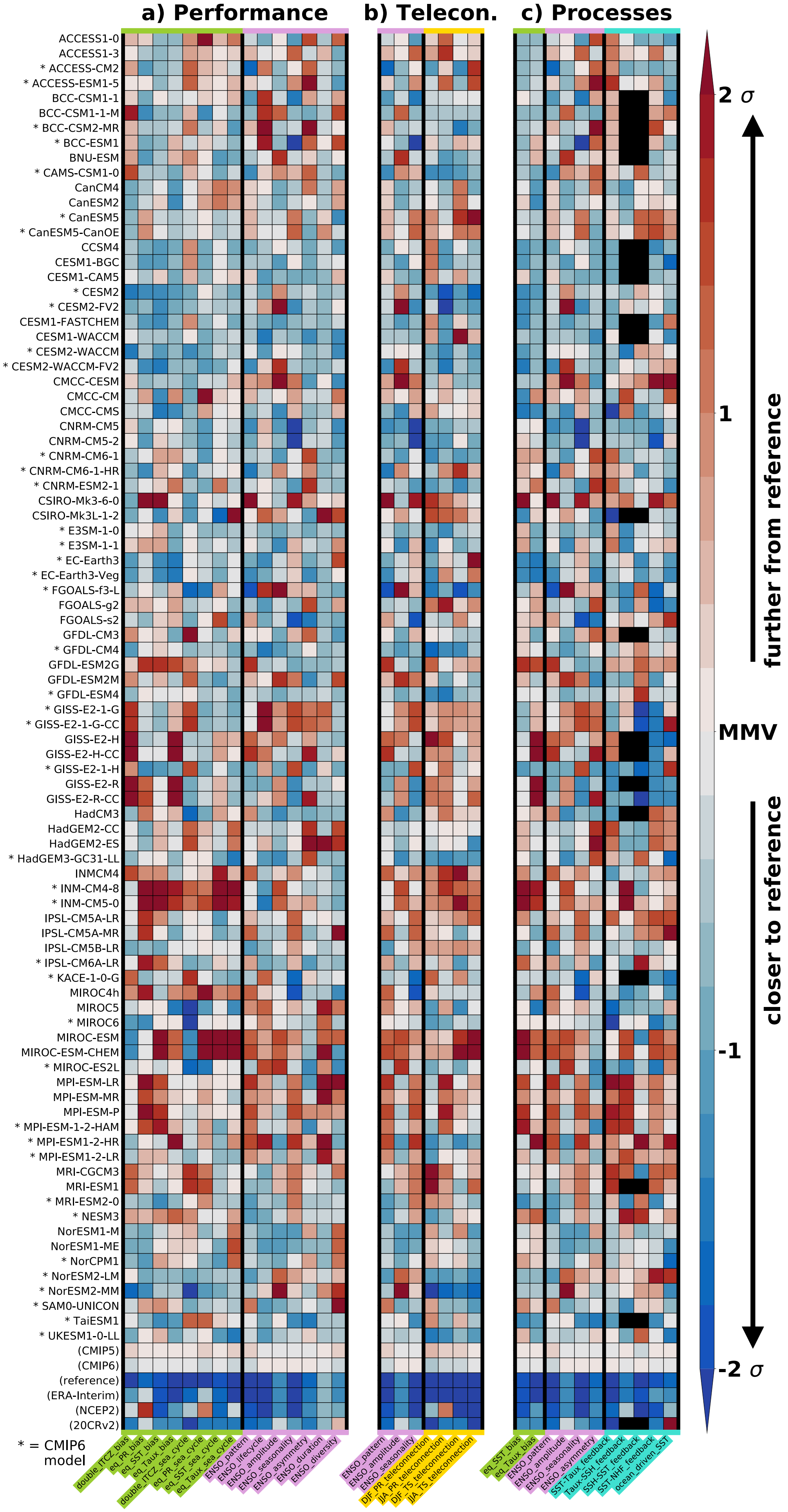

Fig. 2. The master portrait plot that provides the overall summary of the CLIVAR 2020 ENSO metrics to address the three sets of scientific questions defined by the CLIVAR Pacific Region Panel. Portrait plots of metric values relative to the MMV computed with all CMIP models (both CMIP5 and CMIP6) and normalized by the standard deviation (*σ*) of each column. Due to the standardization, here a value of 0 (white) corresponds to the MMV error, while a value of 2 as dark red (-2 as dark blue) corresponds to a model error that is two standard deviations greater (lesser) than the MMV error. The darker the blue (or the lighter the red), the closer the model is to the observational reference for a given metric. Missing data are indicated in black. At the bottom of each column, the MMVs computed with CMIP5 dataset (46 models) and CMIP6 dataset (42 models) are given, then the reference value. The last three rows compare the reference to other alternate datasets largely comprising atmospheric reanalysis (see appendix A). A star (*) before a model name indicates that the model is part of the CMIP6 dataset. Metric names highlighted in light green, light purple, yellow, and turquoise colors correspond to metrics evaluating, respectively, the background climatology, basic ENSO characteristics, teleconnections, and physical processes.

quantity to a portrait plot"). In this portrait plot, the darkest blue colors indicate models that are closest to the reference dataset for a given metric while the darkest red colors indicate models the furthest away from the reference. In addition to displaying individual model results, the bottom six rows of Fig. 2 show, respectively, the CMIP5 mean metric value (MMV; the averaged metric value of CMIP5 models), the CMIP6 MMV, and the reference observational dataset, followed by three additional observational datasets. These additional observational datasets highlight that our knowledge of climate is limited by the short observational records (Deser et al. 2017, 2018) and that in some cases, observational sampling uncertainties are comparable in magnitude to model errors (mainly for precipitation and net surface heat flux).

The portrait plot provides an overview of model fidelity for ENSO, relative to the MMV. For example, scanning across the columns indicates that the CESM1 and CNRM-CM5 model families (see Table A3) tend to outperform the MMV for most of the ENSO performance metrics, while the GISS and MIROC-ESM families tend to underperform for those metrics (Fig. 2). Differences between the CMIP5 and CMIP6 models are also apparent. Clearly, using qualifiers such as "better," "worse" or "out/underperforms" is relative and depends on the context including the science question and the metrics considered. The bottom line is that there is no physically consistent way to combine the metrics of each collection to come up with a single objective ranking. Instead, our goal here is to encourage ENSO literacy via a discussion between model data users and experts of a specific climate phenomenon.

The summary provided by the portrait plot (Fig. 2) is often not enough to answer a specific science question. To facilitate closer examination, the CLIVAR 2020 ENSO metrics package provides several levels of diagnostics, reflecting different levels of complexity. In the case of the ENSO pattern [defined here as the linear regression of sea surface temperature anomalies (SSTA) in the equatorial Pacific onto Niño-3.4 SSTA during boreal winter; Fig. 1], the metric measures that, for example, FGOALS-f3-L (brown) performs better than GISS-E2-H-CC (pink). Additional information related to the model errors is obtained from the zonal position of SSTAs and their latitudinal extent in Fig. 1 (first level of diagnostics). One may also want to know if a model is able to simulate asymmetries between El Niño (warm phase of ENSO) and La Niña (cold phase of ENSO) events (e.g., Hoerling et al. 1997). This is available from the second level of diagnostics (same analysis as in Fig. 1 but with composites of El Niño and La Niña events). Some metrics also include space–time (Hovmöller) diagrams to facilitate a better understanding of the seasonality (e.g., for precipitation biases), a map focusing on a specific region (e.g., for precipitation teleconnections; Figs. 4f–h), or provide access to other variables (e.g., for the total heat flux feedback, its four components: shortwave and longwave radiations and latent and sensible heat fluxes).

Evaluation of ENSO performance

The ENSO performance metrics collection (Fig. 2a) consists of 15 metrics for the background climatology (the time mean and seasonal cycle; color coded in light green) and basic ENSO characteristics (color coded in light purple). The first eight metrics, detailed below, illustrate well-known biases of climate models in the tropical Pacific (e.g., Guilyardi et al. 2020). The equatorial Pacific Ocean displays a "cold tongue," with locally colder sea surface temperature (SST) that extends westward from the coasts of South America. In most models this region is too cold and extends too far westward, resulting in a cold equatorial bias referred to as the "cold tongue bias" (eq_SST_bias). Linked to this SST bias, the modeled easterly trade winds are often shifted westward, resulting in trade winds that are too weak in the central Pacific, and too strong in the western Pacific (eq_Taux_bias). These biases are associated with reduced precipitation in the western Pacific (dry equator bias; eq_PR_bias) and a tendency for the intertropical convergence zone (ITCZ) to be too symmetric with respect to the equator ("double ITCZ" bias; double_ITCZ_bias). The remaining four background climatology metrics gauge the seasonal cycle amplitude of the above features. It is essential to evaluate these systematic model biases before analyzing ENSO itself, because the mean state in the tropical Pacific Ocean strongly influences ENSO characteristics and its teleconnections (e.g., Wang and An 2002; Guilyardi 2006; Sun et al. 2009; Yeh et al. 2018; Bayr et al. 2018, 2019b; Ding et al. 2020). The mean state is also generally better constrained by the available observational records than ENSO itself.

The subsequent seven ENSO performance metrics in Fig. 2a (those color coded in light purple) evaluate some basic ENSO characteristics (e.g., Hoerling et al. 1997; Capotondi et al. 2015, 2020; Guilyardi et al. 2020). ENSO is characterized by SSTA in the central and eastern equatorial Pacific (ENSO_pattern; Fig. 1), with a variability of about 1°C (ENSO_amplitude, i.e., the standard deviation). The maximum SSTA amplitudes are usually reached during boreal winter (ENSO_seasonality) and events develop during boreal spring, lasting for about a year (ENSO_duration). It is frequent that a La Niña occurs quickly after an El Niño but the reverse is not as frequent, resulting in a complex oscillation (ENSO_life cycle). However, ENSO amplitude is not symmetric, as El Niño can reach larger SSTA amplitudes than La Niña (ENSO_asymmetry). Another important aspect of ENSO is that each event is unique in terms of the SSTA pattern (ENSO_diversity).

As an example, the CESM models broadly outperform the MMV (based on the performance metrics collection), whereas the MIROC-ESM models generally underperform the MMV (Fig. 2a). It is important to remind here that "models that underperform" do not mean "models that do not simulate ENSO." For example, the GISS-E2-H-CC model underperforms in term of ENSO pattern (Fig. 1e), but simulates a pattern resembling the observed one but shifted (Figs. 1a,c). The MMV computed with CMIP5 models and CMIP6 models (respectively, sixth and fifth rows from the bottom of the plot) indicate that CMIP6 models tend to outperform the CMIP5 models, particularly for the background climatology. Figure 3 confirms this result, showing that 5 of the 15 ENSO performance metrics have significantly improved at the 95% confidence level (taking into account the different number of models currently available from CMIP5 and CMIP6)—with a reduced double ITCZ bias, improved seasonal cycle of equatorial Pacific precipitation and trade winds, ENSO pattern, and diversity.

