Correspondence to:
C. Pickler,
carolyne.pickler@fau.de

General Circulation Model Selection Technique for Downscaling: Exemplary Application to East Africa

C. Pickler[1] and **T. Mölg[1]**

[1]Climate System Research Group, Institute of Geography, Friedrich-Alexander University Erlangen-Nürnberg (FAU), Erlangen, Germany

Abstract Downscaling is widely used in studies of local and/or regional climate as it yields a greater spatial resolution than general circulation models (GCMs) can provide. It utilizes GCM output or reanalysis data, which is transformed using mathematical relationships or used to force the lateral boundaries of a regional climate model. However, there is no set selection technique to determine which GCM realization(s) to employ. Here, a comprehensive yet easily applicable model selection technique for studies requiring GCM data as a constraint was developed. The technique evaluates, with respect to a reanalysis product and/or observational data, the ability of GCM realizations to reconstruct the mean state of the climate and the space-time climatic anomalies for the atmospheric state variables at three distinct pressure levels. It was applied to the region of East Africa, where GISS-E2-H r6i1p3 was found to perform the strongest. The top ranked realizations were found to better capture processes when evaluated for the example of the Indian Ocean Dipole. Furthermore, the surface air temperature and precipitation from three 10-year regional climate model simulations, one forced by the Modern-Era Retrospective Analysis for Research and Applications version 2 reanalysis, one forced by the top ranked GCM, and one by the lowest ranked one, were compared to gridded observations. Results show that using a top ranked GCM for the boundary conditions leads to a better dynamical downscaling simulation than a low-ranked GCM, suggesting the potential of the proposed technique for future downscaling techniques.

1. Introduction

General circulation models (GCMs), comprising atmosphere, land, and ocean components, are used to simulate large-scale climatic processes over various time periods. However, due to their coarse resolution, they cannot yield comprehensive simulations of climate at regional and local scales. Downscaling provides a great way to address this problem, since it leads to greater spatial resolution than GCMs can provide. It can be divided into two distinct groups: statistical and dynamical. Statistical downscaling involves establishing and using mathematical relationships between large-scale climate and regional or local climate to transform GCM output or reanalysis data to regional or local scale. Dynamical downscaling involves forcing the lateral boundaries of a regional (atmospheric) climate model (RCM) with reanalysis data or GCM output. This technique allows for local-scale processes (e.g., glacier mass balance, river/lake basin water levels) to be physically modeled from large-scale climate dynamics without circumventing any of the main scales (Giorgi, 2019; Giorgi & Bates, 1989).

Downscaling has been extensively used to understand regional and/or local climate. Mölg and Kaser (2011) used dynamical downscaling to quantify changes in glacier mass balance to large-scale circulation for the Kilimanjaro glaciers. Aalbers et al. (2018) elucidated whether changes in western European mean and daily precipitation were the result of internal variability or climate change using regional downscaling of a 16 member initial-condition RCM-GCM ensemble. The Alamo Lake and Bill Williams River basin (United States) were studied using statistical and dynamical downscaling to examine how they would be influenced by future climate scenarios (Shamir et al., 2019). Furthermore, downscaling is also used for vulnerability, impacts, and adaptation assessments, including studies of vector borne diseases (Rodó et al., 2013), pollution-related mortality (J. Sun et al., 2015), and heat stress (Im et al., 2018). Additional examples and a thorough overview of downscaling can be found in Giorgi (2019). It should be noted that significant advancements in the field of regional climate modeling and downscaling can be attributed to the Coordinated Regional climate Downscaling Experiment, which was established by the World Climate Research Programme (WCRP) and its Task Force for Regional Climate Downscaling in 2009. It has provided numerous

regional downscaling products (e.g., Jacob et al., 2014; Mearns et al., 2012) and assessments of downscaling models and techniques (e.g., Akinsanola & Ogunjobi, 2017; Legasa et al., 2020). However, while significant work has been undertaken, some uncertainties still require attention, including those associated with the selection of GCM(s) used. This has been shown to be an important source of uncertainty in downscaling and climate change impact studies, especially on multidecadal time scales (Giorgi & Francisco, 2000; Hawkins & Sutton, 2009; Jobst et al., 2018; Lutz et al., 2016).

This leads to the following important question. How do we select GCMs for downscaling? Studies have been undertaken to create a GCM selection criteria. The majority of these have focused on past and/or present performance. Pierce et al. (2009) presented a GCM selection technique based on 42 performance metrics, including seasonal precipitation, temperature, El Niño-Southern Oscillation (ENSO) and Pacific Decadal Oscillation, for a case study of hydrological change in the Western United States. Overland et al. (2011), using an Arctic case study, presented a two-tier approach to gauge model performance. First, evaluating the large-scale climate physics of the models by examining how well they capture continental-scale processes. Second, assessing model performance on a regional scale by analyzing variables of ecological or societal relevance. McSweeney et al. (2012) built upon this work by stressing the importance of model evaluation. They proposed determining the ability of the model to reproduce the regional climate and then culling the remaining models but ensuring that all plausible projections were retained. In 2015, this work was expanded upon by the creation of a selection criteria for multiple regions (McSweeney et al., 2015). The ability of the model to simulate the regional climate of the regions of interest was evaluated (e.g., annual temperature and precipitation, teleconnections, circulation, storm tracks, and monsoon season) and the least realistic models for each region were discarded. Lutz et al. (2016) combined this past performance approach with an envelope-based technique to select suitable GCMs and applied this to the river basins of the Indus, Ganges, and Brahmaputra. All these studies emphasized the superiority of a multimodel ensemble for diagnostic regional climate change studies. These past studies also placed emphasis on the surface variables of temperature and precipitation. Atmospheric state variables (winds, humidity, and temperature) at varying pressure levels were not the main focus. These can, nonetheless, provide insight into atmospheric variability and, most importantly, are used to force the lateral boundaries of RCMs. Previous research focused mainly on capturing synoptic-scale trends and gave little attention to more local ones, which play an important role in downscaling. Overall, these different selection techniques highlight the absence of a straightforward, more generic one due to the complex nature and wide range of goals in downscaling studies.

To address this issue, we will focus on model selection for a case study of East Africa and the high-elevation climate on Kilimanjaro, Tanzania. This area is of interest as African climate suffers from a lack of research attention (Washington et al., 2006) and is highly vulnerable to climate change (Boko et al., 2007; Niang et al., 2014). Furthermore, high-altitude regions, like Kilimanjaro, are topographically complex and particularly sensitive to climate change (e.g., Ohmura, 2012; Pepin et al., 2015). The Kilimanjaro glaciers provide a fitting case study as they allow for the capture of large-scale and local-scale climate change dynamics and examination of how these scales are linked by the mesoscale atmospheric circulation over the mountain (Mölg & Kaser, 2011; Mölg et al., 2006, 2009, 2012). They have dramatically retreated over the twentieth century (N. Cullen et al., 2013) and been extensively studied on a local scale (e.g., N. J. Cullen et al., 2007; Mölg & Hardy, 2004; Mölg et al., 2008, 2009) and using regional atmospheric modeling (e.g., Collier et al., 2018; Mölg & Kaser, 2011; Mölg et al., 2012). Their large-scale influences have also been studied (Collier et al., 2018; Endris et al., 2016, 2019; Mölg et al., 2006; Thielke & Mölg, 2019) and multiannual automated weather station (AWS) records are available for model evaluation. With both the local, regional, and macroscales in the area being evidently affected by climate change, it provides an ideal location for downscaling.

In this study, we use past performance model evaluation to determine how well GCMs capture the local and regional climate of East Africa and the midtroposphere layers around Kilimanjaro. The goal is to provide a useful selection criterion that can be used for downscaling studies and that is, moreover, easily applicable to other regions. In Section 2, we describe the study area, the GCM data, and the regional and local data used for model evaluation. In Sections 3 and 4, we present the model selection procedure and its results. In Sections 5 and 6, we discuss the potential benefits of the model rankings and provide concluding remarks and future outlook.
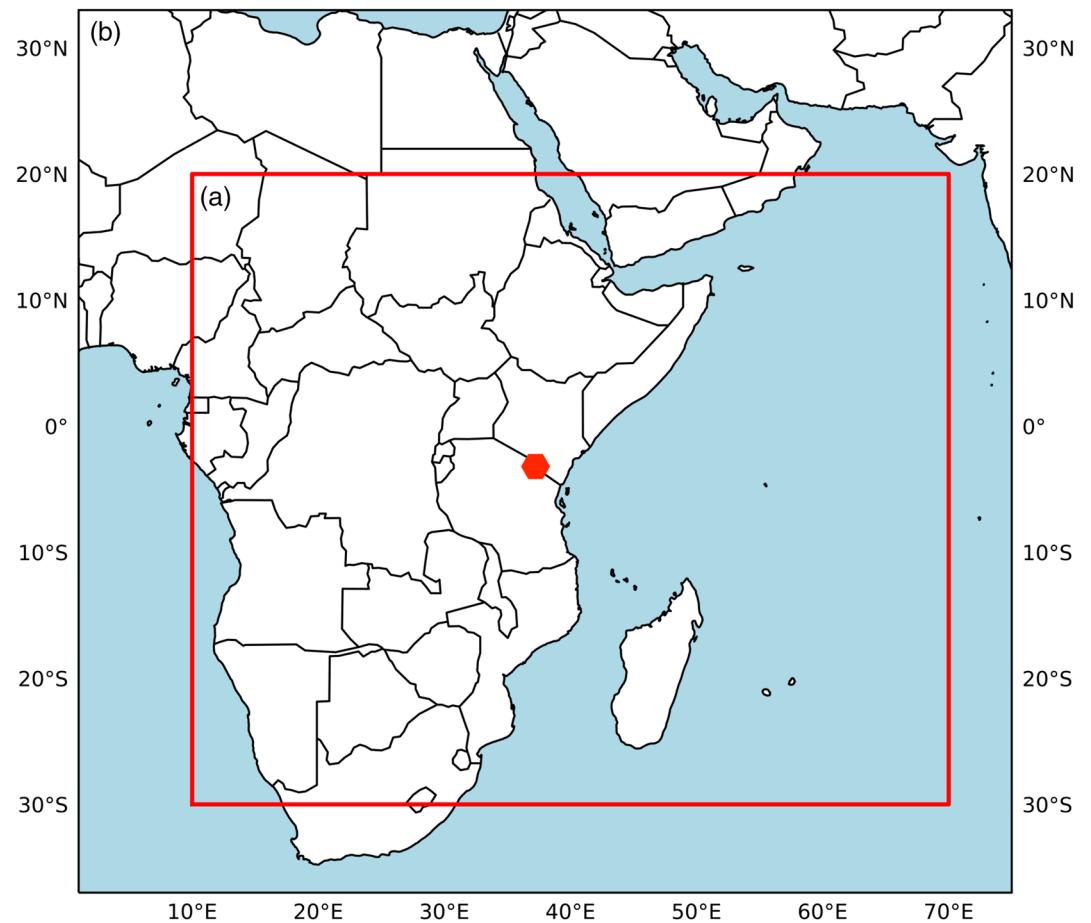
Figure 1. (a) Domain (red box) of model selection case study, where the red hexagon indicates the location of Kilimanjaro and (b) the WRF domain centered over Kilimanjaro used for each case presented in the dynamical downscaling test (Section 5.3). This domain coincides with that used in Collier et al. (2019). WRF, Weather and Research Forecasting.

2. Study Region and Data

The model selection procedure is illustrated for a case study covering 20°N–30°S by 10°–70°E, comprising East Africa, parts of central and southern Africa and the western Indian Ocean (Figure 1a). This large region covers the typical area that would be considered in regional atmospheric and climate modeling of East Africa (Collier et al., 2018; Mölg & Kaser, 2011; Mölg et al., 2012; L. Sun et al., 1999). From hereafter, this will be referred to as the African domain. A bimodal rainfall regime characterizes East Africa with the so called "short rains" in October–December (OND) and "long rains" in March–May (MAM). This rainfall regime is typically associated with the north-south migration of the intertropical convergence zone. Topography and the bodies of water in the region influence the spatial distribution of the rainfall, while the temporal variability is linked to atmosphere-ocean modes, namely ENSO (e.g., Nicholson, 2015; Pohl & Camberlin, 2011) and Indian Ocean Dipole (IOD) (e.g., Clark et al., 2003; Nicholson, 2015). Data for this region came from the Modern-Era Retrospective Analysis for Research and Applications version 2 (MERRA-2) reanalysis data set (Gelaro et al., 2017). Monthly means of temperature, specific humidity, and zonal (U) and meridional winds (V) from 6-hourly data at 200, 500, and 850 hPa at a resolution of 0.5° latitude × 0.625° longitude for January 1980 to December 2012 were utilized. The choice is based on a recent study that found MERRA-2 to be a suitable data set for representing important climatic features of East Africa (Collier et al., 2018).

Kilimanjaro, Africa's highest mountain, lies at the center of this region (Figure 1). It is a massive strato-volcano which formed in connection to the East African rift system. It is located on the Kenya-Tanzania border (3.076°S, 37.353°E) and is composed of three peaks: Shira (4,005 m above mean sea level [MSL]),

Mawenzi (5,140 MSL), and Kibo (5,895 MSL). Modern glaciation is confined to Kibo where a total ice cover loss of 85% from 1912 to 2011 occurred (N. Cullen et al., 2013). These glaciers and their dependencies on, and signals of, mesoscale atmospheric and large-scale climate dynamics have been extensively studied for more than a decade (e.g., Collier et al., 2018; N. J. Cullen et al., 2007; Kaser et al., 2004; Mölg & Hardy, 2004; Mölg & Kaser, 2011; Mölg et al., 2008, 2009). Four AWSs have been installed on the peak. AWS 1 (3.067°S, 37.350°E) is located on the Northern Ice Field at 5,794 MSL and has been operational since March 2000 (Hardy, 2011). In February 2005, AWS 2 and AWS 3 were installed. AWS 2 (3.060°S, 37.349°E) is near the base of a vertical ice cliff face on the Northern Ice Field at 5,763 MSL (Winkler et al., 2010). AWS 3 (3.078°S, 37.354°E) is located on the upper slopes of Kersten Glacier in the Southern Ice Field at 5,873 MSL and has been the subject of significant research (e.g., Collier et al., 2018; Mölg & Kaser, 2011; Mölg et al., 2008, 2009, 2012). AWS 4 (3.081°S, 37.352°E) was installed in January 2009 on the midslopes of the Kersten Glacier at 5,603 MSL (Mölg et al., 2020). For this study, only AWS 1 and AWS 3 are suitable, since their records are sufficiently long and they are not influenced by a special (ice cliff) microclimate like AWS 2. Consideration of AWS 1 and AWS 3 data for the present GCM context is only possible due to their location on the peak of an isolated mountain at almost exactly 500 hPa, where local conditions (of daily or longer means) mostly reflect the large-scale properties of the tropical midtroposphere (Mölg et al., 2009; Pepin et al., 2010). Records for AWS 1 and AWS 3 span March 2000 to September 2017 and February 2005 to December 2013, respectively. The monthly means of the air temperature, specific humidity, and zonal and meridional wind measurements were considered. Details concerning the sensors and data processing can be found in Mölg et al. (2008, 2009) and Hardy (2011).