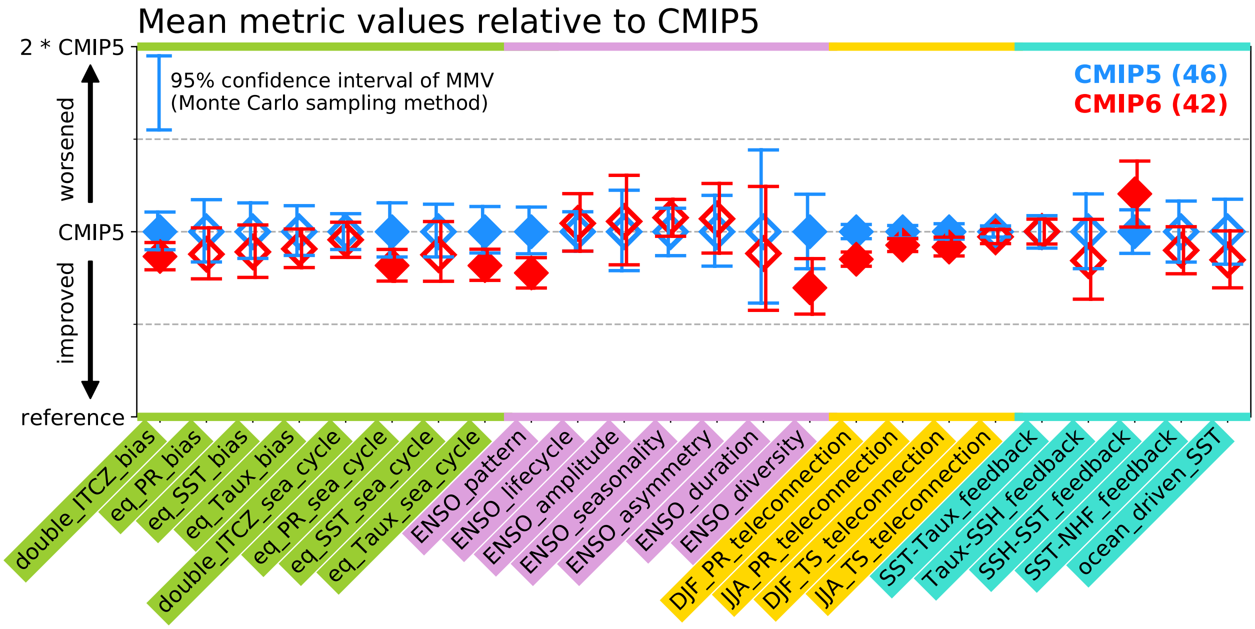

Fig. 3. Eight metrics improve and one is degraded from CMIP5 to CMIP6. Comparison of the MMVs computed with the CMIP5 dataset (46 models, in blue) and CMIP6 dataset (42 models, in red). Values are normalized by the CMIP5 MMV. Blue (red) whiskers show the 95% confidence interval on the MMV computed from CMIP6-sized (CMIP5-sized) samples drawn at random from among the 46 CMIP5 (42 CMIP6) models, using a Monte Carlo method (see appendix B for details). Solid symbols indicate differences between CMIP5 and CMIP6 MMVs are significantly different at the 95% confidence level.

Evaluation of ENSO teleconnections

The ENSO teleconnections metrics collection (Fig. 2b) includes three of the aforementioned performance metrics—the ENSO pattern, amplitude, and seasonality (color coded in light purple)—to provide essential context before considering teleconnections; and four metrics for evaluating the teleconnections themselves (color coded in yellow). During ENSO events, the large SST anomalies occurring in the equatorial Pacific induce a massive reorganization of the atmospheric circulation, influencing precipitation (PR) and surface temperature (TS) globally (e.g., Ropelewski and Halpert 1987), a phenomenon called “teleconnection.” The PR and TS teleconnection pattern metrics during boreal winter [December–February (DJF) average] between 60°S and 60°N are evaluated with a spatial RMSE (PR teleconnection illustrated in Fig. 4 in panel “global”). At first glance, the global map of PR teleconnection during boreal winter of a model among the closest to the reference (Fig. 4c) and a model among the farthest away (Fig. 4d) look very similar, indicating that all climate models reproduce ENSO teleconnections. A closer look highlights that the second model tends to underestimate teleconnections over the Pacific and Atlantic Oceans, and overestimates them over the Indian Ocean (Fig. 4d). A similar evaluation is done during boreal summer [June–August (JJA) average]. These metrics aim to evaluate

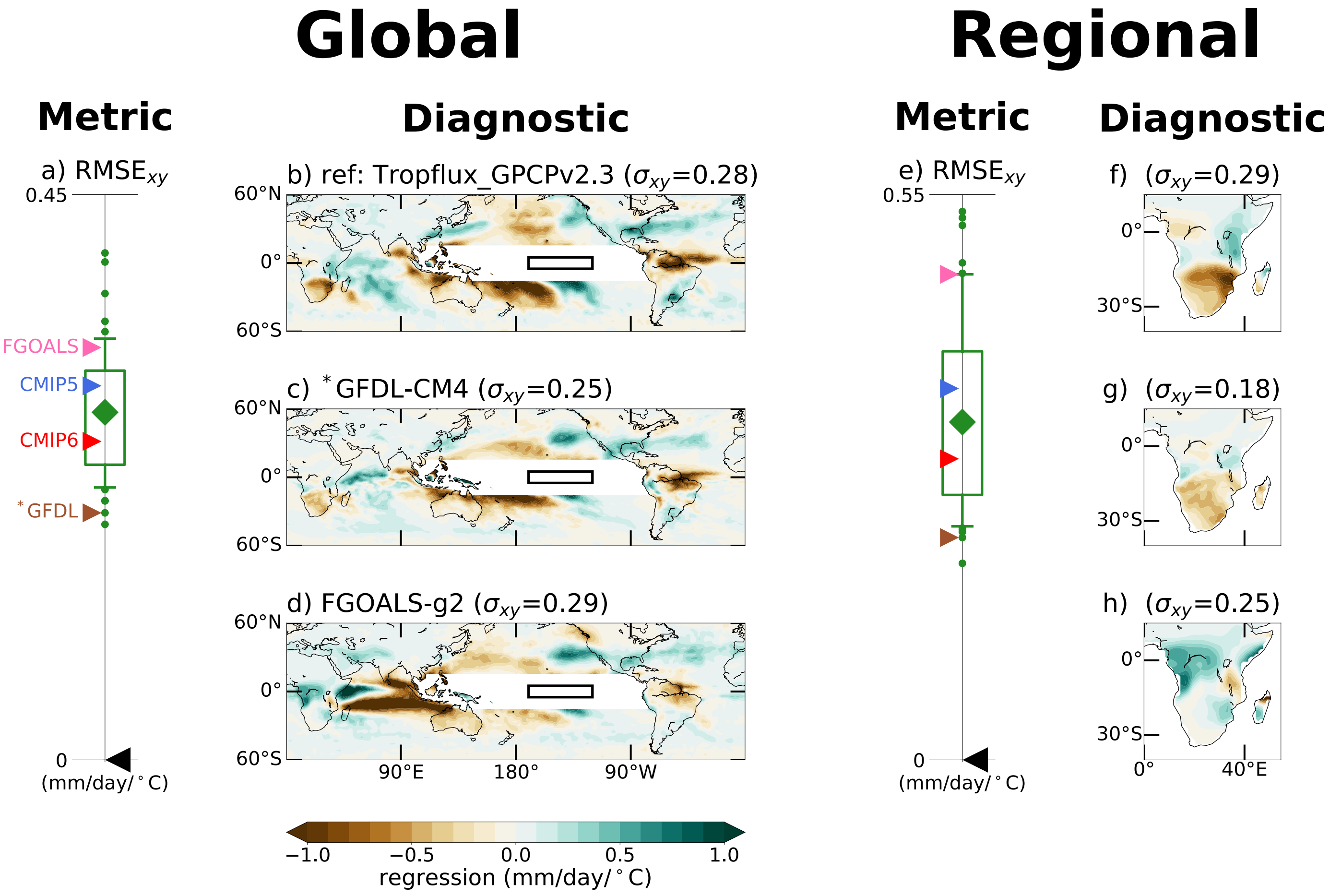

Fig. 4. Examples of diagnostics provided in support of metrics for precipitation teleconnection. ENSO precipitation (PR) teleconnection patterns during boreal winter (DJF): (b)–(d) DJF PR regressed onto DJF Niño-3.4 SSTA. (a) Metric derived from (b)–(d): the spatial (at grid point) RMSE (RMSE$_{xy}$) (DJF_PR_teleconnection; uncentered and biased statistic); (f)–(h) focus on teleconnections over land in the southern half of Africa and (e) associated metric. Teleconnections computed for the observation in (b) and (f), *GFDL CM4 in (c) and (g), and FGOALS-g2 in (d) and (h). In (a) and (e), triangles indicate metric values for the observations (black), *GFDL CM4 (brown), FGOALS-g2 (pink), and MMV computed with CMIP5 dataset (blue) and CMIP6 dataset (red). The equatorial Pacific Ocean (15°S–15°N) is masked out, before computing the metrics, to highlight the remote teleconnection patterns. Spatial standard deviation (σ_{xy}) of each field is given in parentheses in (b)–(d) and (f)–(h).

ENSO-driven anomalies during key seasons (e.g., Power and Delage 2018), but one should keep in mind that they may not entirely be caused by ENSO. For example, there is a debate whether PR anomalies over the tropical Indian Ocean during boreal winter are mainly influenced by local SSTA (Xie et al. 2002; Annamalai et al. 2005) or directly by ENSO SSTA (Izumo et al. 2020).

For the ENSO teleconnections metrics collection (Fig. 2b), the CNRM and NorESM2 models are among the models closest to the reference, while the INM and MIROC-ESM models are among the farthest away. It can also be noted that the CESM and GFDL simulated teleconnections substantially improve from CMIP5 to CMIP6. Three out of four teleconnection metrics (highlighted in gold in Fig. 3) have significantly improved from to CMIP6. This improvement needs to be confirmed by looking at more ensemble members since the pattern and intensity of ENSO-related metrics can vary from one member to the next for these relatively short 165-yr historical simulations (illustrated when discussing Fig. 8; Batehup et al. 2015; Perry et al. 2017). In addition, simplified metrics based on regionally averaged anomalies (e.g., Power and Delage 2018; Perry et al. 2020), which appear to be less sensitive to external (non-ENSO) climate variability, will be introduced in the next version of the CLIVAR 2020 ENSO metrics package to ensure the robustness of the improved teleconnection. Note that a better representation of the teleconnection patterns does not imply a better simulation of the physical processes causing them due to error compensation (Annamalai 2020).

Evaluation of ENSO processes

The ENSO processes metrics collection (Fig. 2c) also includes six previously defined ENSO performance metrics to provide context. The first two metrics (color coded light green) are for evaluating cold tongue and the trade winds biases, which strongly affect the intensity of ENSO feedbacks by shifting the raising branch of the Pacific Walker circulation too far west (e.g., Bayr et al. 2018, 2019a). The next four metrics (color coded light purple) are for the ENSO pattern, amplitude, seasonality, and asymmetry, which provide essential information about ENSO's structure, intensity, seasonality, and nonlinearity, a prerequisite for examining the feedbacks that generate them (five metrics color coded in turquoise). ENSO is the result of the amplification of SSTA through the positive ocean–atmosphere Bjerknes feedback (Bjerknes 1969; Jin 1997): warm SSTA weakens the trade winds, deepening the thermocline, which favors the upwelling of warmer water, increasing the initial SSTA (Fig. 5a). These same processes operate in the opposite sense to amplify cold SSTA. This feedback is evaluated in three steps: the atmospheric branch, linking eastern equatorial Pacific SSTA to remote western/central equatorial Pacific trade winds anomalies (e.g., Bellenger et al. 2014; Bayr et al. 2019a; SST-Taux_feedback; Figs. 5b,c); the ocean–atmospheric branch, linking the remote trade winds anomalies to subsurface temperature anomalies in the eastern equatorial Pacific (e.g., Bayr et al. 2019a; Taux-SSH_feedback; Figs. 5d,e); and the oceanic branch, linking the subsurface temperature anomalies to eastern equatorial Pacific SSTAs (e.g., Bayr et al. 2019a; SSH-SST_feedback; Figs. 5f,g). The Bjerknes feedback is damped by the negative heat flux feedback (e.g., Zebiak and Cane 1987; Lloyd et al. 2012; Bellenger et al. 2014; SST-NHF_feedback; Fig. 5a): warm SSTA increases cloud cover and tends to decrease warming by reducing incoming shortwave radiation, damping the initial SSTA (e.g., Im et al. 2015). Cold SSTA are damped by the same processes operating in an opposite sense. The last metric evaluates the balance between the effect of the Bjerknes feedback and heat fluxes on the development of ENSO SSTA (ocean_driven_SST; see Bayr et al. 2019a for more details).

Because the exploration of ENSO processes is still very much a research topic, this list was purposely kept short, unlike other comprehensive analysis frameworks (e.g., Jin et al. 2006; Graham et al. 2017; Chen et al. 2017; Ray et al. 2018a,b). To explain ENSO variability, more processes need to be considered, including nonlinear dynamical heating, tropical instability waves, equatorial Kelvin and Rossby waves, westerly wind events, involving numerous nonlinearities (An et al. 2020). It is also common to consider separately the influence of the net

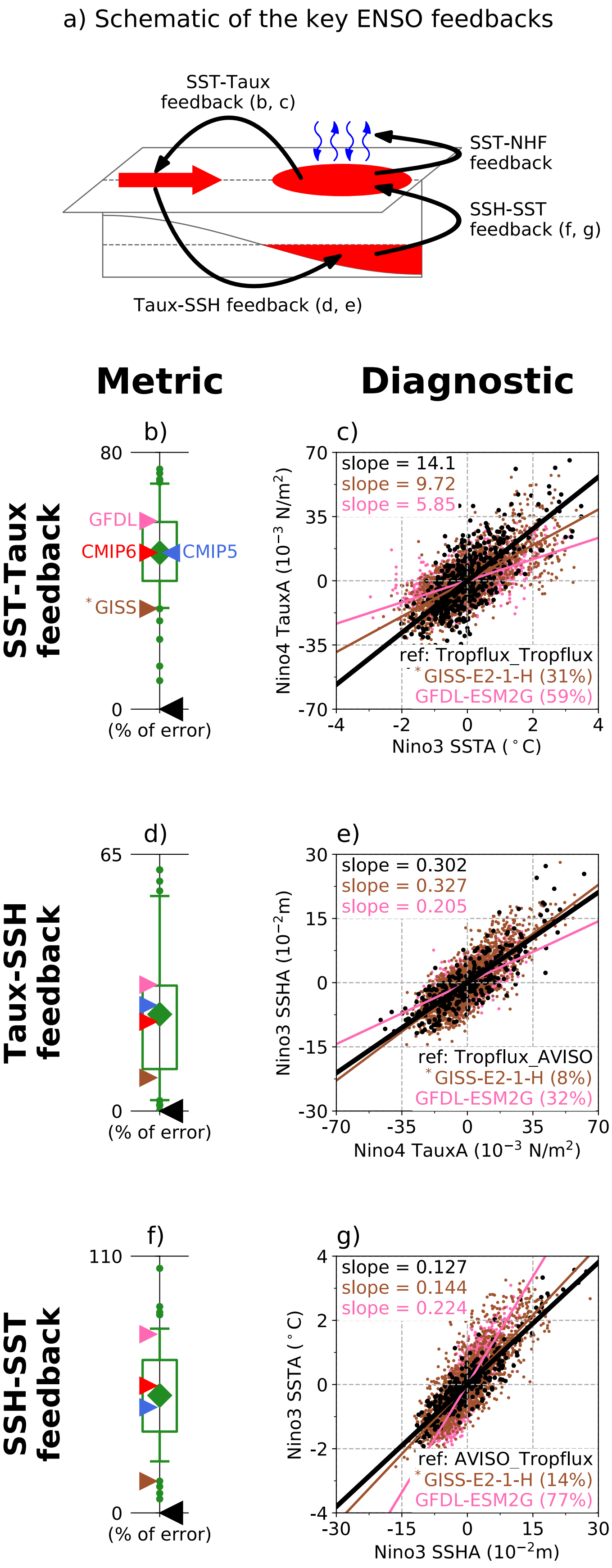

Fig. 5. A figure for ENSO experts, exploring the underlying properties of ENSO feedbacks using several metrics. Schematics of (a) the key ENSO feedbacks; (b),(c) SST-Taux feedback (SST-Taux_feedback), i.e., Niño-4 TauxA regressed onto Niño-3 SSTA; (d),(e) Taux-SSH feedback (Taux-SSH_feedback), i.e., Niño-3 SSHA regressed onto Niño-4 TauxA; and (f),(g) SSH-SST feedback (SSH-SST_feedback), i.e., Niño-3 SSTA regressed onto Niño-3 SSHA. The diagnostic extracted from the scatterplots in (c), (e), and (g) is the slope of the linear regression line (values given in the top-left corner). The metric is then the absolute value of the relative difference between the model and reference slopes [|(model - ref)/ref|] (metric values in % given with the legend). In (b), (d), and (f), triangles indicate metric values for the observations (black), *GISS-E2–1-H (brown), GFDL-ESM2G (pink), and MMV of CMIP5 dataset (blue) and CMIP6 dataset (red). Green boxplots represent the distribution of metric values computed with all CMIP models: whiskers extend to the 10th and 90th percentiles; boxes encompass the 25th and 75th percentiles; a diamond marks the mean (MMV); and dots indicate models that fall outside the whiskers. Regions are defined in Table A2 and Fig. A1.

heat flux components (latent heat flux, sensible heat flux, longwave radiation and shortwave radiation) on ENSO because they can have opposite effects. As we found that there is a better agreement among observational datasets on the net heat flux and its interannual anomalies than for its individual components, only the metrics using the net heat flux were kept. However, this metrics collection can already help to determine whether correct ENSO simulation performance occurs for the right reasons, i.e., from the correct balance of processes and not via error compensation (Guilyardi 2006; Bayr et al. 2019a). For example, the MPI models have close-to-observed ENSO SSTA amplitude, but apparently for the wrong reasons (Fig. 2c), due to compensation between weak atmospheric and ocean–atmospheric branches of the Bjerknes feedback (which would act to weaken ENSO) and a weak surface heat flux damping of SSTAs (which would act to strengthen ENSO). This error compensation between a too weak amplification and a too weak damping is a common problem in climate models and affects about half of the CMIP5 models (Bellenger et al. 2014, Bayr et al. 2018, 2019a). This hampers the simulation of important ENSO properties, such as the phase locking of ENSO to the seasonal cycle (Wengel et al. 2018) or the asymmetry between El Niño and La Niña (Bayr et al. 2018).

None of the five metrics specific to this collection (color coded in turquoise in Fig. 3) have significantly improved from CMIP5 to CMIP6 and one was significantly degraded (highlighting a too strong coupling between subsurface and sea surface temperatures).

Relationships among ENSO metrics

To understand the links among metrics, and to reduce redundancies within the metrics collection, we have computed intermetric correlations across the CMIP model ensemble (Fig. 6). For example, it confirms previous findings indicating that the dry equator bias (eq_PR_bias) and cold tongue bias (eq_SST_bias) biases are linked (correlation of 0.6; e.g.,

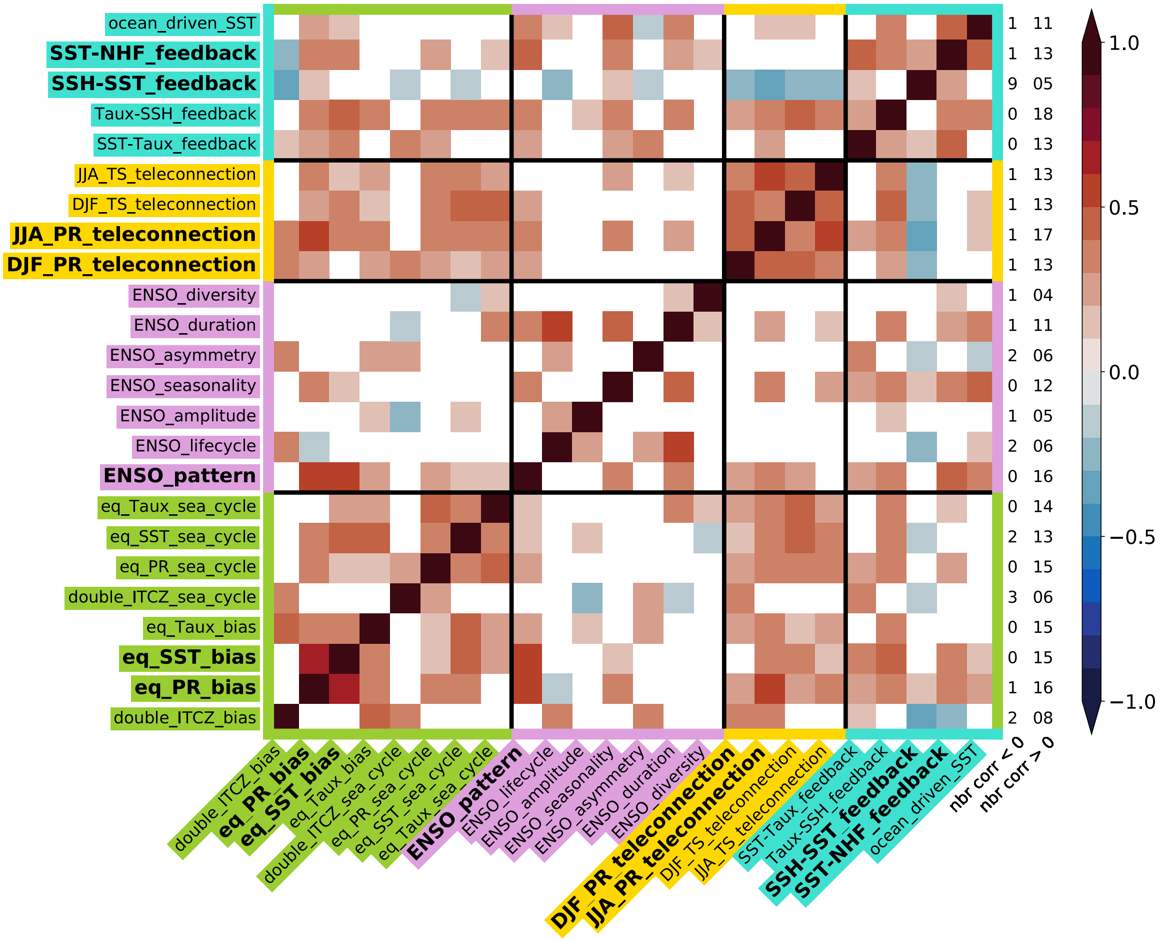

Fig. 6. The analysis of correlations between metrics highlights some known relationships (e.g., precipitation and SST biases) and new surprising ones (such as the oceanic branch of the Bjerknes feedback found to be anticorrelated with several metrics). Intermetric correlations, computed across the CMIP models (both CMIP5 and CMIP6). Only correlations significant at the 95% confidence level are shaded. Numbers to the right of the figure are the number of significant negative correlations and the number of significant positive correlations. Under the null hypothesis that the correlation is 0, significance is estimated using a two-sided *p* value for the Wald test, assuming a *t*-distributed test statistic. The correlation matrix is symmetric about the diagonal. Metrics specifically discussed in the text are highlighted in bold.