GCMs were accessed from the CMIP5 repository (https://esgf-node.llnl.gov/projects/esgf-llnl/) in September 2018 (Taylor et al., 2012). Each realization was independently analyzed for GCMs with more than one initial-condition ensemble member (realization) as they represent a possible outcome with slightly differing initial conditions and timing of events. The atmospheric state variables (U, V, specific humidity, and temperature) were the focus of the selection procedure due to their importance for downscaling studies. Models were culled depending on availability of the monthly means of these variables at 200, 500, and 850 hPa for the historical simulation, a simulation of the recent past (1850–2005) that is consistent with the instrumental era and includes natural and anthropogenic forcings. This left 50 models and 213 realizations. Moreover, since the aim is to provide a selection technique for downscaling, models were culled depending on whether they provided the boundary conditions required for dynamical downscaling (6 hourly) for the historical simulation and, at least, one representative concentration pathway (RCP). This lead to the elimination of 24 GCMs and a total of 180 GCM realizations. The remaining 26 GCMs and their total 33 realizations were subjected to the model selection procedure and can be found in Table 1.

3. Model Selection

Building upon the work of McSweeney et al. (2012, 2015) and Overland et al. (2011), a three-tier evaluation strategy was applied to the realizations in Table 1: (1) whether the global warming signal was captured, (2) how well the tropospheric circulation of the African domain was simulated, and (3) how well the climate of the air layers sampled at the summit of Kilimanjaro was modeled. The first tier is a precondition that evaluates the robustness of the GCM on the global scale, while the following two tiers analyze its ability to reproduce the climate of the region. A schematic of this can be found in Figure 2, which will be explained in Sections 3.1–3.3.

3.1. Global Warming Signal

First, as a precondition to the selection procedure, the robustness of each GCM realization was evaluated by examining their ability to reconstruct the global warming signal. If a model cannot capture this global signal, how can it be considered robust? Studies of global temperature evolution over 1979–2005 using observational data sets and global reanalysis products have shown a warming trend (Foster & Rahmstorf, 2011; Simmons et al., 2017). Using the historical simulation, the ability of realizations to capture this tendency was tested by determining their global temperature trend over 1979–2005. If a realization can simulate a positive temperature trend over this period, then they are deemed to meet the global warming criteria and will continue onto the selection procedure.

Table 1
List of Retained CMIP5 GCMs and Their Respective Realizations, Where, Unless Specified Otherwise, the Realization (r) Has Initialization 1 (i1) and Physics 1 (p1)

Model	Realization	Historical extended to 2012	Resolution	Reference
Access 1.0	r1	–	1.25° × 1.875°	Bi et al. (2013); Dix et al. (2013)
Access 1.3	r1	–	1.25° × 1.875°	Bi et al. (2013); Dix et al. (2013)
BCC-CSM 1.1	r1	X	2.7906° × 2.8125°	Wu et al. (2014)
BCC-CSM 1.1m	r1	X	2.7906° × 2.8125°	Wu et al. (2014)
BNU-ESM	r1	–	2.7906° × 2.8125°	Ji et al. (2014)
CanESM2	r1	X	2.7906° × 2.8125°	Arora et al. (2011); von Salzen et al. (2013)
CCSM4	r6	–	0.9424° × 1.25°	Gent et al. (2011)
CMCC-CM	r1	–	0.7487° × 0.75°	Scoccimarro et al. (2011)
CNRM-CM5	r1	X	1.4008° × 1.40625°	Voldoire et al. (2013)
CSIRO-Mk3.6.0	r1	–	1.8653° × 1.875°	Jeffrey et al. (2013)
FGOALS-g2	r1	–	2.7906° × 2.8125°	Li et al. (2013)
GFDL-CM3	r1, r3, r5	–	2° × 2.5°	Donner et al. (2011); Griffies et al. (2011)
GFDL-ESM2M	r1	–	2.0225° × 2.5°	Dunne et al. (2012, 2013)
GFDL-ESM2G	r1	–	2.0225° × 2°	Dunne et al. (2012, 2013)
GISS-E2-H	r6p3	X	2° × 2.5°	Schmidt et al. (2014)
GISS-E2-R	r6, r6p3	X (only r6p3)	2° × 2.5°	Schmidt et al. (2014)
HadGEM2-ES	r1, r2	X (only r2)*	1.25° × 1.875°	Collins et al. (2011); Martin et al. (2011)
IPSL-CM5A-LR	r1, r2, r3, r4	–	1.8947° × 3.75°	Dufresne et al. (2013)
IPSL-CM5A-MR	r1	–	1.2676° × 2.5°	Dufresne et al. (2013)
IPSL-CM5B-LR	r1	–	1.8947° × 3.75°	Dufresne et al. (2013)
MIROC-ESM	r1	–	2.7906° × 2.8125°	S. Watanabe et al. (2011)
MIROC-ESM-CHEM	r1	X	2.7906° × 2.8125°	S. Watanabe et al. (2011)
MIROC4h	r1	–	0.5616° × 0.5627°	Sakamoto et al. (2012)
MIROC5	r1	X	1.4008° × 1.40625°	M. Watanabe et al. (2010)
MRI-CGCM3	r1	X	1.12148° × 1.125°	Yukimoto et al. (2012)
NorESM1-M	r1	X	1.8947° × 2.5°	Bentsen et al. (2013); Iversen et al. (2013)

Note. Resolution represents the horizontal atmospheric grid resolution and is expressed as latitude by longitude. A X stipulates that the historical simulation has been extended until 2012, while * indicates that it has been extended to 11/2019.

3.2. African Domain

Past performance of the GCM realizations for the African domain was evaluated seasonally (dry: January–February [JF] and June–September [JJAS]; wet: MAM and OND) by comparison to MERRA-2 reanalysis data for the region. Performance was assessed for the historical simulation over the period of 1980–2012. The historical simulation finishes in 2005 and was extended to 2012 using the historicalExt simulation, a simulation continuing the historical simulation. If the historicalExt simulation was not available, the historical simulation was extended using RCP 4.5, as it was the sole RCP available for all the GCM realizations being subjected to the selection procedure (Table 1). Furthermore, it should be noted that all RCP scenarios are almost the same over this period and only begin to diverge at around 2025; therefore, the choice of which RCP scenario used to extend the simulations should not influence the results (Maule et al., 2017). Evaluation of each of the atmospheric state variables (temperature, specific humidity, and U and V winds) was undertaken at 200, 500, and 850 hPa. These variables represent the driving data for dynamical downscaling and studying them at 200, 500, and 850 hPa allows for examination of their performance within the troposphere. They were regridded using bilinear interpolation to match the resolution of the MERRA-2
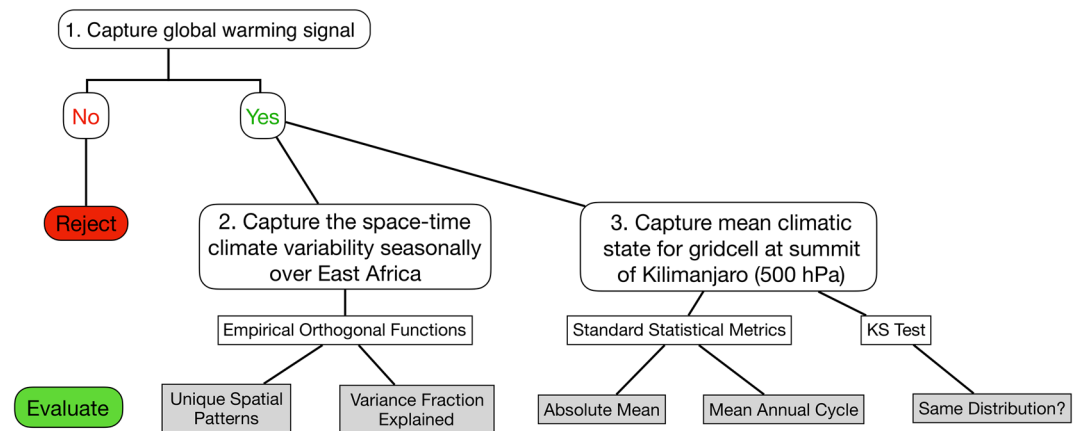
Figure 2. Schematic illustrating model selection procedure, where KS test is the Kolmogorov-Smirnov test. Details concerning each test can be found in Sections 3.1–3.3.

reanalysis data (0.5° latitude × 0.625° longitude) and time means were taken over each season for each year of the 33-year study period. Analysis of past performance focused on the dominant spatial patterns of variability for each season. Empirical orthogonal function (EOF) analysis was used to extract these spatial patterns, as this approach has been well established (Hannachi et al., 2007; Wilks, 2006). Degenerate EOFs were eliminated using North's Rule of Thumb (North et al., 1982) since nonunique spatial patterns cannot be compared. The nondegenerate EOFs and their variance fraction explained with its 95% confidence interval of MERRA-2 were calculated and can be found in Table A1. For the 33 retained GCM historical simulations, the EOFs and their variance fraction explained with its 95% confidence interval corresponding to each of the MERRA-2 nondegenerate EOFs were calculated. For each variable, EOF 1 of MERRA-2 and its variance fraction explained, if nondegenerate, was compared with that of the GCM realization and so on, until all nondegenerate EOFs of MERRA-2 were analyzed. This strict criterion was undertaken, even though the realizations may have inverted the dominant spatial patterns (e.g., the dominant spatial pattern of MERRA-2 is represented by that of EOF 2 in the GCM realization), as it reflects a short-coming the realization has in reproducing the dominant spatial pattern in the first place. To evaluate how well the realization captured the variance fraction explained, the relative error with respect to MERRA-2 was calculated (see Section 3.4). Furthermore, the dominant spatial patterns of variability found in the nondegenerate EOFs from the MERRA-2 reanalysis data were correlated with the respective EOF patterns for each realization using the Pearson correlation. The relative error of these correlations was determined and then combined with that from the associated variance fraction explained, giving each equal weighting (50%), providing one metric evaluating how well the tropospheric circulation of the African domain was simulated. These metrics were combined as we want to determine how well the realization captures the spatial pattern and its associated variance fraction explained for the nondegenerate EOF of interest. This means that we do not want to rank a realization higher if it attributes the correct variance fraction explained but for an incorrect spatial pattern.

3.3. Kilimanjaro

The 33 GCM historical simulations were also compared with observations from AWS 1 and AWS 3 at the summit of Kilimanjaro to determine their ability to reconstruct the large-scale climatic properties of the tropical midtroposphere. Evaluation focused on reconstructing the mean state of the climate system at this location, since models typically have more problems capturing the mean state than the anomalies thereof (Bindoff et al., 2013). With respect to all the model resolutions, AWS 1 and AWS 3 are located in the same grid cell at 500 hPa (mean air pressure at AWS 3, e.g., is 502 hPa). Hence, their measurements were each averaged together, leading to a monthly observational record spanning March 2000 to September 2017. Using this as the assessment period, the realizations were extended to September 2017 using the historicalExt simulation, when available (Table 1), and RCP 4.5. The atmospheric state variables from the GCM historical simulations at 500 hPa for the grid cell at the peak (with regard to the native spatial resolution of the GCM)

were individually evaluated against the observational record for three distinct metrics. First, their ability to reproduce the overall mean state of temperature, specific humidity, and U and V winds was tested. The mean state of the observational record and realizations were calculated, along with their standard deviation with 2σ, for each variable for the entire comparison period. This standard deviation will be referred to as the 95% confidence interval. The absolute mean difference of the realizations to observations was utilized to analyze how well the mean state was reproduced. Second, the nonparametric Kolmogorov-Smirnov (KS) test was then run for each variable and realization to determine whether they came from the same distribution as the observations. Here, realizations are evaluated with respect to their KS statistic (D value) and the null hypothesis that the observations and GCM realization are drawn from the same source. Third, the mean annual cycle of the observations and realizations was compared by calculating the Pearson correlation and the root mean square error (RMSE). To further quantify and rank how each realization performed during these tests, the relative error was calculated.

3.4. Testing Synopsis

The tests performed for model selection evaluated how well the realizations captured the mean climatic state and simulated the anomalies of the space-time patterns. Apart from meeting the precondition of capturing the global warming signal, no one metric was weighted above another as each examines how a realization simulates the climate of the African domain and the large-scale climatic properties of the tropical midtroposphere in the region. Furthermore, to maintain the transferability of the procedure, no preset thresholds or cutoffs were used. These can be selected by the user since they are dependent on the problem being addressed.

The relative error for each metric provided a good tool to evaluate realization performance but is not indicative of performance in absolute terms. It was calculated using the technique outline in Rupp et al. (2013) and resulted in a normalized value ranging between 0 and 1, where 0 represents the best fit and 1 the worst. For each realization, the root mean square (RMS) of the relative errors from each individual test was calculated:

$$\text{RMS} = \sqrt{\frac{1}{n}\sum_{i=1}(\text{modelRE}_i)^2} \tag{1}$$

In this equation, modelRE is the relative error of the model for the given metric (value between 0 and 1) and n is the number of metrics (73). It should be noted that in our case, RMS and RMSE are equivalent as the best case relative error is always 0. This ranking method will highlight which GCM realization performs best on average rather than just which GCM has the lowest overall relative error.

4. Model Selection Results

All realizations were found to capture the global warming signal and were subjected to the model selection procedure. This resulted in the realization rankings presented in Table 2 and Figure 3. No GCM realization markedly performed best in every test. However, GISS-E2-H r6i1p3 performs the strongest, showing the lowest total relative error and RMS. The two realizations of GISS-E2-R (r6i1p1, r6i1p3), which differ in ocean model from GISS-E2-H, also rank highly in tenth and twelfth place. In the following subsections, how well each realization performed for each variable and metric is examined. Due to a large quantity of data material, only main characteristics are highlighted.