Oueslati and Bellon 2015; Brown et al. 2020). It also shows that the ENSO pattern metric (ENSO_pattern) is significantly correlated with 16 of the 23 other metrics, in particular the two aforementioned mean state bias (correlation of 0.6 for both) and the heat flux feedback metric (SST-NHF-feedback; correlation of 0.5). This is consistent with previous studies showing that the cold tongue bias induces an eastward shift of the ENSO pattern (e.g., Bellenger et al. 2014; Brown et al. 2020), and that ENSO SSTA in models are too strongly controlled by surface heat fluxes (e.g., Bayr et al. 2019a). The teleconnection metrics, color coded in yellow, are correlated with each other (0.4 or higher) but surprisingly they tend not to be systematically linked to basic ENSO characteristics (color coded in light purple). Another interesting link arises from the correlation between the dry equator bias and PR teleconnection: the bias has a stronger influence during boreal summer (JJA_PR_teleconnection; correlation of 0.5) than winter (DJF_PR_teleconnection; correlation of 0.2). Of equal interest, the oceanic branch of the Bjerknes feedback (SSH-SST_feedback) is significantly anticorrelated with more than third of the metrics (9 out of 23). Such correlations and anticorrelations clearly generate as many questions as answers, and open new avenues for future research, beyond the scope of this paper.

Selecting models for particular applications

The metrics package can be used to select models that are well suited to address particular scientific questions or societal needs. As there are many possible applications for climate simulations, providing a single objective procedure to select models is not possible. The metrics used to select models and their relative weights in the decision-making need to be decided through a discussion between users and ENSO experts. For example, a user may want to analyze the present PR teleconnection during boreal winter. A first simple approach would be to select the models that best reproduce the corresponding metric (DJF_PR_teleconnection), noting that the definition of “best” requires some leeway to allow for the potential impact of non-ENSO related climatic noise on the observed and modeled teleconnections (e.g., Deser et al. 2018; Perry et al. 2020). Furthermore, as ENSO-driven teleconnections are controlled by changes in the strength of the Walker circulation (e.g., Klein et al. 1999; Diaz et al. 2001; Wang 2002), it is important to select models that also reproduce this process (evaluated by the atmospheric branch of the Bjerknes feedback; SST-Taux_feedback). Using these two metrics, one can retain models that are ranked among the top 50% (a threshold that can be changed depending on user’s needs and the estimated impacts of non-ENSO related climatic noise), giving a set of 17 models. This model subset performs significantly better than the full CMIP ensemble for several metrics (Fig. 7a), including the PR biases and their seasonal cycles, the cold tongue bias, the ENSO pattern and skewness, and summer PR and winter TS teleconnections. Nevertheless, if the user’s interest lies in regional teleconnections over land, this subset is not adequate as it does not reproduce winter PR teleconnections over South Asia and Oceania any better than the CMIP ensemble (Fig. 7b). While this subset also better simulates winter TS teleconnection in this same region (Fig. 7d), it is definitely not suited for summer teleconnection over land (Figs. 7c,e). This example indicates again that the set of models that should be chosen for a particular study need to be tailored for its objectives (e.g., ENSO rainfall signature over South Asia). The metrics package we propose allows to make educated choices, by selecting models based on the most appropriate metrics.

Metric robustness

ENSO properties vary at decadal to centennial time scales (e.g., Wittenberg 2009; Li et al. 2013; McGregor et al. 2013; Carré et al. 2014; Liu et al. 2017; Fedorov et al. 2020), and climate models are increasingly able to simulate such long-term variations of ENSO (e.g., Yeh and Kirtman 2004; Atwood et al. 2017; Guilyardi et al. 2020). Therefore, using only one simulation of the past 165 years may not be enough to robustly evaluate ENSO in climate models. To obtain a better estimate of the model errors, one can use large ensembles of simulations with differing initial conditions

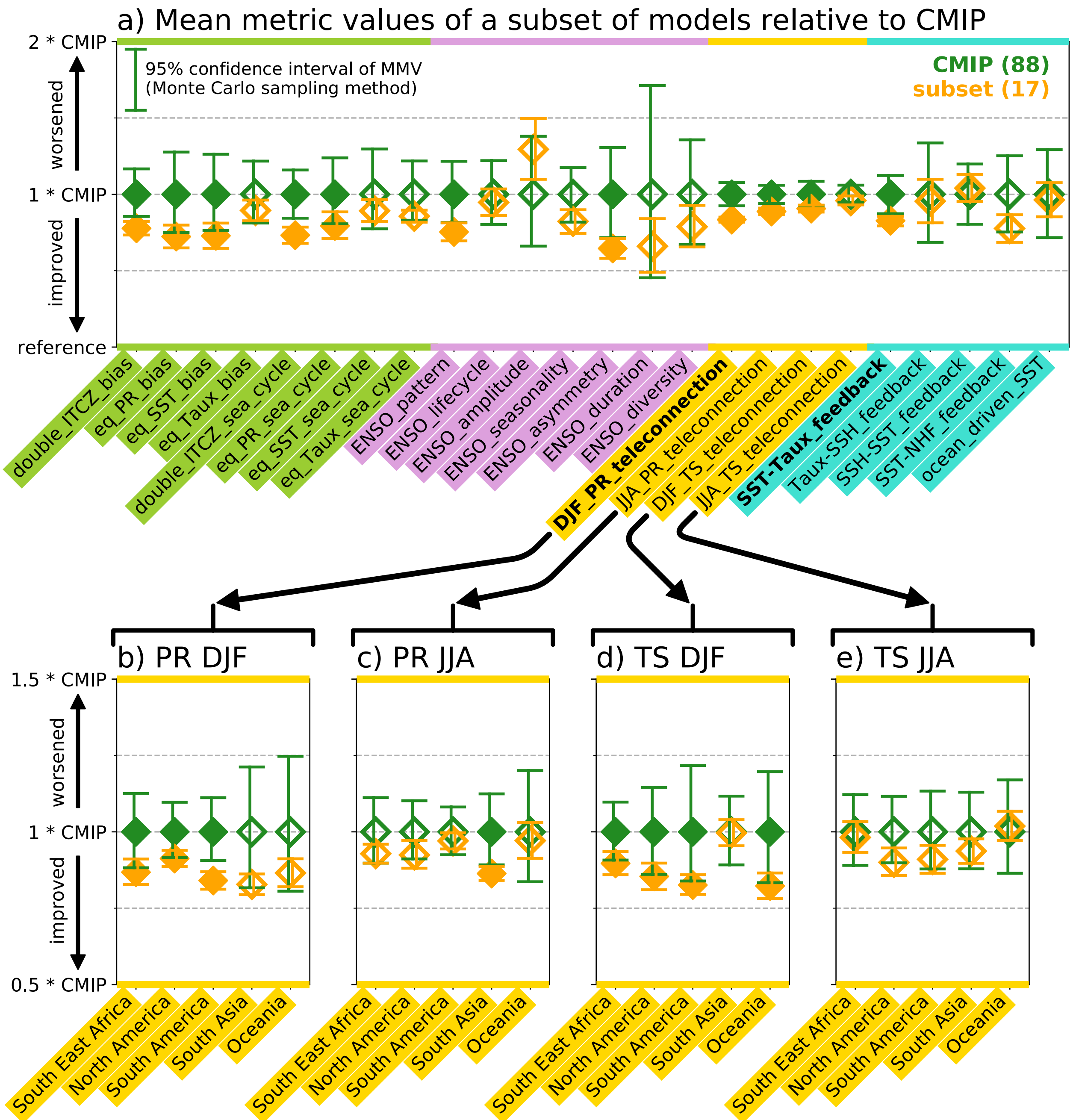

Fig. 7. A subset of 17 models is defined to analyze PR teleconnection during boreal winter. A finer analysis of the performance of this subset, using the associated online diagnostics, highlights that it is not suited for every region, season, or field. This underlines the fact that a single objective procedure to select models does not exist and that the selection must be done via an exchange between users and ENSO experts, hence contributing to ENSO literacy. (a) Comparison of the MMVs computed with all CMIP models (88 models; green) and a subset of models (17 models; orange) that performs well for DJF_PR_teleconnection and SST-Taux_feedback (in bold). The 17 models are (* markers indicate CMIP6 models): *CESM2, *CESM2-FV2, *CESM2-WACCM, *CESM2-WACCM-FV2, CMCC-CM, CNRM-CM5, CNRM-CM5–2, *EC-Earth3, *EC-Earth3-Veg, *FGOALS-f3-L, FGOALS-s2, *GFDL CM4, *GFDL-ESM4, *MIROC-ES2L, *MIROC6, *NESM3, and *NorESM2-MM. (b)–(e) As in (a), but for the regional teleconnection over land (level 2 diagnostics) of DJF_PR_teleconnection in (b), JJA_PR_teleconnection in (c), DJF_TS_teleconnectionin (d), and JJA_TS_teleconnection in (e). Values are normalized by the CMIP value. Green (orange) whiskers show the 95% confidence interval on the MMV computed from subset-sized (CMIP-sized) samples drawn at random from among the 88 CMIP models (17 models of the subset), using a Monte Carlo method (see appendix B for details). Solid symbols indicate differences between CMIP and subset MMVs are significantly different.

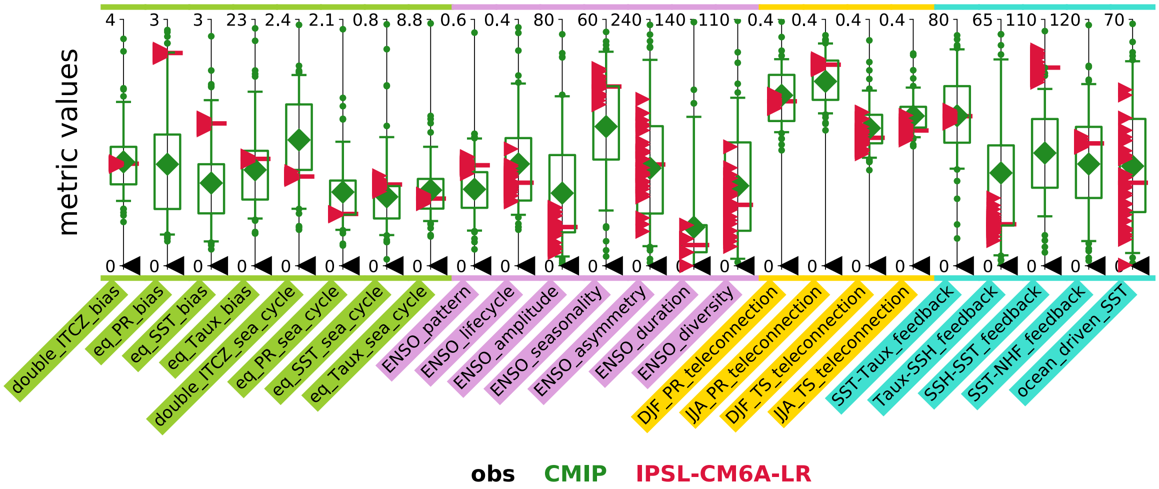

Fig. 8. For some key ENSO metrics, simulations ensembles are needed to provide a robust evaluation. Parallel coordinate plot of metric values showing an ensemble of 32 historical simulations available for the model IPSL-CM6A-LR. Triangles indicate values for the observations (black) and IPSL-CM6A-LR (red); horizontal red lines indicate the average value of the model's ensemble.

(Deser et al. 2020) as, for example, the 32 historical simulations available for the IPSL-CM6A-LR model (Fig. 8; Boucher et al. 2020). Not surprisingly, the background climatology (time-mean and seasonal cycle) metrics show little intermember variability compared to intermodel variability. On the other hand, some ENSO metrics exhibit much larger intermember variations for this model, and likely the case for other models. For example, for the metrics evaluating the ENSO asymmetry (ENSO_asymmetry) and the balance between Bjerknes feedback and heat fluxes during ENSO (ocean_driven_SST), this single model ensemble range covers about two thirds of the CMIP range. The wide range of metric values obtained with the IPSL-CM6A-LR ensemble, especially concerning ENSO diversity (ENSO_diversity) and TS teleconnection during boreal winter (DJF_PR_teleconnection), raises questions regarding the robustness of some our results interpreting ENSO performance changes between CMIP5 and CMIP6. This caveat highlights the need to consider large ensembles when applying our ENSO metrics.

Summary

The CLIVAR 2020 ENSO metrics package has been developed through consensus-seeking discussions among CLIVAR ENSO experts over several years, with an eye toward the trade-offs between comprehensiveness, conciseness, and simplicity of the calculations. It was specifically designed for stakeholders primarily interested in knowing how well ENSO is simulated in climate models for reasons other than understanding ENSO itself. The package is particularly well suited for benchmarking ENSO performance as models evolve, and for identifying the relative strengths and weaknesses of different models.

The metrics collections and their summary portrait plot (Fig. 2) allow rapid and concise comparative evaluation of ENSO in models. It underscores persistent challenges: three of the four key ENSO feedbacks have an average error of 50%, and the average error for ENSO asymmetry exceeds 100%. To better understand the structures and sources of individual model errors, users can leverage different levels of diagnostic information provided by the metric package (e.g., Figs. 1, 4, and 5), including those from which the metrics were derived.

The first comparison of CMIP5 (46 models) and CMIP6 (42 models available at the time of revisions) shows that, of the 24 metrics used to evaluate ENSO in the CLIVAR 2020 ENSO metrics package, 8 significantly improved in CMIP6 compared to CMIP5 and 1 was significantly degraded while the majority remained unchanged (Fig. 3). The main improvements in CMIP6 are reduced biases in the tropical Pacific time-mean, seasonal cycle, ENSO patterns, diversity of ENSO events, and remote teleconnections. On the other hand, the representation of the processes linking subsurface and surface temperature anomalies in the eastern equatorial Pacific was degraded from CMIP5 to CMIP6.

Benchmarking performance changes across model generations will continue as additional model simulations are contributed to the CMIP6 database (see sidebar “Using and accessing the CLIVAR 2020 ENSO metrics Python package”). Work is also underway to quantify the sensitivity of the metrics to internal variability, namely, by diagnosing large ensembles performed by several modeling groups. The role of observational uncertainty is a critical factor for some metrics (e.g., precipitation and total heat flux) and will be addressed in a subsequent version of the package.

Computing intermetric correlations across the large CMIP5 and CMIP6 multimodel ensemble is helpful to detect relationships among metrics, reduce redundancy in the metrics package, and highlight relationships among model errors. For example, the metric evaluating the link between subsurface and surface temperature anomalies is found to be negatively correlated with several other metrics, suggesting that improving this metric in the current generation of models could negatively impact skill in other characteristics, or vice versa. This opens up a host of new scientific questions, which may help to accelerate progress toward improved modeling and understanding of ENSO and its impacts.

The CLIVAR 2020 ENSO metrics package is a pilot implementation written in Python, which in recent years has become very popular among climate scientists. The collaborative development and interface with the PCMDI metrics package provide a proof of concept in the science/IT interface (see sidebar “The three pillars of community model evaluation”). Given the widespread interest in ENSO, it is likely that this package will be adopted by other community diagnostic infrastructures (e.g., the ESMValTool model evaluation framework; Eyring et al. 2016c; Righi et al. 2020). Efforts are underway to identify and address any remaining technical and scientific challenges related to the package, leveraging it to improve models and spur new scientific investigations. Our hope is that this package leads to more productive collaboration among climate experts, climate information users, and software and data engineers, and that it provides an example for other such efforts to build upon.

Acknowledgments. The authors are grateful to H. Annamalai and two anonymous reviewers for helpful comments that substantially improved the manuscript. We thank the International CLIVAR Pacific Panel members and ENSO experts for a number of fruitful discussions over the years, and more specifically Pascale Braconnot, Wenju Cai, Antonietta Capotondi, Alexey Fedorov, Fei-Fei Jin, Matthieu Lengaigne, Xiaopei Lin, as well as the support of the International CLIVAR Project Office, Jose Santos and Jing Li. We acknowledge the support from the Agence Nationale de la Recherche ARISE project, under Grant ANR-18-CE01-0012, and the Belmont project GOTHAM, under Grant ANR-15-JCLI-0004-01, the European Commission’s H2020 Programme “Infrastructure for the European Network for Earth System Modelling Phase 3 (IS-ENES3)” project under Grant Agreement 824084, the Deutsche Forschungsgemeinschaft (DFG) project “Influence of Model Bias on ENSO Projections of the 21st Century” through Grant 429334714, the Australian Research Council through Grant FT160100162, and CMIP modeling groups for their support. Feedback on the model evaluation framework was provided via discussions with the World Climate Research Programme’s Working Group on Numerical Experimentation (WGNE) and the Working Group on Coupled Modeling (WGCM). YYP currently holds a National Research Council (NRC) Research Associateship Award at NOAA/PMEL. Part of the work was performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contract DE-AC52-07 NA27344 via the Regional and Global Climate Modeling Program. Support for the Twentieth Century Reanalysis Project dataset is provided by the U.S. Department of Energy, Office of Science Innovative and Novel Computational Impact on Theory and Experiment (DOE INCITE) program, and Office of Biological and Environmental Research (BER), by the National Oceanic and Atmospheric Administration Climate Program Office, and by the National Oceanic and Atmospheric Administration Climate Program

Office, and by the NOAA Physical Sciences Laboratory. We acknowledge the World Climate Research Programme's Working Group on Coupled Modelling, which is responsible for CMIP, and we thank the climate modeling groups (listed in Table A3 of this paper) for producing and making available their model output. For CMIP, the U.S. Department of Energy's support of the Earth System Grid Federation enables data access and software infrastructure in partnership with the Global Organization for Earth System Science Portals. The Twentieth Century Reanalysis V2 (20CRv2), Global Ocean Data Assimilation System (GODAS), Global Precipitation Climatology Project version 2.3 (GPCPv2.3), and NCEP–DOE AMIP-II Reanalysis (NCEP2) data are provided by the NOAA/OAR/ESRL/PSL, Boulder, Colorado. The altimeter products were produced by SSALTO/DUACS and distributed by AVISO+, with support from CNES. ERA-Interim data are provided by ECMWF. The Simple Ocean Data Assimilation ocean/sea ice reanalysis version 3.4.2 (SODA3.4.2) is supported by the National Science Foundation Physical Oceanography Program, NOAA/GFDL, NOAA/NCEP, NOAA/NESDIS, NASA GMAO, and the NASA MAP and Physical Oceanography programs. The TropFlux data are produced under a collaboration between Laboratoire d'Océanographie: Expérimentation et Approches Numériques (LOCEAN) from Institut Pierre Simon Laplace (IPSL; Paris, France) and National Institute of Oceanography/CSIR (NIO; Goa, India), and supported by Institut de Recherche pour le Développement (IRD; France). TropFlux relies on data provided by the ECMWF interim reanalysis (ERA-Interim) and ISCCP projects. This is PMEL Contribution 5050.

Data availability statement. CMIP5 and CMIP6 data can be accessed at https://esgf-node.llnl.gov/projects/esgf-llnl/. 20CRv2, GODAS, GPCPv2.3, NCEP2 data can be downloaded from https://psl.noaa.gov/, AVISO from www.aviso.altimetry.fr, ERA-Interim from www.ecmwf.int/, SODA3.4.2 from https://www2.atmos.umd.edu/~ocean/, TropFlux from https://incois.gov.in/tropflux/.

The latest release of the CLIVAR 2020 ENSO metrics package can be accessed at https://github.com/CLIVAR-PRP/ENSO_metrics/releases/tag/v1.0-2020, the associated documentation at https://github.com/CLIVAR-PRP/ENSO_metrics/wiki and the results computed with the PCMDI metrics package at https://cmec.llnl.gov/results/enso/.

Table A1. Reference datasets used in the CLIVAR 2020 ENSO metrics package. Field acronyms defined with the first two datasets of the table.

Reference dataset	Reference publication	Period	Fields
20 CRv2	Compo et al. (2011)	1871–2012	Precipitation (PR) Surface temperature (TS) Net heat flux (NHF) Latent heat flux (LHF) Sensible heat flux (SHF) Longwave radiation (LWR) Shortwave radiation (SWR) Zonal wind stress (Taux)
AVISOv6.3 (gridded altimeter products)	—	1993–2018	Sea surface height (SSH)
CMAP	Xie and Arkin (1997)	1979–2018	PR
ERA-Interim (reanalysis)	Dee et al. (2011)	1979–2018	PR, TS, NHF, LHF, SHF, LWR, SWR, Taux
ERSSTv5	Huang et al. (2017)	1854–2018	Sea surface temperature (SST)
GODAS (reanalysis)	Saha et al. (2006)	1980–2018	SSH
GPCPv2.3 (gridded analysis)	Adler et al. (2003)	1979–2018	PR
HadISSTv1.1 (gridded analysis)	Rayner et al. (2003)	1870–2018	SST
NCEP2 (reanalysis)	Kanamitsu et al. (2002)	1979–2018	PR, TS, NHF, LHF, SHF, LWR, SWR, Taux
SODA3.4.2 (reanalysis)	Carton et al. (2018)	1980–2017	SSH
TropFlux (mix between ERA-Interim and ISCCP corrected with observations)	Praveen Kumar et al. (2012, 2013)	1979–2018	SST, NHF, LHF, SHF, LWR, SWR, Taux

Table A2. Regions defined in the CLIVAR 2020 ENSO metrics package—see Fig. A1.

Region	Eastern Pacific	Equatorial Pacific	global60	Niño-3	Niño-3.4	Niño-4
Coordinates	90°–150°W 15°S–15°N	150°E–90°W 5°S–5°N	0°–360°E, 60°S–60°N	90°–150°W 5°S–5°N	120°–170°W 5°S–5°N	160°E–170°W 5°S–5°N

Appendix A: Datasets

The details (names, references, periods, and fields) of the reference datasets used by the metrics package are given in Table A1, and the regions are defined in Table A2 (displayed in Fig. A1). In this paper, the reference is composed of AVISO's sea surface height (SSH), ERA-Interim's TS, GPCPv2.3's PR, and TropFlux's SST, net surface heat flux (NHF), and surface zonal wind stress (Taux). The additional observational datasets displayed in the portrait plot also combine several datasets: the one labeled ERA-Interim is composed of SODA3.4.2's SSH (this ocean reanalysis is forced by ERA-Interim) and ERA-Interim for all other variables; the one labeled NCEP2 is composed of GODAS's SSH (this ocean reanalysis is forced by NCEP2) and NCEP2 for all other variables; the one labeled 20CRv2 only uses this reanalysis.

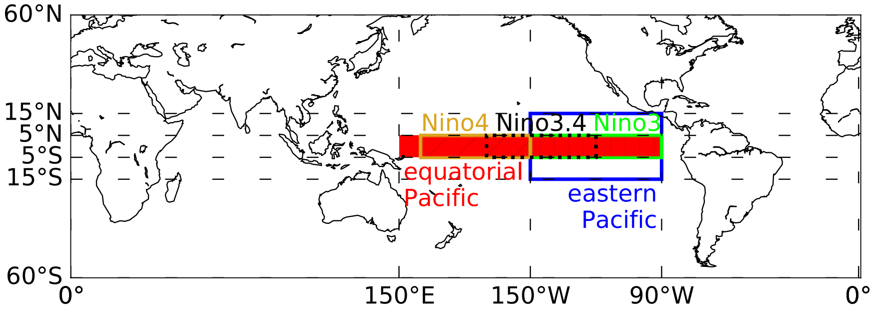

Fig. A1. Regions defined in the CLIVAR 2020 ENSO metrics package (see Table A2; the global60 region is the map displayed here).