4.1. Temperature

Some distinct tendencies are observed in the model selection results for temperature. GCM realizations performed well when examining the variance fraction metric for EOF 1. For all pressure levels and seasons, the vast majority (≥95%) of the variance fraction explained from the realizations are found within the confidence interval of the respective MERRA-2 variance fraction. This means that the 95% confidence interval of the GCM realization and that of MERRA-2 overlap. This can be observed for JJAS in Figure 4. The realizations encountered difficulties with respect to EOF 2–EOF 4 at 200 and 500 hPa as the amount of

Table 2
RMS of the Relative Errors for GCM Realizations Subjected to the Model Selection Procedure, Displayed From Highest Ranked to Lowest Ranked

	RMS_Obs	RMS_EOF_Temp	RMS_EOF_hus	RMS_EOF_U	RMS_EOF_V	RMS_tot
Model	12	32	13	14	2	73
GISS-E2-H r6i1p3	0.47 (20)	0.35 (2)	0.39 (5)	0.43 (12)	0.52 (16)	0.40
NorESM1-M r1i1p1	0.43 (12)	0.37 (5)	0.50 (21)	0.40 (7)	0.54 (20)	0.42
BCC-CSM1-1 r1i1p1	0.40 (7)	0.37 (4)	0.49 (18)	0.47 (16)	0.56 (23)	0.42
CCSM4 r6i1p1	0.44 (15)	0.39 (9)	0.50 (20)	0.42 (10)	0.39 (6)	0.43
IPSL-CM5B-LR r1i1p1	0.45 (16)	0.40 (11)	0.51 (25)	0.42 (8)	0.33 (5)	0.43
MRI-CGCM3 r1i1p1	0.40 (8)	0.34 (1)	0.60 (32)	0.43 (11)	0.69 (28)	0.44
GFDL-CM3 r3i1p1	0.34 (2)	0.51 (23)	0.47 (13)	0.30 (1)	0.48 (10)	0.44
BNU-ESM r1i1p1	0.42 (11)	0.47 (18)	0.49 (17)	0.36 (4)	0.27 (1)	0.44
ACCESS1-0 r1i1p1	0.54 (29)	0.40 (12)	0.46 (12)	0.44 (13)	0.32 (3)	0.44
GISS-E2-R r6i1p1	0.52 (27)	0.36 (3)	0.47 (15)	0.53 (24)	0.50 (15)	0.45
GFDL-ESM2M r1i1p1	0.45 (17)	0.47 (19)	0.45 (11)	0.35 (3)	0.72 (29)	0.45
GISS-E2-R r6i1p3	0.58 (32)	0.40 (10)	0.32 (1)	0.54 (27)	0.61 (24)	0.46
MIROC-ESM-CHEM r1i1p1	0.38 (6)	0.46 (17)	0.35 (3)	0.60 (31)	0.49 (11)	0.46
MIROC5 r1i1p1	0.45 (18)	0.48 (21)	0.40 (6)	0.47 (17)	0.50 (13)	0.46
CMCC-CM r1i1p1	0.51 (26)	0.38 (7)	0.51 (23)	0.50 (20)	0.79 (33)	0.46
CNRM-CM5 r1i1p1	0.53 (28)	0.39 (8)	0.48 (16)	0.60 (32)	0.33 (4)	0.48
HadGEM2-ES r1i1p1	0.46 (19)	0.41 (13)	0.57 (29)	0.56 (29)	0.40 (7)	0.48
CSIRO-Mk3-6-0 r1i1p1	0.48 (22)	0.41 (14)	0.54 (27)	0.54 (26)	0.54 (21)	0.48
GFDL-CM3 r5i1p1	0.41 (10)	0.49 (22)	0.51 (22)	0.48 (18)	0.50 (14)	0.48
FGOALS-g2 r1i1p1	0.55 (30)	0.45 (15)	0.57 (30)	0.42 (9)	0.40 (8)	0.48
MIROC-ESM r1i1p1	0.41 (9)	0.48 (20)	0.38 (4)	0.63 (33)	0.52 (18)	0.49
HadGEM2-ES r2i1p1	0.49 (25)	0.45 (16)	0.55 (28)	0.54 (28)	0.32 (2)	0.49
BCC-CSM1-1-m r1i1p1	0.48 (21)	0.55 (26)	0.44 (9)	0.38 (6)	0.75 (31)	0.50
ACCESS1-3 r1i1p1	0.49 (23)	0.37 (6)	0.66 (33)	0.57 (30)	0.53 (19)	0.50
IPSL-CM5A-LR r3i1p1	0.38 (5)	0.55 (27)	0.41 (7)	0.50 (21)	0.76 (32)	0.50
GFDL-ESM2G r1i1p1	0.56 (31)	0.53 (24)	0.51 (24)	0.34 (2)	0.64 (25)	0.50
CanESM2 r1i1p1	0.59 (33)	0.55 (25)	0.52 (26)	0.37 (5)	0.41 (9)	0.52
GFDL-CM3 r1i1p1	0.38 (4)	0.58 (28)	0.60 (31)	0.45 (14)	0.55 (22)	0.53
IPSL-CM5A-LR r2i1p1	0.44 (14)	0.60 (29)	0.47 (14)	0.47 (15)	0.66 (26)	0.53
IPSL-CM5A-LR r1i1p1	0.33 (1)	0.66 (33)	0.34 (2)	0.51 (22)	0.52 (17)	0.54
IPSL-CM5A-MR r1i1p1	0.43 (13)	0.61 (30)	0.44 (8)	0.49 (19)	0.73 (30)	0.54
MIROC4h r1i1p1	0.49 (24)	0.61 (31)	0.45 (10)	0.52 (23)	0.50 (12)	0.55
IPSL-CM5A-LR r4i1p1	0.35 (3)	0.64 (32)	0.49 (19)	0.53 (25)	0.66 (27)	0.56

Note. RMS_Obs is the RMS for the mean climatic state analysis, RMS_EOF_Temp, RMS_EOF_hus, RMS_EOF_U, and RMS_EOF_V are the RMS for the seasonal space-time anomaly tests at 200, 500, and 850 hPa for each respective atmospheric state variable. RMS_tot is the RMS for all tests. The number located below each RMS indicates the number of metrics contained within. The number within the parentheses represents the ranking for the metrics being considered.

realizations within the MERRA-2 confidence interval decreased significantly for the variance fraction explained. With the exception of EOF 2 in JF and OND at 200 hPa, EOF 2 in MAM at 500 hPa, and EOF 4 in JF at 500 hPa, only about half of the realizations fall within the confidence interval of the respective MERRA-2 value. This trend did not hold true at 850 hPa, where the variance fraction explained of the realizations for all the nondegenerate EOFs were always in good agreement with MERRA-2 (≥89% of realizations within its confidence interval). Different tendencies were observed when the correlation between dominant spatial
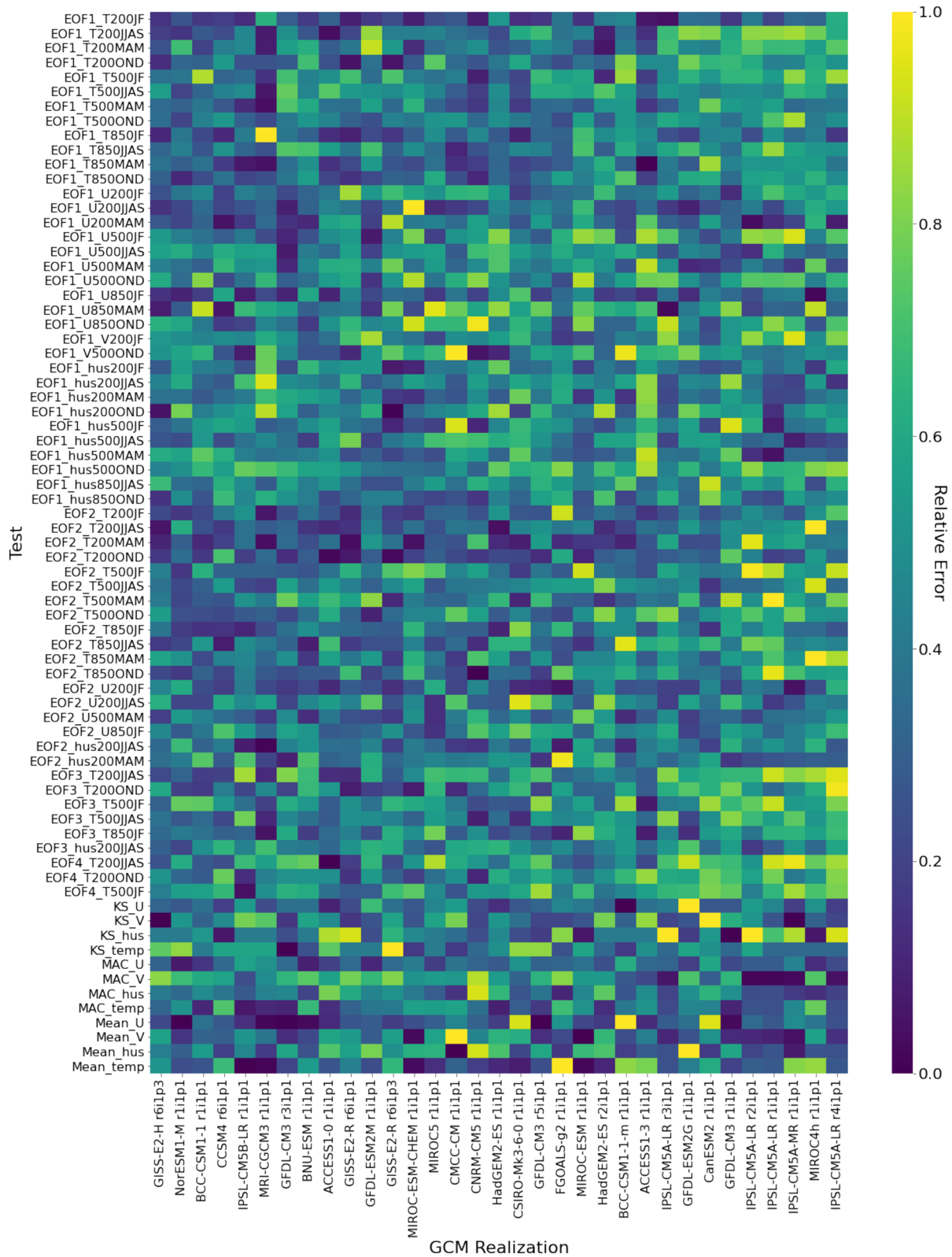
patterns of variability found in each realization and the respective nondegenerate EOF from MERRA-2 reanalysis data was examined, meaning that realizations are able to attribute the correct variance fraction explained to an EOF but encounter difficulties reconstructing the spatial pattern. This could point toward the spatial patterns being inverted in the realizations (EOF 1 occurs as EOF 2 and vice versa). An example of the spatial patterns observed can be found in Figure 5 for EOF 1 in JJAS. The spatial correlation observed between the realizations and MERRA-2 is strongest at 200 hPa and decreases toward 850 hPa (Figures 4 and 5). This could be related to the increasing complexity of the spatial patterns toward higher pressure levels. For this metric, no realization clearly performs strongest. When the variance fraction explained and the spatial patterns are combined into a single metric, the RMS of the relative errors ranges from 0.34 to 0.66 (average: 0.47) (Table 2). Here, the lower ranked realizations display greater RMSs than the higher ones, with the top 10 realizations having an average RMS of 0.40 and the bottom 10 having an average of 0.57.

Upon examination of the absolute mean difference metric at 500 hPa in the Kilimanjaro grid cell, all realizations perform very well, except FGOALS-g2 r1i1p1, which does not lie within the confidence interval of the observational record (see Figure 3: only yellow pixel in the bottom line). For the mean annual cycle, the majority of realizations reproduce its shape and magnitude (RMSE ≤ 1K). The KS test concludes that, for all the realizations, the null hypothesis that the realization and observations are drawn from the same distribution cannot be rejected.

4.2. Specific Humidity

The tendencies observed for temperature are not as clear in the specific humidity evaluation. With regard to the variance fraction explained metric associated with all the EOFs, the majority of realizations are within the confidence interval of the respective MERRA-2 value. EOF 1 at 500 hPa is the exception, where 27 realizations lie outside the confidence interval. The majority of realizations (≥97%) also have difficulties capturing the dominant spatial patterns. They typically explain less than 50% of the variance of MERRA-2. When combining these two metrics into one, the RMS of the relative errors ranges from 0.32 to 0.66, for a total of 13 tests (Table 2). Unlike temperature, the higher ranked models do not have smaller RMSs than the lower ranked ones with results being fairly similar for all realizations.

All the realizations are a good fit with regard to the mean specific humidity metric at 500 hPa over Kilimanjaro and are all found within the confidence interval of the observational record. For the mean annual cycle metric, the majority of realizations capture its shape and magnitude (average Pearson correlation coefficient: 0.8, RMSE: 0.73 g/kg). It should be noted that unlike for temperature, FGOALS-g2 r1i1p1 is a very good fit for the mean annual cycle (Figure 3). In the KS test, GDFL-CM3 r3i1p1 performs best but, as with temperature, the test's null hypothesis cannot be rejected for any realization.

4.3. Winds

The majority of realizations (≥79%) are able to capture the variance fraction explained for U and V winds. The exception lies in EOF 1 at 850 hPa in OND, where only about half of the realizations capture the variance fraction explained. Similarly to specific humidity, no clear tendencies are observed in the correlation of the dominant spatial patterns. The majority of realizations explain less than 50% of the variance of MERRA-2 for U and V winds. However, for the U winds of EOF 1 at 850 hPa in JF, 22 realizations have a Pearson correlation greater than 0.7, with IPSL-CM5A-MR r1i1p1 explaining ~74% of the variance of MERRA-2. As with specific humidity, the RMSs for U and V winds are fairly similar for all realizations (Table 2 and Figure 3).

Figure 3. GCM realization ranking presented from best to worst (left to right) as determined by the relative error of the realizations for each metric, where darker is a better fit. EOF represents the metric evaluating how well the GCM reconstructed the dominant spatial pattern and its variance fraction explained. The number at the end of these labels refers to which nondegenerate EOF is being analyzed and the following letter and number represent the atmospheric state variable (hus, specific humidity; T, temperature; U, U winds; and V, V winds), the atmospheric pressure level (200, 500, or 850 hPa), and the season (JF, MAM, JJAS, and OND). KS represents the Kolmogorov-Smirnov test, mean represents the mean state test, and MAC the mean annual cycle test with appended acronyms representing the atmospheric state variable being evaluated. EOF, empirical orthogonal function; GCM, general circulation model; JF, January–February; JJAS, June–September; MAM, March–May; OND, October–December.
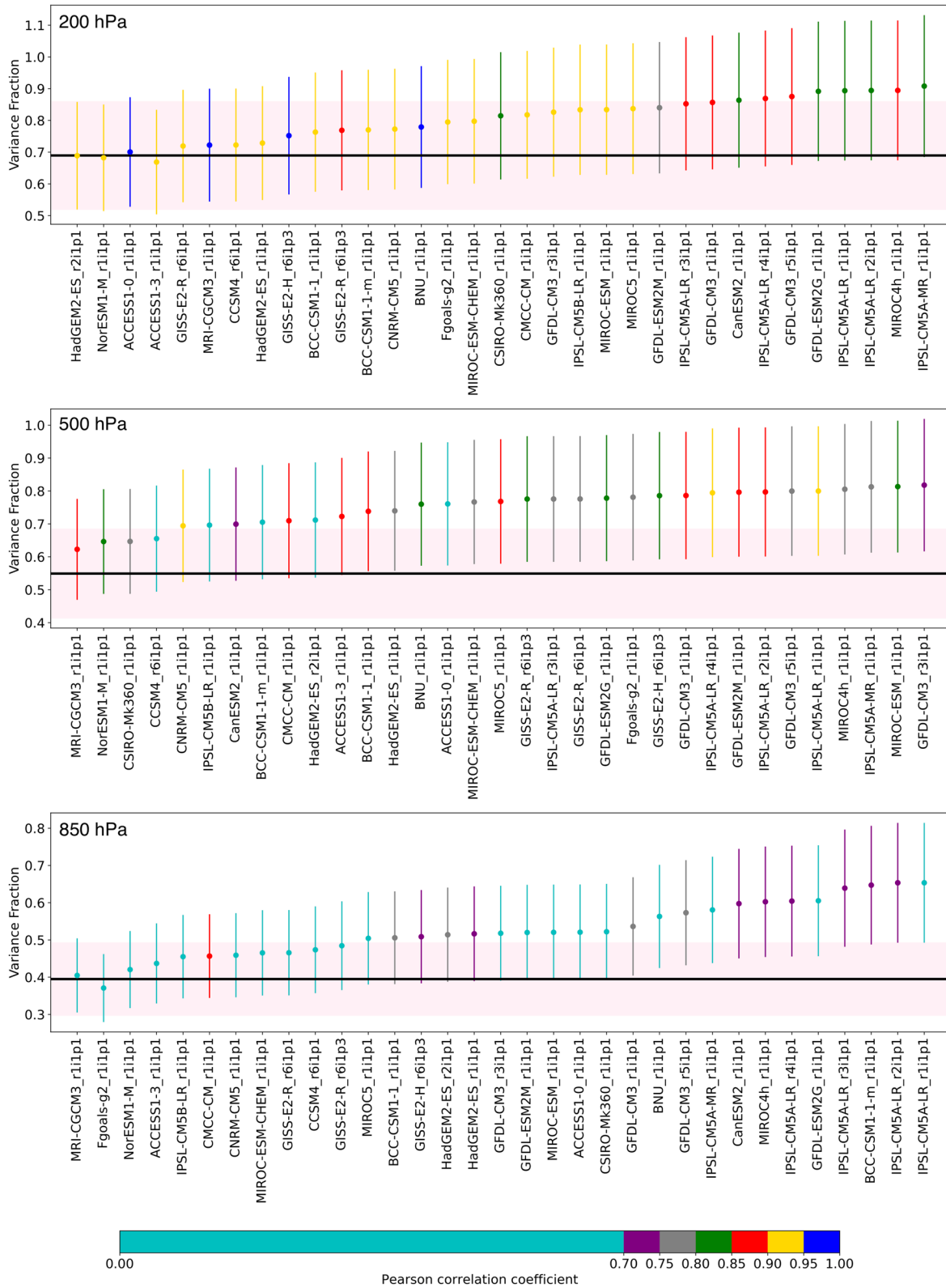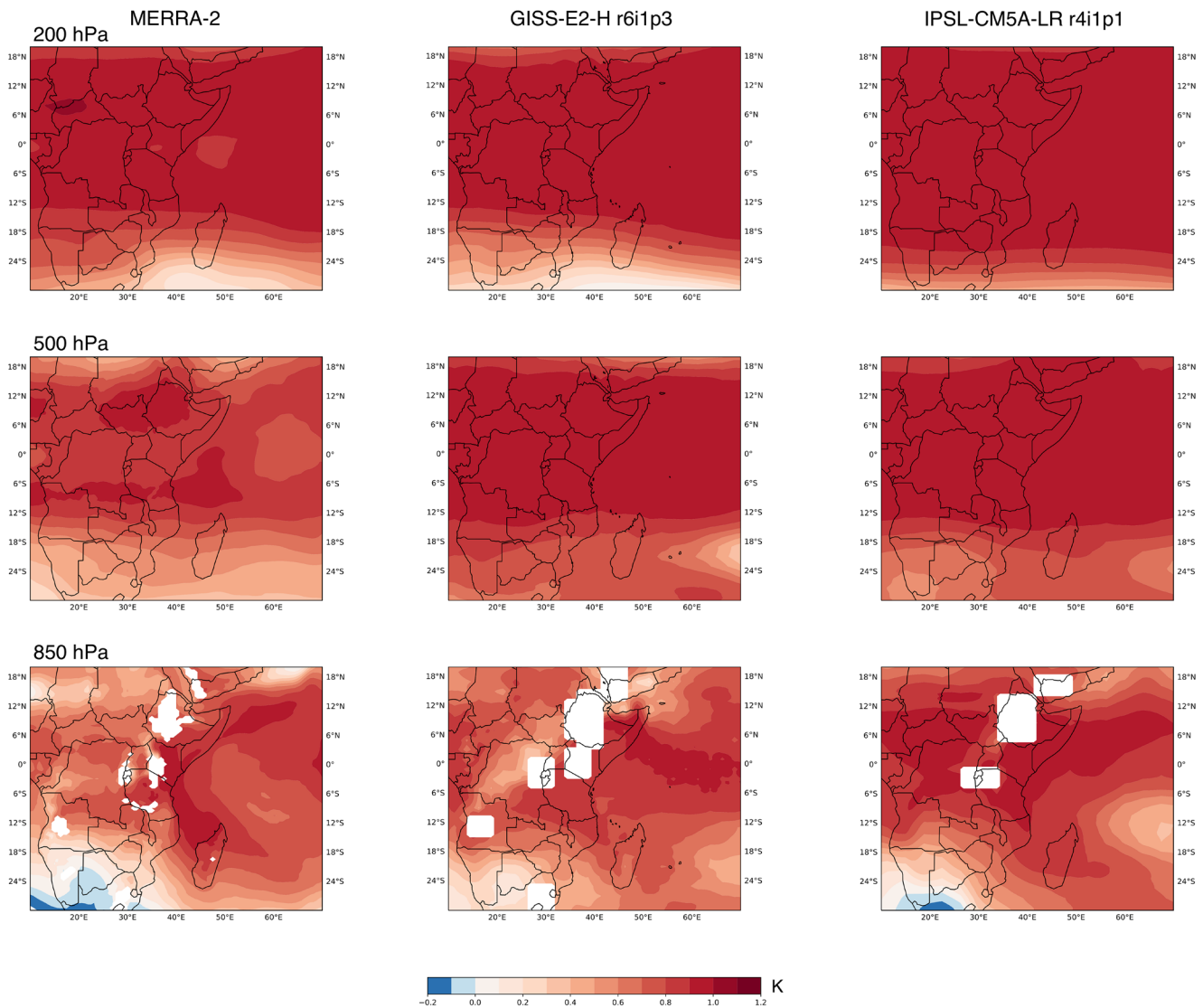
Figure 5. Examples of dominant spatial patterns (EOF 1) used for EOF spatial correlation analysis for temperature in JJAS. Figures on the left are the dominant spatial patterns of MERRA-2, the middle column are those of the highest ranked model, GISS-E2-H r6i1p3, and those on the right are from the lowest ranked model, IPSL-CM5A-LR r4i1p1. Absence of data (white regions) in plots at 850 hPa represents areas where the 850 hPa surface intersects with the land surface in the realization or MERRA-2. EOF, empirical orthogonal function; JJAS, June–September; MERRA-2, Modern-Era Retrospective Analysis for Research and Applications version 2.

The mean winds of the realizations at 500 hPa are all in the confidence interval of the observational record from the summit. However, V winds are on average underestimated. Realizations encounter difficulties reconstructing the shape of the mean annual cycle for V winds but are able to reproduce its magnitude. Only CanESM2, CSIRO-Mk3.6.0, and all IPSL realizations correctly reproduce the shape of the mean annual cycle (Figure 3). Realizations, however, perform better for the mean annual cycle of the U winds. The majority of realizations are able to reconstruct its shape, with the exception of IPSL-CM5B-LR r1i1p1 and MRI-CGCM3 r1i1p1 (Pearson correlation coefficient ≤ 0.2) and the KS test concludes that the null hypothesis can only be rejected for GFDL-ESM2G r1i1p1. On the other hand, for V winds, the null hypothesis that

Figure 4. Example of variance fraction and EOF spatial correlation analysis for EOF 1 of temperature in JJAS. The variance fraction of the GCM realizations is presented with respect to the MERRA-2 (black line with shading, 95% confidence interval) and is ranked from best fit (left) to worst (right). The variance fraction of each realization is presented with its 95% confidence interval. The varying colors associated with each realization indicate the Pearson correlation coefficient. EOF, empirical orthogonal function; GCM, general circulation model; JJAS, June–September; MERRA-2, Modern-Era Retrospective Analysis for Research and Applications version 2.

the observations and the realization come from the same distribution can be rejected for 15 realizations (all Access realizations, BNU-ESM r1i1p1, CanESM2 r1i1p1, CCSM4 r6i1p1, CMCC-CM r1i1p1, GFDL-CM3 r1i1p1, GISS-E2-R r6i1p3, all HadGEM2-ES realizations, IPSL-CM5A-LR r2i1p1, IPSL-CM5B-LR r1i1p1, MIROC5 r1i1p1, and NorESM1-M r1i1p1).

5. Discussion

5.1. Processes Analysis

The selection procedure cannot explicitly tell if the realizations were able to represent the key processes of the region. This means that, solely from the statistics of the selection results, we cannot conclude that the highest ranking models were best at capturing the most important processes of East Africa. One of the processes playing a major role in the region's climate is the IOD, interannually due to its influence on rainfall (Black, 2005; Clark et al., 2003) and on greater temporal time scales by affecting climate change in the region (Marchant et al., 2007). Other processes influence East African climate but, for brevity purposes, we focused on the how well the realizations represent the IOD using the method outlined in Thielke and Mölg (2019). A brief description can be found in Appendix B. The results of the associated three tests can be found in Figure 6, where it can be observed that a larger percent of higher ranked models are able to capture the IOD than lower ranked ones. There were 10 failed tests in the upper third of the models, with CCSM4 r6i1p1 representing three of these. In the middle third, there were 13 failed tests and 18 failed tests for the bottom third. This agrees with Thielke and Mölg (2019), who found 12 realizations of the GISS-E2-H to pass all tests. This points to the ability of the selection procedure to rank models that can capture one of the key processes in East Africa higher than those that cannot. However, these tests and others characterizing the additional processes in the region are not incorporated into the model selection procedure in order to maintain its generic application potential. They should, thus, be used along with the model selection procedure as an independent evaluation to stress the importance that GCM realizations capture the key processes of the region of interest.

5.2. Model Selection Sensitivity Evaluation

5.2.1. Reanalysis-Based Climate of the African Domain

The majority of the tests comprised in the model selection procedure are related to how well the GCM realizations simulate the climate of the African domain. These involve evaluating the space-time climate anomalies with respect to MERRA-2, chosen as it was found to be a suitable data set for representing important climatic features of the African domain and East Africa (Collier et al., 2018). As the top ranked GCM realization, GISS-E2-H r6i1p3, and its high ranking counterparts, GISS-E2-R r6i1p1 and r6i1p3, were produced at the same institution as MERRA-2, questions arise concerning sensitivity to the reanalysis product used for the selection procedure. To evaluate this, ERA5 (Hersbach et al., 2020), produced by the European Centre for Medium-Range Weather Forecasts (ECMWF), was used in place of MERRA-2 in the model selection procedure. First, the nondegenerate EOFs of ERA5 for each season and atmospheric state variable at 200, 500, and 850 hPa were calculated for 1980–2012 in the African domain (Table C1). Using the nondegenerate EOFs of ERA5 and following the procedure outlined in Section 3.2, the metrics for how well the GCM realizations simulate the climate of the African domain were calculated. These were combined with those from the 12 tests addressing how well the climate of the air layers sampled at the summit of Kilimanjaro was reconstructed. The procedure from Section 3.4 was then used to provide the model ranking found in Figure 7. It should be noted that the 12 tests evaluating the climate of the air layers sampled at the summit of Kilimanjaro have not been affected by the switch from MERRA-2 to ERA5 as they involve comparison with AWS data.

The ranking obtained using MERRA-2 (Figure 3) and that using ERA5 (Figure 7) are rather similar. The majority of the top 10 and bottom 10 ranked GCM realizations coincide between the two rankings. The similarities between ranking of the GISS realizations indicate that the choice of reanalysis product does not infer some genealogical bias. This point is further illustrated upon closer examination of the rankings associated with ERA5, which do not favor European models using parts of the ECMWF IFS, coming from the same institution as ERA5.
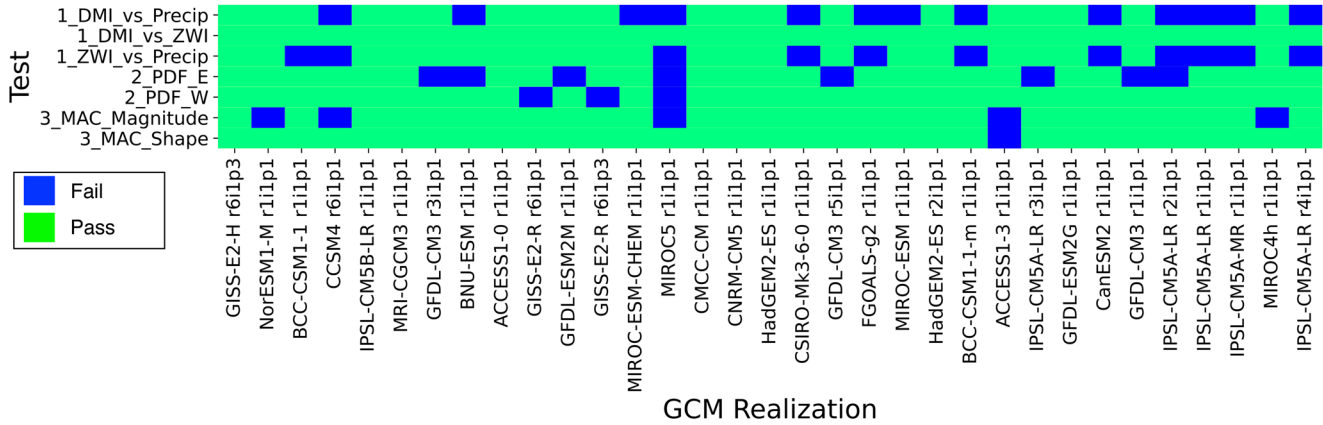
Figure 6. Heatmap of the results of processes analysis for the GCM realizations. The realizations are presented from highest ranked (left) to lowest. A green square indicates a model passed the processes analysis test, while blue indicates it failed the criteria outlined in Appendix B and Thielke and Mölg (2019). The number in front of each test name refers to the ability to capture IOD teleconnections (1), IOD magnitude (2), and East African precipitation (3). For tests starting with 1, the name indicates the evaluation of the linear correlation between the two indices, where precip refers to OND precipitation. For tests from 2, PDF is the probability density function and east and west indicate the eastern and western poles of the Indian Ocean sea surface temperature anomalies. For 3, MAC represents the mean annual cycle of the East African precipitation test, evaluating its shape and magnitude. GCM, general circulation model; IOD, Indian Ocean Dipole; OND, October–December.