We analyze 46 CMIP5 (Taylor et al. 2012) models and 42 CMIP6 (Eyring et al. 2016a) models, using simulations forced by estimates of historical atmospheric composition and land use (1850–2005 for CMIP5 and 1850–2014 for CMIP6). Only the first ensemble member of each model is used from their historical simulation. List of CMIP models and their acronyms are described in Table A3. The 32 historical simulations available for the IPSL-CM6A-LR model

Table A3. CMIP model acronyms and versions. A star (*) after a model name indicates that the model is part of the CMIP6 dataset; "atm." stands for atmosphere, "chem." for chemistry, "res." for resolution, and "stratos. dyn." for stratospheric dynamics.

Model names	Model versions
ACCESS: Australian Community Climate and Earth System Simulator	ACCESS1.0
	ACCESS1.3
	ACCESS-CM2*
	ACCESS-ESM1–5* (with carbon cycle)
BCC_CSM: Beijing Climate Center, Climate System Model	BCC_CSM1–1
	BCC_CSM1–1-M (medium res.)
	BCC_CSM2-MR* (medium res.)
BCC_ESM: Beijing Climate Center, Earth System Model	BCC_ESM1*
BNU-ESM: Beijing Normal University–Earth System Model	BNU-ESM
CAMS-CSM: Chinese Academy of Meteorological Sciences–Climate System Model	CAMS-CSM1–0*
CanCM: Canadian Coupled Global Climate Model	CanCM4

Table A3. (*Continued*).

Model names	Model versions
CanESM: Canadian Earth System Model	CanESM2 CanESM5* CanESM5-CanOE (with ocean biology)
CESM: Community Earth System Model [formerly Community Climate System Model (CCSM)]	CCSM4 CESM1(CAM5) CESM1(BGC) (with carbon cycle) CESM1(FASTCHEM) (with superfast chem.) CESM1(WACCM) (with stratos. dyn. and ozone chem.) CESM2* CESM2-FV2* (low res.) CESM2-WACCM* (with stratos. dyn. and ozone chem.) CESM2-FV2-WACCM* (low res., with stratos. dyn. and ozone chem.)
CMCC-CM: Centro Euro-Mediterraneo sui Cambiamenti Climatici Climate Model	CMCC-CM CMCC-CMS (with stratos. dyn.)
CMCC: Centro Euro-Mediterraneo sui Cambiamenti Climatici Earth System Model	CMCC-CESM
CNRM-CM: Centre National de Recherches Météorologiques Coupled Global Climate Model	CNRM-CM5 CNRM-CM5–2 CNRM-CM6–1* CNRM-CM6–1-HR* (high res.)
CNRM-ESM: Centre National de Recherches Météorologiques Earth System Model	CNRM-ESM2–1*
CSIRO: Commonwealth Scientific and Industrial Research Organization	CSIRO-Mk3.6.0 CSIRO-Mk3L-1–2 (low res.)
EC-Earth: European Consortium Earth System Model	EC-Earth3 EC-Earth3-Veg* (with dynamic vegetation)
FGOALS: Flexible Global Ocean–Atmosphere–Land System Model	FGOALS-f3-L* FGOALS-g2 FGOALS-s2 (with dynamic vegetation)
GFDL CM: Geophysical Fluid Dynamics Laboratory Climate Model	GFDL CM3 GFDL CM4*
GFDL-ESM: Geophysical Fluid Dynamics Laboratory Earth System Model	GFDL-ESM2G (GOLD ocean model) GFDL-ESM2M (MOM ocean model) GFDL-ESM4*
GISS: Goddard Institute for Space Studies	GISS-E2–1-G* (GISS ocean model) GISS-E2–1-G-CC* (GISS ocean model with carbon cycle) GISS-E2-H (HYCOM oce. model) GISS-E2-H-CC (HYCOM ocean model with carbon cycle) GISS-E2–1-H* (HYCOM ocean model) GISS-E2-R (Russell ocean model) GISS-E2-R-CC (Russell ocean model with carbon cycle)
HadGEM: Hadley Global Environment Model	HadCM3 HadGEM3-GC31-LL* (low res.) HadGEM2-CC (with carbon cycle)
HadGEM-ES: Hadley Global Environment Model Earth System Model	HadGEM2-ES

Table A3. (*Continued*).

Model names	Model versions
INMCM: Russian Institute for Numerical Mathematics Climate Model	INMCM4 INM-CM4–8* INM-CM5–0*
IPSL-CM: Institut Pierre Simon Laplace Climate Model	IPSL-CM5A-LR (low res.) IPSL-CM5A-MR (medium res.) IPSL-CM5B-LR (low res.) IPSL-CM6A-LR* (low res.)
KACE: Korean Advanced Climate Earth System Model	KACE-1–0-G*
MIROC: Model for Interdisciplinary Research on Climate	MIROC4h MIROC5 MIROC6*
MIROC-ESM: Model for Interdisciplinary Research on Climate, Earth System Model	MIROC-ESM MIROC-ESM-CHEM (with atm. chem.) MIROC-ES2L*
MPI-ESM: Max Planck Institute Earth System Model	MPI-ESM-LR (low res.) MPI-ESM-MR (medium res.) MPI-ESM-P (hybrid resolutions of LR and MR) MPI-ESM-1–2-HAM* (low res., with atm. chem.) MPI-ESM-1–2-HR* (high res.) MPI-ESM-1–2-LR* (low res.)
MRI-CGCM: Meteorological Research Institute Coupled General Circulation Model	MRI-CGCM3
MRI-ESM: Meteorological Research Institute Earth System Model	MRI-ESM1 MRI-ESM2–0*
NESM: Nanjing University of Information Science and Technology Earth System Model	NESM3*
NorCPM: Norwegian Climate Prediction Model (based of CESM)	NorCPM1*
NorESM: Norwegian Earth System Model (based of CESM)	NorESM1-M (medium res.) NorESM1-ME (medium res., with capability to be fully emission driven) NorESM2-LM* (low atm. res. and medium ocean res.) NorESM2-MM* (medium res.)
SAM: Seoul National University Atmosphere Model (based on CESM1)	SAM0-UNICON* (with a unified convection scheme)
TaiESM: Taiwan Earth System Model	TaiESM1*
UKESM: Met Office, Natural Environment Research Council (NERC) centers and U.K. universities Earth System Model	UKESM1–0-LL (low res.)

(Boucher et al. 2020) are also used to highlight the need to use large ensembles for certain metrics.

Appendix B: Methods

All the metrics presented here are calculated from monthly means. The monthly mean time series are linearly detrended, and anomalies are computed relative to the dataset's seasonal cycle. When necessary (e.g., to analyze spatial data), the data are interpolated onto a generic 1° latitude × 1° longitude grid. ENSO is defined based on SSTA during December (5-month triangular-weighted moving average), spatially averaged over the Niño-3.4 region (120°–170°W, 5°S–5°N; Fig. A1). The ENSO pattern (ENSO_pattern), life cycle (ENSO_life cycle), and duration

(ENSO_duration) metrics are based on linear regressions and the 5-month triangular-weighted moving average is applied to each field. The teleconnection metrics (DJF_PR_teleconnection, JJA_PR_teleconnection, DJF_TS_teleconnection, JJA_TS_teleconnection) are also based on linear regressions but using 3-month averaged anomalies. When individual ENSO events are detected, we define an El Niño event (the warm phase of ENSO) by the December value of smoothed (5-month triangular-weighted moving average) Niño-3.4 SSTA exceeding 0.75 standard deviation (STD; relative to the dataset). A La Niña event (the cold phase of ENSO) is defined by the December value of smoothed Niño-3.4 SSTA falling below -0.75 STD. All of

Table B1. Metrics defined in the CLIVAR 2020 ENSO metrics package. "abs. rel. diff." stands for absolute value of the relative difference (see sidebar "From a measured quantity to a portrait plot"), EN for El Niño and LN for La Niña, and "telecon." for teleconnection. Seasons abbreviations, MAM, JJA, NDJ, and DJF correspond to the seasons March–May, June–August, November–January, and December–February, respectively. See Table A1 for definitions of field acronyms and reference datasets, and Table A2 and Fig. A1 for the definition of the regions. Full details about each metric can be found here: https://github.com/CLIVAR-PRP/ENSO_metrics/wiki.

	Short name	Description	Algorithmic definition	Fields	Units
Climatology (light green)	double_ITCZ_bias	Eastern Pacific meridional PR bias ("double ITCZ bias")	RMSE	PR	mm day^{-1}
	eq_PR_bias	Equatorial Pacific zonal PR bias	RMSE	PR	mm day^{-1}
	eq_SST_bias	Equatorial Pacific zonal SST bias ("cold tongue bias")	RMSE	SST	°C
	eq_Taux_bias	Equatorial Pacific zonal Taux bias	RMSE	Taux	10^{-3} N m^{-2}
	double_ITCZ_sea_cycle	STD of eastern Pacific meridional PR seasonal cycle	RMSE	PR	mm day^{-1}
	eq_PR_sea_cycle	STD of equatorial Pacific zonal PR seasonal cycle	RMSE	PR	mm day^{-1}
	eq_SST_sea_cycle	STD of equatorial Pacific zonal SST seasonal cycle	RMSE	SST	°C
	eq_Taux_sea_cycle	STD of equatorial Pacific zonal Taux seasonal cycle	RMSE	Taux	10^{-3} N m^{-2}
Basic characteristics (light purple)	ENSO_pattern	ENSO pattern: equatorial Pacific zonal SSTA (regression)	RMSE	SST	°C °C^{-1}
	ENSO_life cycle	ENSO life cycle: SSTA evolution in Niño-3.4 (regression)	RMSE	SST	°C °C^{-1}
	ENSO_amplitude	ENSO amplitude: STD of Niño-3.4 SSTA	STD (abs. rel. diff.)	SST	%
	ENSO_seasonality	ENSO seasonal timing: STD of Niño-3.4 SSTA NDJ/MAM	STD (abs. rel. diff.)	SST	%
	ENSO_asymmerty	ENSO asymmetry: Skewness of Niño-3.4 SSTA	SKE (abs. rel. diff.)	SST	%
	ENSO_duration	ENSO event duration: Number of consecutive months with ENSO life cycle > 0.25 (Niño-3.4; based on regression)	Number of months (abs. rel. diff.)	SST	%
	ENSO_diversity	ENSO spatial pattern diversity: Interquartile range (IQR) of all EN's (LN's) zonal position of the maximum (minimum) SSTA	IQR (abs. rel. diff.)	SST	%
Telecon. (yellow)	DJF_PR_teleconnection	ENSO PR telecon. pattern in global60 during DJF (based on regression)	RMSE	PR, SST	mm day^{-1} °C^{-1}
	JJA_PR_teleconnection	ENSO PR telecon. pattern in global60 during JJA (based on regression)	RMSE	PR, SST	mm day^{-1} °C^{-1}
	DJF_TS_teleconnection	ENSO TS telecon. pattern in global60 during DJF (based on regression)	RMSE	SST, TS	°C °C^{-1}
	JJA_TS_teleconnection	ENSO TS telecon. pattern in global60 during JJA (based on regression)	RMSE	SST, TS	°C °C^{-1}
Processes (turquoise)	SST-Taux_feedback	Atmospheric Bjerknes feedback (Niño-3 SSTA, Niño-4 TauxA)	Slope (abs. rel. diff.)	SST, Taux	%
	Taux-SSH_feedback	Ocean–atmospheric Bjerknes feedback (Niño-4 TauxA, Niño-3 SSHA)	Slope (abs. rel. diff.)	SSH, Taux	%
	SSH-SST_feedback	Oceanic Bjerknes feedback in Niño-3 (SSH, SST)	Slope (abs. rel. diff.)	SSH, SST	%
	SST-NHF_feedback	Total heat flux feedback in Niño-3 (SST, NHF)	Slope (abs. rel. diff.)	SST, NHF	%
	ocean_driven_SST	Ocean-driven SST change in Niño-3 (EN and LN)	dSSToc = dSST – dSSTnhf (abs. rel. diff.)	SST, NHF	%

these processing choices are based upon well-established approaches in the large body of ENSO literature.