5.2.2. Number of Metrics

To scrutinize the rather large the number of metrics (73), testing was undertaken to vary the number associated with the space-time climatic anomaly in the African domain. These metrics were chosen as they represent the vast majority of the metrics utilized. Two different approaches were employed: (1) eliminating metrics associated with an EOF with a variance fraction below a certain threshold and (2) considering only certain EOFs. Each of these are made up of a number of scenarios (Table 3). It should be noted that the scenario using a variance fraction threshold of 20% and that where only EOF 1 is considered are equivalent.

Using the technique elaborated in Section 3.4, the results obtained from each of the scenarios were used to produce a respective model ranking. To evaluate whether these rankings were better at capturing key processes of the region, the results of the IOD analysis (Section 5.1) were ordered with respect to them. This showed that as the number of metrics decreased, so did the ability of the higher ranked models to better simulate the IOD than lower ranked ones. This is clear when examining the number of failed IOD tests associated with the upper third, middle third, and bottom third of models for each scenario (Figure 8). These results highlight the importance of the lesser dominant EOFs (EOF 2–EOF 4), which explain a smaller fraction of the variance, and the value of the metrics included within the selection procedure.

5.2.3. Ranking Method

As elaborated in Section 3.4, the GCM realizations were ranked by calculating the RMS of their relative errors. This resulted in the ranking found in Figure 3. Utilizing this technique means that the model that performs best on average ranks highest. This model might perform poorly in a test, such as the top ranked GISS-E2-H r6i1p3 in the KS test for temperature (large D value) and mean annual cycle test for V winds, but still ranks high as it performs best on average, that is has the lowest RMS. To evaluate the robustness of this ranking method, the GCM realizations were also ranked according to another method, lowest cumulative relative error (Equation 2):

$$\text{Sum} = \sum_{i=1}^{n} \text{modelRE}_i \tag{2}$$

where Sum represents the cumulative relative error for a GCM realization, n is the number of metrics (73), and modelRE is the relative error of the GCM realization for the respective metric. The difference between the two techniques is minimal as the rankings have remained fairly similar (Table D1), confirming the
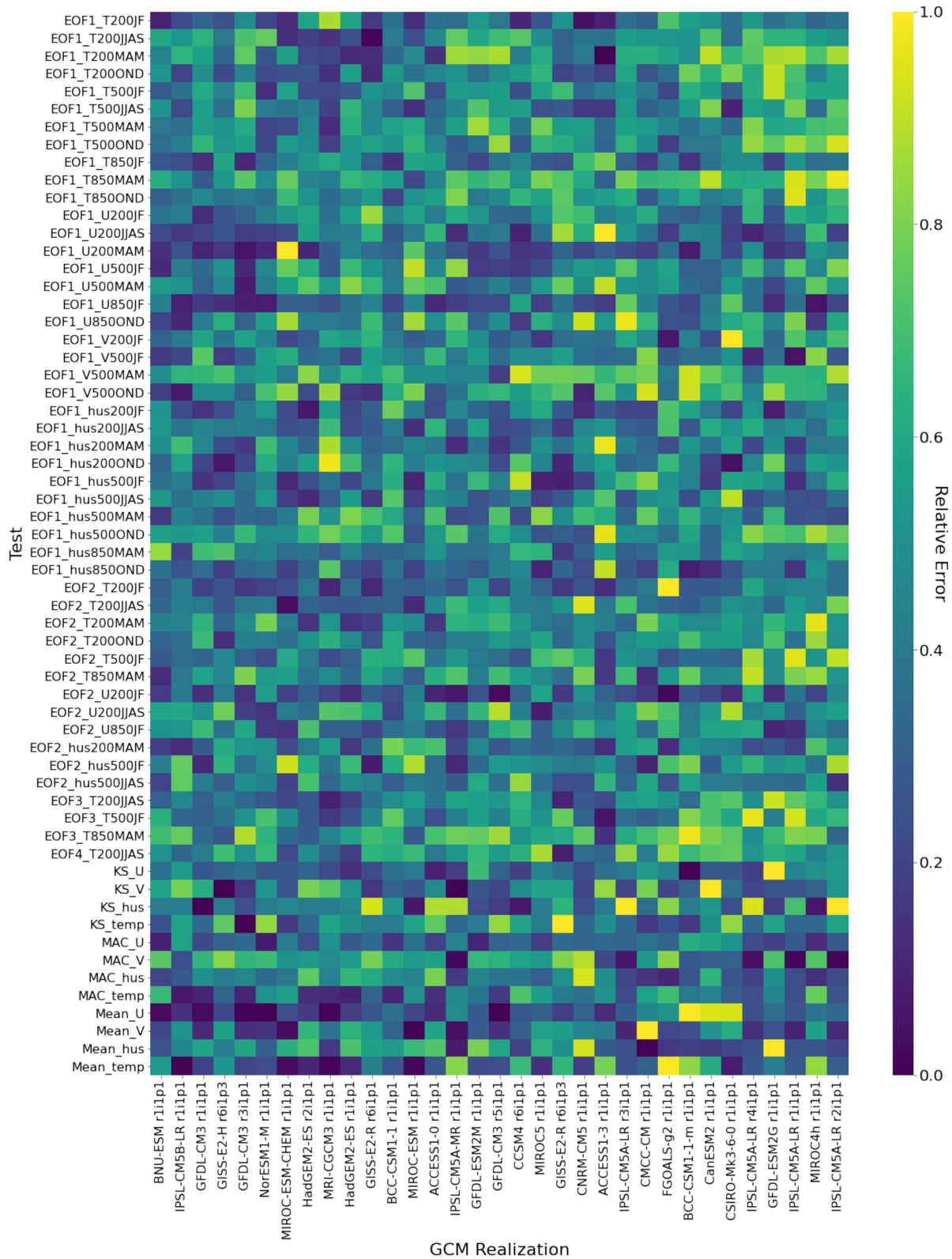
Table 3
Approaches to Test the Number of Metrics and Their Respective Scenarios

| Approach	Scenario	Number of eliminated metrics	Total number of metrics remaining
(1) Variance fraction threshold	10% threshold	9	64
	15% threshold	17	56
	20% threshold	25	48
(2) Only certain EOFs considered	EOF 1	25	48
	EOF 1 and EOF 2	9	64

findings of Rupp et al. (2013). This illustrates that, in most cases, the model with the lowest relative error is also the model with the lowest RMS.

5.3. Dynamical Downscaling

Furthermore, our concept and motivation imply that higher ranked realizations should result in a better downscaling product. To get a sense for this hypothesis, we undertook three dynamical downscaling simulations for East Africa, one with MERRA-2, one with a "good," and one with a "bad" boundary condition, which come from the highest, GISS-E2-H r6i1p3, and lowest ranked, IPSL-CM5A-LR r4i1p1, GCM realizations, respectively. The three resultant downscaled climates can then be evaluated by comparing them to the observational data set CRU TS 4.04 (Harris et al., 2020). We specified a 10-year period and a 50-km target resolution centered over East Africa and Kilimanjaro to keep the computational expense for the discussion within limits.

In this regard, output from each GCM realization and MERRA-2 were used to drive the lateral boundaries of the Weather and Research Forecasting (WRF) model, version 4.1.2, over a domain typically used for regional atmospheric and climatic modeling of East Africa (33°N–37°S, 1°–75°E) from January 1996 to December 2005 (Figure 1b). Note that the WRF domain is slightly larger than that used for the model selection procedure as we are strictly following the setup of the most recent WRF study of Collier et al. (2019). More details and an overview of the configuration, which was taken from published WRF studies on East Africa (Collier et al., 2018, 2019), can be found in Appendix E. Over equatorial East Africa (5°N–5°S, 30°–40°E), monthly downscaled precipitation and surface air temperature from the WRF simulations using MERRA-2 and the "good" and "bad" GCM realizations as boundary conditions were evaluated spatially and temporally with respect to the monthly precipitation and surface air temperature from the observational data set CRU TS 4.04. The downscaled climates were evaluated twofold: (1) ability to simulate the mean annual cycle and (2) ability to capture the seasonal spatial patterns. For the mean annual cycle analysis, the spatial average for surface air temperature and precipitation were taken, individually, for the simulations driven by the two GCM realizations and MERRA-2. The mean annual cycle was then calculated and compared in terms of its magnitude, shape, and bias with that obtained from CRU TS 4.04. The magnitude analysis involved taking the RMSE, while the shape comparison employed the Pearson correlation. For the seasonal spatial pattern evaluation, the study period was split up seasonally (JF, MAM, JJAS, and OND) and the seasonal average was taken temporally for over the entire period. The seasonal spatial patterns from the simulations driven by the realizations and MERRA-2 were correlated to those from CRU TS 4.04 using the Pearson correlation. The results of these analyses can be found in Figure 9.

Figure 7. GCM realization ranking using ERA5, presented from best to worst (left to right) as determined by the relative error of the realizations for each metric, where darker is a better fit. EOF represents the metric evaluating how well the GCM reconstructed the dominant spatial pattern and its variance fraction explained. The number at the end of these labels refers to which nondegenerate EOF is being analyzed and the following letter and number represent the atmospheric state variable (hus, specific humidity; T, temperature; U, U winds; and V, V winds), the atmospheric pressure level (200, 500, or 850 hPa), and the season (JF, MAM, JJAS, OND). KS represents the Kolmogorov-Smirnov test, mean represents the mean state test, and MAC the mean annual cycle test with appended acronyms representing the atmospheric state variable being evaluated. EOF, empirical orthogonal function; GCM, general circulation model; JF, January–February; JJAS, June–September; MAM, March–May; OND, October–December.
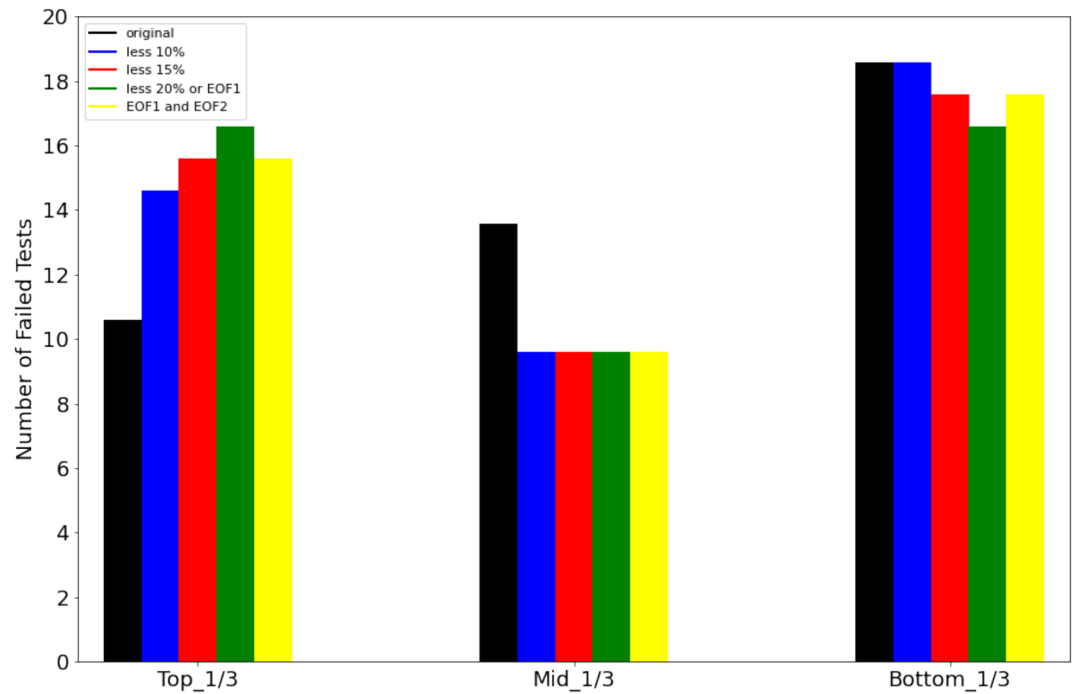
Figure 8. Number of failed tests during the IOD processes analysis for the original ranking and that of each scenario. Results are presented for the upper third of the models, middle third, and bottom third based on the associated ranking. IOD, Indian Ocean Dipole.

For temperature, using the top ranked GCM, GISS-E2-H r6i1p3, and/or MERRA-2 leads to a better downscaling result than IPSL-CM5A-LR r4i1p1. This is apparent when examining the mean annual cycle, where the simulations driven by GISS-E2-H r6i1p3 and MERRA-2 are better able to capture its magnitude and shape and for almost all 12 months have a lower bias with respect to the downscaled surface temperature from CRU TS 4.04. The exception lies in May, where IPSL-CM5A-LR r4i1p1 has a slightly lower bias than MERRA-2. For the seasonal spatial patterns of temperature, GISS-E2-H r6i1p3 and MERRA-2 perform clearly better than IPSL-CM5A-LR r4i1p1 over all four seasons (Figure 9).

This also holds true for the seasonal patterns of precipitation. However, for JJAS, the three perform similarly with IPSL-CM5A-LR r4i1p1 doing better than MERRA-2 but worse than GISS-E2-H r6i1p3. For the mean annual cycle, MERRA-2 is best at reproducing the shape, while it encounters difficulty with the magnitude. GISS-E2-H r6i1p3 is better able to capture the magnitude than IPSL-CM5A-LR r4i1p1 and MERRA-2. Examination of its mean annual cycle bias shows that GISS-E2-H r6i1p3 has a lower bias over 7 of the 12 months. It should also be noted that IPSL-CM5A-LR r4i1p1 is better able to capture the magnitude of the mean annual cycle than MERRA-2. All three WRF simulations encounter difficulties reconstructing the OND wet season, where they significantly overestimate precipitation. This could be associated with the physical cumulus parametrization (Kain-Fritsch) used in the WRF simulations (Otieno et al., 2020). It could also be related to a known issue that CMIP5 GCMs have in reconstructing the mean annual cycle for precipitation in East Africa, with underestimating the long rains and overestimating the short ones. A too weak Atlantic meridional overturning circulation in the GCMs leading to a sea surface temperature (SST) bias over the Indian Ocean is postulated to result in the difficulties capturing the rainy seasons (Yang et al., 2015).

Overall, the dynamical downscaling results show that using a top ranked GCM for the boundary conditions leads to a better dynamical downscaling result over equatorial East Africa than a low-ranked GCM, a result that is most obvious for surface air temperature, where a clear distinction between the top ranked realizations and the lower ones exists in the GCM selection results (Table 2). However, the benefit of GCM boundary selection is also evident for precipitation.

Temperature
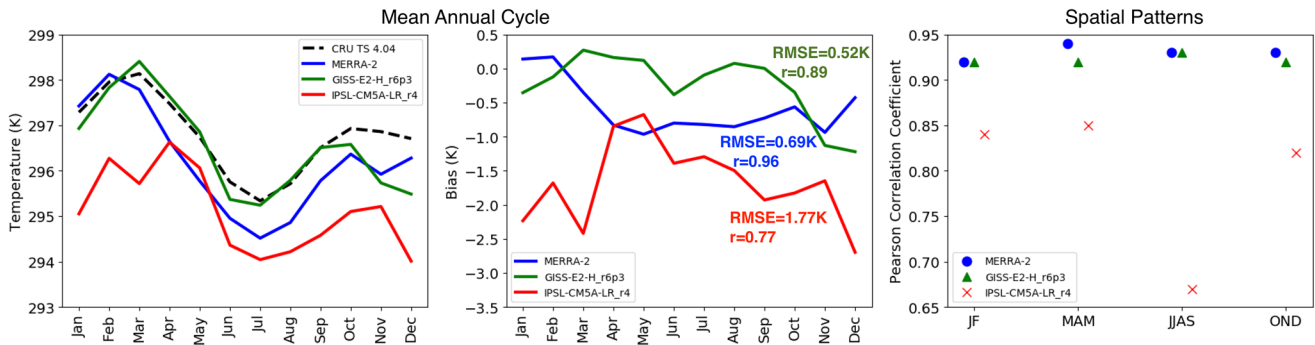
Precipitation
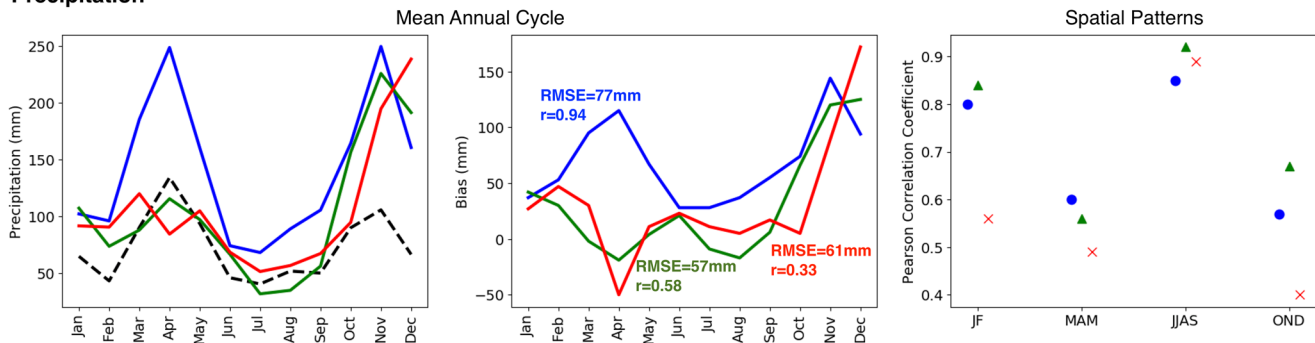
Figure 9. Comparison of surface air temperature and precipitation (different rows) from three 10-year WRF simulations to CRU TS 4.04 for the period of 1996–2005 for equatorial East Africa. GISS-E2-H r6i1p3 is the top ranked realization (green) and IPSL-CM5A-LR r4i1p1 is the lowest ranked one (red). The first two columns represent the mean annual cycle analysis and the third focuses on reconstructing seasonal spatial patterns. The numbers on the plots of the second column represent the RMSE and Pearson correlation coefficient (r) evaluating the mean annual cycle. RMSE, root mean square error; WRF, Weather and Research Forecasting.

6. Conclusions

We presented a model selection procedure to determine the suitability of GCM realizations for downscaling studies. Our results indicated that no one realization performed markedly best in all the tests but it was able to rank higher models that captured a key process in the region and produced a better downscaling product, which is a promising prospect.

In evaluation or selection of CMIP5 models, only one realization per model is typically analyzed (e.g., McSweeney et al., 2012; Ongoma et al., 2019). However, studies of climate extremes indicate the importance of including all realizations due internal climate variability amplifying or obscuring extreme events (Fischer et al., 2014; Perkins & Fischer, 2013). In Figure 3, it can be observed that most realizations from the same GCM perform similarly in the model selection procedure, as expected since the key difference between realizations is internal climate variability. Upon first glance, this does not appear to hold true for GFDL-CM3 realizations which rank 28th (r1i1p1), 7th (r3i1p1) and 19th (r5i1p1). However, upon closer examination of their RMSs (Table 2), it can be seen that they are rather similar with the main difference being GFDL-CM3 r3i1p1 is better able to capture (smaller RMS) the dominant spatial patterns and their variance fraction explained for U winds over the African domain.

The selection procedure focused on analysis of the atmospheric state variables at varying atmospheric pressure levels. These variables are not usually the focus of model selection procedures but are important as they are used to drive the lateral boundaries of RCMs in dynamical downscaling studies and their evaluation, on at least one atmospheric pressure level, is required to adequately select a GCM for this purpose (Jury et al., 2015). They also tend to provide more reliable seasonal forecasts than SSTs and teleconnections (Nicholson, 2017). In particular, the ability to reproduce the mean climatic state is also analyzed with respect to observations at a fixed atmospheric pressure level. This step is important and usually overlooked

as models have a harder time capturing the mean climatic state than the space-time anomalies (Bindoff et al., 2013). By including it, it provides a thorough examination of the ability of the realization to reproduce the climate of a specific region of interest. All these analyzes are not computationally expensive but result in a comprehensive overview of the climate for the region of interest at different atmospheric pressure levels, time scales, and for various aspects of the climatic state.

The model selection procedure shed light on some of the strengths and weaknesses of the GCM realizations. In all metrics, the realizations performed strongest for temperature. This has also been observed in studies of the surface temperature of the region (Aloysius et al., 2016; Ongoma et al., 2018). For specific humidity, they were able to capture the mean climatic state but encountered difficulties reconstructing the dominant spatial patterns of variability. This could be related to the issues CMIP5 GCMs have in simulating precipitation over the region (Aloysius et al., 2016; Yang et al., 2015). For these two variables, the realizations encountered less difficulties simulating the variance fraction explained of each nondegenerate EOF than reconstructing their spatial patterns. This points to the possibility that the spatial patterns are inverted in the realizations. Realizations had the most issues when attempting to reconstruct the tendencies associated with winds, particularly V winds. Furthermore, the GCM realizations were able to better reproduce the variance fraction explained and spatial patterns for EOF 1 than for EOF 2–EOF 4. This indicates that the realizations display a greater skill in reconstructing the dominant spatial pattern (EOF 1, explaining the largest variance fraction). It also points to the importance of evaluating the lesser dominant spatial patterns (EOF 2–EOF 4, explaining a smaller variance fraction) as they distinguish which realizations are able to better simulate the climate of the region of interest. This is further highlighted when examining the number of EOF metrics utilized. All these findings highlight areas of interest and possible improvement in climate models. Overall, the results clearly indicate that realizations perform better for some metrics but it is important not to weigh these tests more favorably as all atmospheric state variables and climate tendencies are important to the climate of the region.

The rankings produced by the model selection procedure do not explicitly tell whether a realization was able to reconstruct the key processes of the region of interest. To evaluate this for the case study of East Africa, their ability to reconstruct the IOD was analyzed and it was found that a larger percentage of higher ranked models were better able to capture the IOD than lower ranked ones. This illustrates the ability of the selection procedure to rank realizations that can capture a key process of the region of interest higher than those that cannot. However, to maintain the generic potential of the selection procedure, this process analysis step was not incorporated within the testing scheme. It should, nevertheless, be used in conjunction as an independent indicator of model skill.

Furthermore, the assumption and motivation behind the selection procedure was that higher ranked realizations would produce a better downscaling product than lower ranked ones. This was confirmed by running three WRF simulations forced respectively by MERRA-2, the top ranking realization (GISS-E2-H r6i1p3), and the lowest ranked one (IPSL-CM5A-LR r4i1p1) over East Africa for the period of 1996–2005 and comparing the simulated monthly surface temperature and precipitation over equatorial East Africa to CRU TS 4.04. From the examination of the mean annual cycle and seasonal spatial patterns, one can see that the top ranked realization led to a better downscaling product than the lower ranked one, a result more evident for temperature than precipitation.

The selection procedure is, however, unfortunately biased toward the group of the space-time anomaly tests. This is due to the number of nondegenerate EOFs from the seasonal analysis at 200, 500, and 850 hPa. The analyses of their dominant spatial patterns of variability and their importance resulted in 61 tests. This is a significantly higher amount of tests than the 12 for the mean climatic state. This could possibly be compensated for by weighting but that adds a further degree of subjectivity to the selection procedure.

The 12 mean climatic state tests were undertaken with respect to an observational data set that spans March 2000 to September 2017, leading to a small overlap with the historical simulations. This data set was utilized as decadal AWS records at high altitude are scarce and its unique location, on the summit of Kilimanjaro, allowed for the capture of the large-scale climatic properties of the tropical midtroposphere, and comparison to CMIP5 output. For analysis, the historical simulations had to be extended using RCP 4.5 and/or the historicalExt simulation to allow for analysis. However, the small overlap period means that from 2006 onward

these tests are evaluating how well RCP 4.5 and/or historicalExt did in simulating the mean climatic state rather than how well the historical simulation did. While similar results are expected if the historical simulation were present for this period, this should be noted. Due to the bias toward the space-time anomaly test, the relatively shorter overlap period should not significantly impact the model selection results.

While this methodology was presented for the case study of East Africa, it is comprehensive and easily applicable. Nevertheless, its skill has only been evaluated over East Africa and should be thoroughly examined for other regions. The evaluation of the space-time anomaly only requires selection of a suitable reanalysis product for the study area. The mean climatic state analysis presented was based on observations of the tropical midtroposphere from the summit of Kilimanjaro. This is rather case study specific but can be easily adjusted depending on the problem being addressed by using observational data, radiosondes, or satellite data for the target region and considering the scale they represent. Moreover, we focused the discussion on a regional atmospheric modeling experiment to demonstrate the added value of the methodology for dynamical downscaling. However, it should also be explored in statistical downscaling which is known to benefit from physically sound data inputs (e.g., Hewitson & Crane, 2006), as captured by the four state variables temperature, humidity, and winds.

Appendix A: Nondegenerate EOFs for MERRA-2

Table A1
Nondegenerate EOFs and Their Percent Variance Explained for MERRA-2 Reanalysis Data for the African Domain

Variable	Season	Pressure level (hPa)	Number of nondegenerate EOFs	Variance explained (%)
Temperature	JF	200	2	79.1, 9.7
		500	4	50.0, 17.4, 10.2, 6.2
		850	3	43.9, 17.9, 10.3
	MAM	200	2	78.6, 13.2
		500	2	64.0, 10.7
		850	2	41.3, 16.5
	JJAS	200	4	69.0, 14.1, 6.2, 3.0
		500	3	54.9, 18.2, 8.2
		850	2	39.5, 16.7
	OND	200	4	77.9, 9.0, 4.0, 2.4
		500	2	56.3, 14.4
		850	2	37.9, 19.3
Specific humidity	JF	200	1	50.7
		500	1	29.4
		850	0	–
	MAM	200	2	52.1, 14.6
		500	1	33.7
		850	0	–
	JJAS	200	3	59.5, 12.7, 7.5
		500	1	41.3
		850	1	22.9
	OND	200	1	57.0
		500	1	51.2
		850	1	28.2

Table A1
Continued

Variable	Season	Pressure level (hPa)	Number of nondegenerate EOFs	Variance explained (%)
U winds	JF	200	2	55.8, 14.2
		500	1	38.2
		850	2	30.3, 18.3
	MAM	200	1	49.1
		500	2	28.3, 16.1
		850	1	29.0
	JJAS	200	2	59.7, 13.4
		500	1	27.5
		850	0	–
	OND	200	0	–
		500	1	33.1
		850	1	42.2
V winds	JF	200	1	35.3
		500	0	–
		850	0	–
	MAM	200	0	–
		500	0	–
		850	0	–
	JJAS	200	0	–
		500	0	–
		850	0	–
	OND	200	0	–
		500	1	24.2
		850	0	–

Note. If multiple nondegenerate EOFs are present, first value in the variance explained will correspond to the first nondegenerate EOF, the second will correspond to the second EOF, and so on.