A brief description of each metric can be found in Table B1, and the CLIVAR 2020 ENSO metrics package is fully documented on its Github site (https://github.com/CLIVAR-PRP/ENSO_metrics/wiki).

In Figs. 3 and 7, a nonparametric Monte Carlo method is used to estimate the statistical significance: 10,000 random selections (with replacement) of any given sample are generated, providing the 2.5th and 97.5th percentiles of the distribution to obtain the 95% confidence level. For example, in Fig. 3 the confidence interval on CMIP5 MMV is computed with 10,000 random selections (with replacement) of 42 of the 46 CMIP5 models and that of CMIP6 MMV is computed similarly with 46 of the 42 CMIP6 models. The difference between CMIP5 and CMIP6 is considered significant when the CMIP5 MMV is not included in the 95% confidence interval of CMIP6 and the CMIP6 MMV is not included in the 95% confidence interval of CMIP5.

References

Adler, R. F., and Coauthors, 2003: The version-2 Global Precipitation Climatology Project (GPCP) monthly precipitation analysis (1979–present). *J. Hydrometeor.*, **4**, 1147–1167, https://doi.org/10.1175/1525-7541(2003)004<1147:TVGPCP>2.0.CO;2.

Ahn, M.-S., D. Kim, K. R. Sperber, I.-S. Kang, E. Maloney, D. Waliser, and H. Hendon, 2017: MJO simulation in CMIP5 climate models: MJO skill metrics and process-oriented diagnosis. *Climate Dyn.*, **49**, 4023–4045, https://doi.org/10.1007/s00382-017-3558-4.

An, S.-I., E. Tziperman, Y. Okumura, and T. Li, 2020: Irregularity and asymmetry. *El Niño Southern Oscillation in a Changing Climate*, *Geophys. Monogr.*, Vol. 252, Amer. Geophys. Union, 153–172, https://doi.org/10.1002/9781119548164.ch7.

Annamalai, H., 2020: ENSO Precipitation Anomalies along the Equatorial Pacific: Moist Static Energy Framework Diagnostics. *J. Climate*, **33**, 9103–9127, https://doi.org/10.1175/JCLI-D-19-0374.1.

——, P. Liu, and S.-P. Xie, 2005: Southwest Indian Ocean SST variability: Its local effect and remote influence on Asian monsoons. *J. Climate*, **18**, 4150–4167, https://doi.org/10.1175/JCLI3533.1.

Atwood, A. R., D. S. Battisti, A. T. Wittenberg, W. H. G. Roberts, and D. J. Vimont, 2017: Characterizing unforced multi-decadal variability of ENSO: A case study with the GFDL CM2.1 coupled GCM. *Climate Dyn.*, **49**, 2845–2862, https://doi.org/10.1007/s00382-016-3477-9.

Batehup, R., S. McGregor, and A. J. E. Gallant, 2015: The influence of non-stationary teleconnections on paleoclimate reconstructions of ENSO variance using a pseudoproxy framework. *Climate Past*, **11**, 1733–1749, https://doi.org/10.5194/cp-11-1733-2015.

Bayr, T., M. Latif, D. Dommenget, C. Wengel, J. Harlaß, and W. Park, 2018: Mean-state dependence of ENSO atmospheric feedbacks in climate models. *Climate Dyn.*, **50**, 3171–3194, https://doi.org/10.1007/s00382-017-3799-2.

——, C. Wengel, M. Latif, D. Dommenget, J. Lübbecke, and W. Park, 2019a: Error compensation of ENSO atmospheric feedbacks in climate models and its influence on simulated ENSO dynamics. *Climate Dyn.*, **53**, 155–172, https://doi.org/10.1007/s00382-018-4575-7.

——, D. I. V. Domeisen, and C. Wengel, 2019b: The effect of the equatorial Pacific cold SST bias on simulated ENSO teleconnections to the North Pacific and California. *Climate Dyn.*, **53**, 3771–3789, https://doi.org/10.1007/s00382-019-04746-9.

Bellenger, H., E. Guilyardi, J. Leloup, M. Lengaigne, and J. Vialard, 2014: ENSO representation in climate models: From CMIP3 to CMIP5. *Climate Dyn.*, **42**, 1999–2018, https://doi.org/10.1007/s00382-013-1783-z.

Bjerknes, J., 1969: Atmospheric teleconnections from the equatorial Pacific. *Mon. Wea. Rev.*, **97**, 163–172, https://doi.org/10.1175/1520-0493(1969)097<0163:ATFTEP>2.3.CO;2.

Boucher, O., and Coauthors, 2020: Presentation and evaluation of the IPSL-CM6A-LR climate model. *J. Adv. Model. Earth Sys.*, **12**, e2019MS002010, https://doi.org/10.1029/2019MS002010.

Brown, J. R., and Coauthors, 2020: Comparison of past and future simulations of ENSO in CMIP5/PMIP3 and CMIP6/PMIP4 models. *Climate Past*, **16**, 1777–1805, https://doi.org/10.5194/cp-16-1777-2020.

Cai, W., and Coauthors, 2014: Increasing frequency of extreme El Niño events due to greenhouse warming. *Nat. Climate Change*, **4**, 111–116, https://doi.org/10.1038/nclimate2100.

Capotondi, A., and Coauthors, 2015: Understanding ENSO diversity. *Bull. Amer. Meteor. Soc.*, **96**, 921–938, https://doi.org/10.1175/BAMS-D-13-00117.1.

——, A. T. Wittenberg, J.-S. Kug, K. Takahashi, and M. McPhaden, 2020: ENSO diversity. *El Niño Southern Oscillation in a Changing Climate*, *Geophys. Monogr.*, Vol. 252, Amer. Geophys. Union, 65–86, https://doi.org/10.1002/9781119548164.ch4.

Carré, M., J. P. Sachs, S. Purca, A. J. Schauer, P. Braconnot, R. A. Falcón, M. Julien, and D. Lavallée, 2014: Holocene history of ENSO variance and asymmetry in the eastern tropical Pacific. *Science*, **345**, 1045–1048, https://doi.org/10.1126/science.1252220.

Carton, J. A., G. A. Chepurin, and L. Chen, 2018: SODA3: A new ocean climate reanalysis. *J. Climate*, **31**, 6967–6983, https://doi.org/10.1175/JCLI-D-18-0149.1.

Chen, C., M. A. Cane, A. T. Wittenberg, and D. Chen, 2017: ENSO in the CMIP5 simulations: Life cycles, diversity, and responses to climate change. *J. Climate*, **30**, 775–801, https://doi.org/10.1175/JCLI-D-15-0901.1.

Compo, G. P., and Coauthors, 2011: The Twentieth Century Reanalysis project. *Quart. J. Roy. Meteor. Soc.*, **137**, 1–28, https://doi.org/10.1002/qj.776.

Dee, D. P., and Coauthors, 2011: The ERA-Interim reanalysis: Configuration and performance of the data assimilation system. *Quart. J. Roy. Meteor. Soc.*, **137**, 553–597, https://doi.org/10.1002/qj.828.

Deser, C., I. R. Simpson, K. A. McKinnon, and A. S. Phillips, 2017: The Northern Hemisphere extra-tropical atmospheric circulation response to ENSO: How well do we know it and how do we evaluate models accordingly? *J. Climate*, **30**, 5059–5082, https://doi.org/10.1175/JCLI-D-16-0844.1.

——, ——, A. S. Phillips, and K. A. McKinnon, 2018: How well do we know ENSO's climate impacts over North America, and how do we evaluate models accordingly? *J. Climate*, **31**, 4991–5014, https://doi.org/10.1175/JCLI-D-17-0783.1.

——, and Coauthors, 2020: Insights from Earth system model initial-condition large ensembles and future prospects. *Nat. Climate Change*, **10**, 277–286, https://doi.org/10.1038/s41558-020-0731-2.

Diaz, H. F., M. P. Hoerling, and J. K. Eischeid, 2001: ENSO variability, teleconnections and climate change. *Int. J. Climatol.*, **21**, 1845–1862, https://doi.org/10.1002/joc.631.

Ding, H., M. Newman, M. A. Alexander, A. T. Wittenberg, 2020: Relating CMIP5 model biases to seasonal forecast skill in the tropical Pacific. *Geophys. Res. Lett.*, **47**, e2019GL086765, https://doi.org/10.1029/2019GL086765.

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

——, and Coauthors, 2016b: Towards improved and more routine Earth system model evaluation in CMIP. *Earth Syst. Dyn.*, **7**, 813–830, https://doi.org/10.5194/esd-7-813-2016.

——, and Coauthors, 2016c: ESMValTool (v1.0)—A community diagnostic and performance metrics tool for routine evaluation of Earth system models in CMIP. *Geosci. Model Dev.*, **9**, 1747–1802, https://doi.org/10.5194/gmd-9-1747-2016.

Fedorov, A., S. Hu, A. T. Wittenberg, A. Levine, and C. Deser, 2020: ENSO low-frequency modulations and mean state interactions. *El Niño Southern Oscillation in a Changing Climate*, *Geophys. Monogr.*, Vol. 252, Amer. Geophys. Union, 173–198, https://doi.org/10.1002/9781119548164.ch8.

Flato, G., and Coauthors, 2013: Evaluation of climate models. *Climate Change 2013: The Physical Science Basis*, T. F. Stocker et al., Eds., Cambridge University Press, 741–866, https://doi.org/10.1017/CBO9781107415324.020.

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

——, C. Doutriaux, P. J. Durack, K. E. Taylor, Y. Zhang, D. N. Williams, E. Mason, and J. Servonnat, 2016: A more powerful reality test for climate models. *Eos, Trans. Amer. Geophys. Union*, **97**, https://doi.org/10.1029/2016EO051663.

Graham, F. S., A. T. Wittenberg, J. N. Brown, S. J. Marsland, and N. J. Holbrook, 2017: Understanding the double peaked El Niño in coupled GCMs. *Climate Dyn.*, **48**, 2045–2063, https://doi.org/10.1007/s00382-016-3189-1.

Guilyardi, E., 2006: El Niño-mean state-seasonal cycle interactions in a multi-model ensemble. *Climate Dyn.*, **26**, 329–348, https://doi.org/10.1007/s00382-005-0084-6.

——, A. T. Wittenberg, M. Balmesada, W. Cai, M. Collins, M. J. McPhaden, M. Watanabe, and S.-W. Yeh, 2016: Fourth CLIVAR workshop on the evaluation of ENSO processes in climate models: ENSO in a changing climate. *Bull. Amer. Meteor. Soc.*, **96**, 921–938, https://doi.org/10.1175/BAMS-D-15-00287.1.

——, A. Capotondi, M. Lengaigne, S. Thual, and A. T. Wittenberg, 2020: ENSO modelling: History, progress and challenges. *El Niño Southern Oscillation in a Changing Climate, Geophys. Monogr.*, Vol. 252, Amer. Geophys. Union, 199–226, https://doi.org/10.1002/9781119548164.ch9.

Hoerling, M. P., A. Kumar, and M. Zhong, 1997: El Niño, La Niña, and the nonlinearity of their teleconnections. *J. Climate*, **10**, 1769–1786, https://doi.org/10.1175/1520-0442(1997)010<1769:ENOLNA>2.0.CO;2.

Huang, B., and Coauthors, 2017: Extended Reconstructed Sea Surface Temperature, version 5 (ERSSTv5): Upgrades, validations, and intercomparisons. *J. Climate*, **30**, 8179–8205, https://doi.org/10.1175/JCLI-D-16-0836.1.