Appendix B: Brief Outline of IOD Processes Analysis

The method outlined in Thielke and Mölg (2019) was utilized to analyze how well the GCM realizations captured the IOD for 1948–2005, a time period chosen due the availability of the NCEP/NCAR reanalysis data set and overlap with the CMIP5 historical simulation. First, the ability to capture the relevant IOD teleconnections was determined by examining the linear correlations between the September–October–November (SON) Dipole Mode Index (DMI), the SON Zonal Wind Index (ZWI), and the OND precipitation. The DMI referred to the area-averaged difference in SST anomalies for the western (50°–70°E, 10°N–10°S) and eastern (90°–110°E, 0°–10°S) poles of the Indian Ocean. The ZWI was the area-averaged (60°–90°E, 5°N–5°S) zonal wind component at 850 hPa. The OND precipitation was evaluated for the East African region (25°–40°E, 10°N–15°S). Realizations with a *p*-value greater than 0.05 did not meet this criterion. Second, the ability to reconstruct the IOD magnitude was examined using the probability density functions for the western and eastern poles of the Indian Ocean SST anomalies. Realizations not within one standard deviation of the mean of three observational SST data sets (version 3b and 4 of the Extended Reconstructed SST [ERSST] data set [Huang et al., 2015a, 2015b; Liu et al., 2015; Smith et al., 2008; Xue et al., 2003] and COBE-SST2 [Hirahara et al., 2019]) are deemed to fail at reconstructing the magnitude of the IOD. Finally, evaluation of the representation of East African precipitation by examining, with respect to the monthly land surface precipitation from the CRU TS 4.03 (Harris et al., 2020), how the realization captures the mag-

nitude and shape of the mean annual cycle. Realizations with a RMSE above the standard deviation of the observational data set and a Pearson correlation with a *p*-value greater than 0.05 failed this test.

Appendix C: Nondegenerate EOFs for ERA5

Table C1
Nondegenerate EOFs and Their Percent Variance Explained for ERA5 Reanalysis Data for the African Domain

Variable	Season	Pressure level (hPa)	Number of nondegenerate EOFs	Variance explained (%)
Temperature	JF	200	2	76.6, 9.7
		500	3	48.4, 19.0, 10.3
		850	1	46.7
	MAM	200	2	76.2, 14.2
		500	1	65.1
		850	3	35.3, 19.1, 11.0
	JJAS	200	4	64.3, 16.1, 7.1, 3.7
		500	1	61.3
		850	0	–
	OND	200	2	70.8, 13.2
		500	1	55.8
		850	1	34.0
Specific humidity	JF	200	2	37.5, 15.7
		500	2	29.2, 14.6
		850	0	–
	MAM	200	1	50.4
		500	1	27.3
		850	1	26.9
	JJAS	200	1	62.4
		500	2	36.5, 19.3
		850	0	–
	OND	200	1	54.8
		500	1	51.0
		850	1	32.7
U winds	JF	200	2	58.1, 14.7
		500	1	43.6
		850	2	31.8, 16.8
	MAM	200	1	44.8
		500	1	28.3, 30.4
		850	0	–
	JJAS	200	2	63.8, 12.5
		500	0	–
		850	0	–
	OND	200	0	–
		500	0	–
		850	1	39.9

Table C1
Continued

Variable	Season	Pressure level (hPa)	Number of nondegenerate EOFs	Variance explained (%)
V winds	JF	200	1	37.4
		500	1	30.4
		850	0	–
	MAM	200	0	–
		500	1	25.3
		850	0	–
	JJAS	200	0	–
		500	0	–
		850	0	–
	OND	200	0	–
		500	1	24.1
		850	0	–

Note. If multiple nondegenerate EOFs are present, first value in the variance explained will correspond to the first nondegenerate EOF, the second will correspond to the second EOF, and so on.

Appendix D: GCM Ranking Sensitivity Testing

Table D1
GCM Realizations Rankings Produced Using the RMS of the Relative Errors and the Sum of the Relative Error, RE_tot

Model	Ranking RMS	Ranking sum	RE_tot
ACCESS1-0 r1i1p1	9	9	29.058
ACCESS1-3 r1i1p1	24	22	32.037
BCC-CSM1-1 r1i1p1	3	5	27.887
BCC-CSM1-1-m r1i1p1	23	23	32.635
BNU-ESM r1i1p1	8	11	29.434
CCSM4 r6i1p1	4	4	27.725
CMCC-CM r1i1p1	15	14	30.269
CNRM-CM5 r1i1p1	16	15	30.164
CSIRO-Mk3-6-0 r1i1p1	18	17	31.023
CanESM2 r1i1p1	27	27	34.547
FGOALS-g2 r1i1p1	20	18	31.587
GFDL-CM3 r1i1p1	28	30	35.245
GFDL-CM3 r3i1p1	7	7	28.350
GFDL-CM3 r5i1p1	19	19	31.859
GFDL-ESM2G r1i1p1	26	26	33.557
GFDL-ESM2M r1i1p1	11	12	29.531
GISS-E2-H r6i1p3	**1**	**1**	**26.197**
GISS-E2-R r6i1p1	10	8	28.852
GISS-E2-R r6i1p3	12	10	29.348
HadGEM2-ES r1i1p1	17	21	31.943
HadGEM2-ES r2i1p1	22	24	32.686
IPSL-CM5A-LR r1i1p1	30	28	34.696

Table D1
Continued

Model	Ranking RMS	Ranking sum	RE_tot
IPSL-CM5A-LR r2i1p1	29	31	35.333
IPSL-CM5A-LR r3i1p1	25	25	32.845
IPSL-CM5A-LR r4i1p1	**33**	**33**	**36.744**
IPSL-CM5B-LR r1i1p1	5	6	27.907
IPSL-CM5A-MR r1i1p1	31	29	35.129
MIROC-ESM r1i1p1	21	20	31.890
MIROC-ESM-CHEM r1i1p1	13	13	29.948
MIROC4h r1i1p1	32	32	36.731
MIROC5 r1i1p1	14	16	30.932
MRI-CGCM3 r1i1p1	6	2	26.636
NorESM1-M r1i1p1	2	3	27.286

Note. Top and bottom ranked GCM realizations are presented in bold.

Appendix E: Dynamical Downscaling Test

Two GCM realizations (GISS-E2-H r6i1p3 and IPSL-CM5A-LR r4i1p1) and MERRA-2 were, respectively, used to force the lateral boundaries of WRF over a domain spanning 33°N–37°S by 1°–75°E for 1996–2005. This single-domain consists of 50 km grid spacing centered over Kilimanjaro (Figure 1b). The choice of configuration, including physics and parametrization schemes, followed the detailed WRF studies of Collier et al. (2018, 2019). To prevent model drift, grid analysis nudging was utilized and applied over the domain (Figure 1b) to the horizontal winds, potential temperature, and humidity above the planetary boundary layer. Apart from humidity, the default nudging coefficients were used (Mölg & Kaser, 2011). As per Otte et al. (2012), the humidity nudging coefficient was decreased by a factor of 10. The simulations were run over 2 years with 8 days of spin-up and then combined to provide results for the entire 1996–2005 period, utilizing over 50,400 core hours. While 8 days is sufficient for atmospheric variables to reach equilibrium, it is not sufficient for the soil subsystem. This can require over 1 year. However, Jerez et al. (2020) have shown that this can be ignored unless a study where the deep soil-atmosphere feedback is critical is being undertaken.

Data Availability Statement

These data were accessed from the ESGF repository at https://esgf-node.llnl.gov/search/cmip5/. The MERRA-2 data used in this publication were provided by the Global Modeling and Assimilation Office (GMAO) at the NASA Goddard Space Flight Center. They are available at MDISC (https://disc.gsfc.nasa.gov/datasets/M2I6NPANA_5.12.4/summary).

References

Aalbers, E. E., Lenderink, G., van Meijgaard, E., & van den Hurk, B. J. J. M. (2018). Local-scale changes in mean and heavy precipitation in Western Europe, climate change or internal variability? *Climate Dynamics*, *50*(11), 4745–4766. https://doi.org/10.1007/s00382-017-3901-9

Akinsanola, A., & Ogunjobi, K. (2017). Evaluation of present-day rainfall simulations over West Africa in CORDEX regional climate models. *Environmental Earth Sciences*, *76*(10), 366. https://doi.org/10.1007/s12665-017-6691-9

Aloysius, N. R., Sheffield, J., Saiers, J. E., Li, H., & Wood, E. F. (2016). Evaluation of historical and future simulations of precipitation and temperature in central Africa from CMIP5 climate models. *Journal of Geophysical Research: Atmospheres*, *121*, 130–152. https://doi.org/10.1002/2015JD023656

Arora, V., Scinocca, J., Boer, G., Christian, J., Denman, K., Flato, G., et al. (2011). Carbon emission limits required to satisfy future representative concentration pathways of greenhouse gases. *Geophysical Research Letters*, *38*, L05805. https://doi.org/10.1029/2010GL046270

Bentsen, M., Bethke, I., Debernard, J., Iversen, T., Kirkevåg, A., Seland, Ø., et al. (2013). The Norwegian earth system model, NorESM1-M—Part 1: Description and basic evaluation of the physical climate. *Geoscientific Model Development*, *6*(3), 687–720. https://doi.org/10.5194/gmd-6-687-2013

Bi, D., Dix, M., Marsland, S. J., O'Farrell, S., Rashid, H., Uotila, P., et al. (2013). The ACCESS coupled model: Description, control climate and evaluation. *Australian Meteorological and Oceanographic Journal*, *63*(1), 41–64.

Bindoff, N., Stott, P., AchutaRao, K., Allen, M., Gillett, N., Gutzler, D., et al. (2013). Detection and attribution of climate change: From global to regional. In T. F Stocker, D. Qin, G.-K. Plattner, M. Tignor, S .K. Allen & J. Boschung, et al. (Eds.), *Climate change 2013: The physical science basis*. Cambridge University Press.

Black, E. (2005). The relationship between Indian Ocean sea–surface temperature and East African rainfall. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *363*(1826), 43–47. https://doi.org/10.1098/rsta.2004.1474

Boko, M., Niang, I., Nyong, A., Vogel, C., Githeko, A., Medany, M., et al. (2007). Africa. In M. Parry, O. Canziani, J. Palutikof, P. van der Linden, & C. Hanson (Eds.), *Climate change 2007: Impacts, adaptation and vulnerability. Contribution of Working Group II to the fourth assessment report of the Intergovernmental Panel on Climate Change* (pp. 433–467). Cambridge University Press.

Clark, C. O., Webster, P. J., & Cole, J. E. (2003). Interdecadal variability of the relationship between the Indian Ocean zonal mode and East African coastal rainfall anomalies. *Journal of Climate*, *16*(3), 548–554. https://doi.org/10.1175/1520-0442(2003)016(0548:IVOTRB)2.0.CO;2

Collier, E., Mölg, T., & Sauter, T. (2018). Recent atmospheric variability at Kibo Summit, Kilimanjaro, and its relation to climate mode activity. *Journal of Climate*, *31*(10), 3875–3891. https://doi.org/10.1175/JCLI-D-17-0551.1

Collier, E., Sauter, T., Mölg, T., & Hardy, D. (2019). The influence of tropical cyclones on circulation, moisture transport, and snow accumulation at Kilimanjaro during the 2006–2007 season. *Journal of Geophysical Research: Atmospheres*, *124*, 6919–6928. https://doi.org/10.1029/2019JD030682

Collins, W., Bellouin, N., Doutriaux-Boucher, M., Gedney, N., Halloran, P., Hinton, T., et al. (2011). Development and evaluation of an Earth-System model—HadGEM2. *Geoscientific Model Development*, *4*(4), 1051–1075. https://doi.org/10.5194/gmd-4-1051-2011

Cullen, N., Sirguey, P., Mölg, T., Kaser, G., Winkler, M., & Fitzsimons, S. (2013). A century of ice retreat on Kilimanjaro: The mapping reloaded. *The Cryosphere*, *7*(2), 419–431. https://doi.org/10.5194/tc-7-419-2013

Cullen, N. J., Mölg, T., Kaser, G., Steffen, K., & Hardy, D. R. (2007). Energy-balance model validation on the top of Kilimanjaro, Tanzania, using eddy covariance data. *Annals of Glaciology*, *46*, 227–233. https://doi.org/10.3189/172756407782871224

Dix, M., Vohralik, P., Bi, D., Rashid, H., Marsland, S., O'Farrell, S., et al. (2013). The ACCESS coupled model: Documentation of core CMIP5 simulations and initial results. *Australian Meteorological and Oceanographic Journal*, *63*(1), 83–99.

Donner, L. J., Wyman, B. L., Hemler, R. S., Horowitz, L. W., Ming, Y., Zhao, M., et al. (2011). The dynamical core, physical parameterizations, and basic simulation characteristics of the atmospheric component AM3 of the GFDL global coupled model CM3. *Journal of Climate*, *24*(13), 3484–3519. https://doi.org/10.1175/2011JCLI3955.1

Dufresne, J.-L., Foujols, M.-A., Denvil, S., Caubel, A., Marti, O., Aumont, O., et al. (2013). Climate change projections using the IPSL-CM5 Earth System Model: From CMIP3 to CMIP5. *Climate Dynamics*, *40*(9), 2123–2165. https://doi.org/10.1007/s00382-012-1636-1

Dunne, J. P., John, J. G., Adcroft, A. J., Griffies, S. M., Hallberg, R. W., Shevliakova, E., et al. (2012). GFDL's ESM2 global coupled climate–carbon earth system models. Part I: Physical formulation and baseline simulation characteristics. *Journal of Climate*, *25*(19), 6646–6665. https://doi.org/10.1175/JCLI-D-11-00560.1

Dunne, J. P., John, J. G., Shevliakova, E., Stouffer, R. J., Krasting, J. P., Malyshev, S. L., et al. (2013). GFDL's ESM2 global coupled climate–carbon earth system models. Part II: Carbon system formulation and baseline simulation characteristics. *Journal of Climate*, *26*(7), 2247–2267. https://doi.org/10.1175/JCLI-D-12-00150.1

Endris, H. S., Lennard, C., Hewitson, B., Dosio, A., Nikulin, G., & Artan, G. A. (2019). Future changes in rainfall associated with ENSO, IOD and changes in the mean state over Eastern Africa. *Climate Dynamics*, *52*(3), 2029–2053. https://doi.org/10.1007/s00382-018-4239-7

Endris, H. S., Lennard, C., Hewitson, B., Dosio, A., Nikulin, G., & Panitz, H.-J. (2016). Teleconnection responses in multi-GCM driven CORDEX RCMs over Eastern Africa. *Climate Dynamics*, *46*(9), 2821–2846. https://doi.org/10.1007/s00382-015-2734-7

Fischer, E. M., Sedláček, J., Hawkins, E., & Knutti, R. (2014). Models agree on forced response pattern of precipitation and temperature extremes. *Geophysical Research Letters*, *41*, 8554–8562. https://doi.org/10.1002/2014GL062018

Foster, G., & Rahmstorf, S. (2011). Global temperature evolution 1979–2010. *Environmental Research Letters*, *6*(4), 044022. https://doi.org/10.1088/1748-9326/6/4/044022

Gelaro, R., McCarty, W., Suárez, M. J., Todling, R., Molod, A., Takacs, L., et al. (2017). The Modern-Era Retrospective Analysis for Research and Applications, version 2 (MERRA-2). *Journal of Climate*, *30*(14), 5419–5454. https://doi.org/10.1175/JCLI-D-16-0758.1

Gent, P. R., Danabasoglu, G., Donner, L. J., Holland, M. M., Hunke, E. C., Jayne, S. R., et al. (2011). The community climate system model version 4. *Journal of Climate*, *24*(19), 4973–4991. https://doi.org/10.1175/2011JCLI4083.1

Giorgi, F. (2019). Thirty years of regional climate modeling: Where are we and where are we going next? *Journal of Geophysical Research: Atmospheres*, *124*, 5696–5723. https://doi.org/10.1029/2018JD030094

Giorgi, F., & Bates, G. T. (1989). The climatological skill of a regional model over complex terrain. *Monthly Weather Review*, *117*(11), 2325–2347. https://doi.org/10.1175/1520-0493(1989)117(2325:TCSOAR)2.0.CO;2

Giorgi, F., & Francisco, R. (2000). Evaluating uncertainties in the prediction of regional climate change. *Geophysical Research Letters*, *27*(9), 1295–1298. https://doi.org/10.1029/1999GL011016

Griffies, S. M., Winton, M., Donner, L. J., Horowitz, L. W., Downes, S. M., Farneti, R., et al. (2011). The GFDL CM3 coupled climate model: Characteristics of the ocean and sea ice simulations. *Journal of Climate*, *24*(13), 3520–3544. https://doi.org/10.1175/2011JCLI3964.1

Hannachi, A., Jolliffe, I., & Stephenson, D. (2007). Empirical orthogonal functions and related techniques in atmospheric science: A review. *International Journal of Climatology*, *27*(9), 1119–1152. https://doi.org/10.1002/joc.1499

Hardy, D. (2011). Kilimanjaro. In V. P. Singh, P. Singh, & U. K. Haritashya (Eds.), *Encyclopedia of snow, ice and glaciers* (pp. 672–679). Springer.