Im, S.-H., S.-I. An, S. T. Kim, and F.-F. Jin, 2015: Feedback processes responsible for El Niño-La Niña amplitude asymmetry. *Geophys. Res. Lett.*, **42**, 5556–5563, https://doi.org/10.1002/2015GL064853.

Izumo, T., J. Vialard, M. Lengaigne, I. Suresh, 2020: Relevance of relative sea surface temperature for tropical rainfall interannual variability. *Geophys. Res. Lett.*, **47**, e2019GL086182, https://doi.org/10.1029/2019GL086182.

Jin, F.-F., 1997: An equatorial ocean recharge paradigm for ENSO. Part I: Conceptual model. *J. Atmos. Sci.*, **54**, 811–829, https://doi.org/10.1175/1520-0469(1997)054<0811:AEORPF>2.0.CO;2.

——, S. T. Kim, and L. Bejarano, 2006: A coupled-stability index for ENSO. *Geophys. Res. Lett.*, **33**, L23708, https://doi.org/10.1029/2006GL027221.

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

Kessler, W. S., and Coauthors, 2019: Second report of TPOS 2020. GOOS Rep. 234, 265 pp., http://tpos2020.org/project-reports/second-report/.

Klein, S. A., B. J. Soden, and N. Lau, 1999: Remote sea surface temperature variations during ENSO: Evidence for a tropical atmospheric bridge. *J. Climate*, **12**, 917–932, https://doi.org/10.1175/1520-0442(1999)012<0917:RSSTVD>2.0.CO;2.

Lee, J., K. R. Sperber, P. J. Gleckler, C. J. W. Bonfils, and K. E. Taylor, 2019: Quantifying the agreement between observed and simulated extratropical modes of interannual variability. *Climate Dyn.*, **52**, 4057–4089, https://doi.org/10.1007/s00382-018-4355-4.

Lee, S.-K., P. N. DiNezio, E.-S. Chung, S.-W. Yeh, A. T. Wittenberg, and C. Wang, 2014: Spring persistence, transition and resurgence of El Niño. *Geophys. Res. Lett.*, **41**, 8578–8585, https://doi.org/10.1002/2014GL062484.

Li, J., and Coauthors, 2013: El Niño modulations over the past seven centuries. *Nat. Climate Change*, **3**, 822–826, https://doi.org/10.1038/nclimate1936.

Liu, Y., and Coauthors, 2017: Recent enhancement of central Pacific El Niño variability relative to last eight centuries. *Nat. Commun.*, **8**, 15386, https://doi.org/10.1038/ncomms15386.

Lloyd, J., E. Guilyardi, and H. Weller, 2012: The role of atmosphere feedbacks during ENSO in the CMIP3 models. Part III: The shortwave flux feedback. *J. Climate*, **25**, 4275–4293, https://doi.org/10.1175/JCLI-D-11-00178.1.

Maloney, E. D., and Coauthors, 2019: Process-oriented evaluation of climate and weather forecasting models. *Bull. Amer. Meteor. Soc.*, **100**, 1665–1686, https://doi.org/10.1175/BAMS-D-18-0042.1.

McGregor, S., A. Timmermann, M. H. England, O. Elison Timm, and A. T. Wittenberg, 2013: Inferred changes in El Niño–Southern Oscillation variance over the past six centuries. *Climate Past*, **9**, 2269–2284, https://doi.org/10.5194/cp-9-2269-2013.

McPhaden, M. J., S. E. Zebiak, and M. H. Glantz, 2006: ENSO as an integrating concept in Earth science. *Science*, **314**, 1740–1745, https://doi.org/10.1126/science.1132588.

Oueslati, B., and G. Bellon, 2015: The double ITCZ bias in CMIP5 models: Interaction between SST, large-scale circulation and precipitation. *Climate Dyn.*, **44**, 585–607, https://doi.org/10.1007/s00382-015-2468-6.

Perry, S. J., S. McGregor, A. Sen Gupta, and M. H. England, 2017: Future changes to El Niño–Southern Oscillation temperature and precipitation teleconnections. *Geophys. Res. Lett.*, **44**, 10 608–10 616, https://doi.org/10.1002/2017GL074509.

———, ———, ———, ———, and N. Maher, 2020: Projected late 21st century changes to the regional impacts of the El Niño-Southern Oscillation. *Climate Dyn.*, **54**, 395–412, https://doi.org/10.1007/s00382-019-05006-6.

Power, S. B., and F. P. Delage, 2018: El Niño–Southern Oscillation and associated climatic conditions around the world during the latter half of the twenty-first century. *J. Climate*, **31**, 6189–6207, https://doi.org/10.1175/JCLI-D-18-0138.1.

Praveen Kumar, B., J. Vialard, M. Lengaigne, V. S. N. Murty, and M. J. McPhaden, 2012: TropFlux: Air-sea fluxes for the global tropical oceans—Description and evaluation. *Climate Dyn.*, **38**, 1521–1543, https://doi.org/10.1007/s00382-011-1115-0.

———, ———, ———, ———, ———, M. F. Cronin, F. Pinsard, and K. Gopala Reddy, 2013: TropFlux wind stresses over the tropical oceans: Evaluation and comparison with other products. *Climate Dyn.*, **40**, 2049–2071, https://doi.org/10.1007/s00382-012-1455-4.

Ray, S., A. T. Wittenberg, S. M. Griffies, and F. Zeng, 2018a: Understanding the equatorial Pacific cold tongue time-mean heat budget. Part I: Diagnostic framework. *J. Climate*, **31**, 9965–9985, https://doi.org/10.1175/JCLI-D-18-0152.1.

———, ———, ———, and ———, 2018b: Understanding the equatorial Pacific cold tongue time-mean heat budget. Part II: Evaluation of the GFDL-FLOR coupled GCM. *J. Climate*, **31**, 9987–10 011, https://doi.org/10.1175/JCLI-D-18-0153.1.

Rayner, N. A., D. E. Parker, E. B. Horton, C. K. Folland, L. V. Alexander, D. P. Rowell, E. C. Kent, and A. Kaplan, 2003: Global analyses of sea surface temperature, sea ice, and night marine air temperature since the late nineteenth century. *J. Geophys. Res.*, **108**, 4407, https://doi.org/10.1029/2002JD002670.

Righi, M., and Coauthors, 2020: ESMValTool v2.0—Technical overview. *Geosci. Model Dev.*, **13**, 1179–1199, https://doi.org/10.5194/gmd-13-1179-2020.

Ropelewski, C. F., and M. S. Halpert, 1987: Global and regional scale precipitation patterns associated with the El Niño/Southern Oscillation. *Mon. Wea. Rev.*, **115**, 1606–1626, https://doi.org/10.1175/1520-0493(1987)115<1606:GARSPP>2.0.CO;2.

Saha, S., and Coauthors, 2006: The NCEP Climate Forecast System. *J. Climate*, **19**, 3483–3517, https://doi.org/10.1175/JCLI3812.1.

Santoso, A., M. J. McPhaden, and W. Cai, 2017: The defining characteristics of ENSO extremes and the strong 2015/2016 El Niño. *Rev. Geophys.*, **55**, 1079–1129, https://doi.org/10.1002/2017RG000560.

Sillmann, J., V. V. Kharin, X. Zhang, F. W. Zwiers, and D. Bronaugh, 2013: Climate extremes indices in the CMIP5 multimodel ensemble: Part I. Model evaluation in the present climate. *J. Geophys. Res. Atmos.*, **118**, 1716–1733, https://doi.org/10.1002/JGRD.50203.

Sperber, K. R., H. Annamalai, I.-S. Kang, A. Kitoh, A. Moise, A. Turner, B. Wang, and T. Zhou, 2013: The Asian summer monsoon: An intercomparison of CMIP5 vs. CMIP3 simulations of the late 20th century. *Climate Dyn.*, **41**, 2711–2744, https://doi.org/10.1007/s00382-012-1607-6.

Stoner, A. M., K. Hayhoe, and D. J. Wuebbles, 2009: Assessing general circulation model simulations of atmospheric teleconnection patterns. *J. Climate*, **22**, 4348–4372, https://doi.org/10.1175/2009JCLI2577.1.

Sun, D.-Z., Y. Yu, and T. Zhang, 2009: Tropical water vapor and cloud feedbacks in climate models: A further assessment using coupled simulations. *J. Climate*, **22**, 1287–1304, https://doi.org/10.1175/2008JCLI2267.1.

Taylor, K. E., R. J. Stouffer, and G. A. Meehl, 2012: An overview of CMIP5 and the experiment design. *Bull. Amer. Meteor. Soc.*, **93**, 485–498, https://doi.org/10.1175/BAMS-D-11-00094.1.

Timmermann, A., and Coauthors, 2018: El Niño–Southern Oscillation complexity. *Nature*, **559**, 535–545, https://doi.org/10.1038/s41586-018-0252-6.

Wang, B., and S.-I. An, 2002: A mechanism for decadal changes of ENSO behavior: Roles of background wind changes. *Climate Dyn.*, **18**, 475–486, https://doi.org/10.1007/s00382-001-0189-5.

Wang, C., 2002: Atmospheric circulation cells associated with the El Niño–Southern Oscillation. *J. Climate*, **15**, 399–419, https://doi.org/10.1175/1520-0442(2002)015<0399:ACCAWT>2.0.CO;2.

Wengel, C., M. Latif, W. Park, J. Harlaß, and T. Bayr, 2018: Seasonal ENSO phase locking in the Kiel Climate Model: The importance of the equatorial cold sea surface temperature bias. *Climate Dyn.*, **50**, 901–919, https://doi.org/10.1007/s00382-017-3648-3.

Wittenberg, A. T., 2004: Extended wind stress analyses for ENSO. *J. Climate*, **17**, 2526–2540, https://doi.org/10.1175/1520-0442(2004)017<2526:EWSAFE>2.0.CO;2.

———, 2009: Are historical records sufficient to constrain ENSO simulations? *Geophys. Res. Lett.*, **36**, L12702, https://doi.org/10.1029/2009GL038710.

———, A. Rosati, T. L. Delworth, G. A. Vecchi, and F. Zeng, 2014: ENSO modulation: Is it decadally predictable? *J. Climate*, **27**, 2667–2681, https://doi.org/10.1175/JCLI-D-13-00577.1.

Xie, P., and P. A. Arkin, 1997: Global precipitation: A 17-year monthly analysis based on gauge observations, satellite estimates, and numerical model outputs. *Bull. Amer. Meteor. Soc.*, **78**, 2539–2558, https://doi.org/10.1175/1520-0477(1997)078<2539:GPAYMA>2.0.CO;2.

Xie, S., H. Annamalai, F. A. Schott, and J. P. McCreary, 2002: Structure and mechanisms of south Indian Ocean climate variability. *J. Climate*, **15**, 864–878, https://doi.org/10.1175/1520-0442(2002)015<0864:SAMOSI>2.0.CO;2.

Yeh, S.-W., and B. P. Kirtman, 2004: Tropical Pacific decadal variability and ENSO amplitude modulation in a CGCM. *J. Geophys. Res.*, **109**, C11009, https://doi.org/10.1029/2004JC002442.

———, and Coauthors, 2018: ENSO atmospheric teleconnections and their response to greenhouse gas forcing. *Rev. Geophys.*, **56**, 185–206, https://doi.org/10.1002/2017RG000568.

Zebiak, S. E., and M. A. Cane, 1987: A model El Niño–Southern Oscillation. *Mon. Wea. Rev.*, **115**, 2262–2278, https://doi.org/10.1175/1520-0493(1987)115<2262:AMENO>2.0.CO;2.