Harris, I., Osborn, T. J., Jones, P., & Lister, D. (2020). Version 4 of the CRU TS monthly high-resolution gridded multivariate climate dataset. *Scientific Data*, *7*(1), 1–18. https://doi.org/10.1038/s41597-020-0453-3

Hawkins, E., & Sutton, R. (2009). The potential to narrow uncertainty in regional climate predictions. *Bulletin of the American Meteorological Society*, *90*(8), 1095–1108. https://doi.org/10.1175/2009BAMS2607.1

Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz-Sabater, J., et al. (2020). The ERA5 global reanalysis. *Quarterly Journal of the Royal Meteorological Society*, *146*(730), 1999–2049. https://doi.org/10.1002/qj.3803

Hewitson, B., & Crane, R. G. (2006). Consensus between GCM climate change projections with empirical downscaling: Precipitation downscaling over South Africa. *International Journal of Climatology*, *26*(10), 1315–1337. https://doi.org/10.1002/joc.1314

Hirahara, S., Ishii, M., & Fukuda, Y. (2019). Centennial-scale sea surface temperature analysis and its uncertainty. *Journal of Climate*, *32*(16), 57–75. https://doi.org/10.1175/JCLI-D-12-00837.1

Huang, B., Banzon, V. F., Freeman, E., Lawrimore, J., Liu, W., Peterson, T. C., et al. (2015a). *2015: Extended Reconstructed Sea Surface Temperature (ERSST), version 4*. NOAA National Centers for Environmental Information. https://doi.org/10.7289/V5KD1VVF

Huang, B., Banzon, V. F., Freeman, E., Lawrimore, J., Liu, W., Peterson, T. C., et al. (2015b). Extended Reconstructed Sea Surface Temperature version 4 (ERSST. v4). Part I: Upgrades and intercomparisons. *Journal of Climate*, *28*(3), 911–930. https://doi.org/10.1175/JCLI-D-14-00006.1

Im, E.-S., Kang, S., & Eltahir, E. A. (2018). Projections of rising heat stress over the western Maritime Continent from dynamically downscaled climate simulations. *Global and Planetary Change*, *165*, 160–172. https://doi.org/10.1016/j.gloplacha.2018.02.014

Iversen, T., Bentsen, M., Bethke, I., Debernard, J., Kirkevåg, A., Seland, Ø., et al. (2013). The Norwegian earth system model, NorESM1-M—Part 2: Climate response and scenario projections. *Geoscientific Model Development*, *6*(2), 389. https://doi.org/10.5194/gmd-6-389-2013

Jacob, D., Petersen, J., Eggert, B., Alias, A., Christensen, O. B., Bouwer, L. M., et al. (2014). EURO-CORDEX: New high-resolution climate change projections for European impact research. *Regional Environmental Change*, *14*(2), 563–578. https://doi.org/10.1007/s10113-013-0499-2

Jeffrey, S., Rotstayn, L., Collier, M., Dravitzki, S., Hamalainen, C., Moeseneder, C., et al. (2013). Australia's CMIP5 submission using the CSIRO Mk3. 6 model. *Australian Meteorological and Oceanographic Journal*, *63*, 1–13.

Jerez, S., López-Romero, J. M., Turco, M., Lorente-Plazas, R., Gómez-Navarro, J. J., Jiménez-Guerrero, P., & Montávez, J. P. (2020). On the spin-up period in WRF simulations over Europe: Trade-offs between length and seasonality. *Journal of Advances in Modeling Earth Systems*, *12*, e2019MS001945. https://doi.org/10.1029/2019MS001945

Ji, D., Wang, L., Feng, J., Wu, Q., Cheng, H., Zhang, Q., et al. (2014). Description and basic evaluation of Beijing Normal University Earth system model (BNU-ESM) version 1. *Geoscientific Model Development*, *7*(5), 2039–2064. https://doi.org/10.5194/gmd-7-2039-2014

Jobst, A. M., Kingston, D. G., Cullen, N. J., & Schmid, J. (2018). Intercomparison of different uncertainty sources in hydrological climate change projections for an alpine catchment (upper Clutha River, New Zealand). *Hydrology and Earth System Sciences*, *22*(6), 3125. https://doi.org/10.5194/hess-22-3125-2018

Jury, M. W., Prein, A. F., Truhetz, H., & Gobiet, A. (2015). Evaluation of CMIP5 models in the context of dynamical downscaling over Europe. *Journal of Climate*, *28*(14), 5575–5582. https://doi.org/10.1175/JCLI-D-14-00430.1

Kaser, G., Hardy, D. R., Mölg, T., Bradley, R. S., & Hyera, T. M. (2004). Modern glacier retreat on Kilimanjaro as evidence of climate change: Observations and facts. *International Journal of Climatology*, *24*(3), 329–339. https://doi.org/10.1002/joc.1008

Legasa, M., Manzanas, R., Fernández, J., Herrera, S., Iturbide, M., Moufouma-Okia, W., et al. (2020). Assessing multidomain overlaps and grand ensemble generation in CORDEX regional projections. *Geophysical Research Letters*, *47*, e2019GL086799. https://doi.org/10.1029/2019GL086799

Li, L., Lin, P., Yu, Y., Wang, B., Zhou, T., Liu, L., et al. (2013). The Flexible Global Ocean–Atmosphere–Land System model, grid-point version 2: FGOALS-g2. *Advances in Atmospheric Sciences*, *30*(3), 543–560. https://doi.org/10.1007/s00376-012-2140-6

Liu, W., Huang, B., Thorne, P. W., Banzon, V. F., Zhang, H.-M., Freeman, E., et al. (2015). Extended reconstructed sea surface temperature version 4 (ERSST. v4): Part II. Parametric and structural uncertainty estimations. *Journal of Climate*, *28*(3), 931–951. https://doi.org/10.1175/JCLI-D-14-00007.1

Lutz, A. F., ter Maat, H. W., Biemans, H., Shrestha, A. B., Wester, P., & Immerzeel, W. W. (2016). Selecting representative climate models for climate change impact studies: An advanced envelope-based selection approach. *International Journal of Climatology*, *36*(12), 3988–4005. https://doi.org/10.1002/joc.4608

Marchant, R., Mumbi, C., Behera, S., & Yamagata, T. (2007). The Indian Ocean Dipole—The unsung driver of climatic variability in East Africa. *African Journal of Ecology*, *45*(1), 4–16. https://doi.org/10.1111/j.1365-2028.2006.00707.x

Martin, T. H. D. T. G. M., Bellouin, N., Collins, W. J., Culverwell, I. D., Halloran, P. R., Hardiman, S. C., et al. (2011). The HadGEM2 family of Met Office Unified Model climate configurations. *Geoscientific Model Development*, *4*(3), 723–757. https://doi.org/10.5194/gmd-4-723-2011

Maule, C. F., Mendlik, T., & Christensen, O. B. (2017). The effect of the pathway to a two degrees warmer world on the regional temperature change of Europe. *Climate Services*, *7*, 3–11. https://doi.org/10.1016/j.cliser.2016.07.002

McSweeney, C. F., Jones, R. G., & Booth, B. B. (2012). Selecting ensemble members to provide regional climate change information. *Journal of Climate*, *25*(20), 7100–7121. https://doi.org/10.1175/JCLI-D-11-00526.1

McSweeney, C. F., Jones, R. G., Lee, R. W., & Rowell, D. (2015). Selecting CMIP5 GCMs for downscaling over multiple regions. *Climate Dynamics*, *44*(11-12), 3237–3260. https://doi.org/10.1007/s00382-014-2418-8

Mearns, L. O., Arritt, R., Biner, S., Bukovsky, M. S., McGinnis, S., Sain, S., et al. (2012). The North American regional climate change assessment program: Overview of phase I results. *Bulletin of the American Meteorological Society*, *93*(9), 1337–1362. https://doi.org/10.1175/BAMS-D-11-00223.1

Mölg, T., Cullen, N. J., Hardy, D. R., Kaser, G., & Klok, L. (2008). Mass balance of a slope glacier on Kilimanjaro and its sensitivity to climate. *International Journal of Climatology*, *28*(7), 881–892. https://doi.org/10.1002/joc.1589

Mölg, T., Cullen, N. J., Hardy, D. R., Winkler, M., & Kaser, G. (2009). Quantifying climate change in the tropical midtroposphere over East Africa from glacier shrinkage on Kilimanjaro. *Journal of Climate*, *22*(15), 4162–4181. https://doi.org/10.1175/2009JCLI2954.1

Mölg, T., Großhauser, M., Hemp, A., Hofer, M., & Marzeion, B. (2012). Limited forcing of glacier loss through land-cover change on Kilimanjaro. *Nature Climate Change*, *2*(4), 254. https://doi.org/10.1038/nclimate1390

Mölg, T., & Hardy, D. R. (2004). Ablation and associated energy balance of a horizontal glacier surface on Kilimanjaro. *Journal of Geophysical Research*, *109*, D16104. https://doi.org/10.1029/2003JD004338

Mölg, T., Hardy, D. R., Collier, E., Kropač, E., Schmid, C., Cullen, N. J., et al. (2020). Mesoscale atmospheric circulation controls of local meteorological elevation gradients on Kersten Glacier near Kilimanjaro summit. *Earth System Dynamics*, *11*(3), 653–672. https://doi.org/10.5194/esd-11-653-2020

Mölg, T., & Kaser, G. (2011). A new approach to resolving climate–cryosphere relations: Downscaling climate dynamics to glacier-scale mass and energy balance without statistical scale linking. *Journal of Geophysical Research*, *116*, D16101. https://doi.org/10.1029/2011JD015669

Mölg, T., Renold, M., Vuille, M., Cullen, N. J., Stocker, T. F., & Kaser, G. (2006). Indian Ocean zonal mode activity in a multicentury integration of a coupled AOGCM consistent with climate proxy data. *Geophysical Research Letters*, *33*, L18710. https://doi.org/10.1029/2006GL026384

Niang, I., Ruppel, M., Abdrabo, A., Essel, C., Lennard, J., Padgham, J., & Urquhart, P. (2014). Africa. In V. R. Barros, C. B. Field, D. J. Dokken, M. D. Mastrandrea, K. J. Mach & T. E. Bilir, et al. (Eds.), *Climate change 2014: Impacts, adaptation, and vulnerability. Part b: Regional aspects. Contribution of Working Group II to the fifth assessment report of the Intergovernmental Panel on Climate Change* (pp. 1199–1265). Cambridge University Press.

Nicholson, S. E. (2015). Long-term variability of the East African 'short rains' and its links to large-scale factors. *International Journal of Climatology*, *35*(13), 3979–3990. https://doi.org/10.1002/joc.4259

Nicholson, S. E. (2017). Climate and climatic variability of rainfall over eastern Africa. *Reviews of Geophysics*, *55*, 590–635. https://doi.org/10.1002/2016RG000544

North, G. R., Bell, T. L., Cahalan, R. F., & Moeng, F. J. (1982). Sampling errors in the estimation of empirical orthogonal functions. *Monthly Weather Review*, *110*(7), 699–706. https://doi.org/10.1175/1520-0493(1982)110<0699:SEITEO>2.0.CO;2

Ohmura, A. (2012). Enhanced temperature variability in high-altitude climate change. *Theoretical and Applied Climatology*, *110*(4), 499–508. https://doi.org/10.1007/s00704-012-0687-x

Ongoma, V., Chen, H., & Gao, C. (2018). Projected changes in mean rainfall and temperature over East Africa based on CMIP5 models. *International Journal of Climatology*, *38*(3), 1375–1392.

Ongoma, V., Chen, H., & Gao, C. (2019). Evaluation of CMIP5 twentieth century rainfall simulation over the equatorial East Africa. *Theoretical and Applied Climatology*, *135*, 893–910. https://doi.org/10.1007/s00704-018-2392-x

Otieno, G., Mutemi, J., Opijah, F., Ogallo, L., & Omondi, M. (2020). The sensitivity of rainfall characteristics to cumulus parameterization schemes from a WRF model. Part I: A case study over East Africa during wet years. *Pure and Applied Geophysics*, *177*(2), 1095–1110. https://doi.org/10.1007/s00024-019-02293-2

Otte, T. L., Nolte, C. G., Otte, M. J., & Bowden, J. H. (2012). Does nudging squelch the extremes in regional climate modeling? *Journal of Climate*, *25*(20), 7046–7066. https://doi.org/10.1175/JCLI-D-12-00048.1

Overland, J. E., Wang, M., Bond, N. A., Walsh, J. E., Kattsov, V. M., & Chapman, W. L. (2011). Considerations in the selection of global climate models for regional climate projections: The Arctic as a case study. *Journal of Climate*, *24*(6), 1583–1597. https://doi.org/10.1175/2010JCLI3462.1

Pepin, N., Bradley, R., Diaz, H., Baraër, M., Caceres, E., Forsythe, N., et al. (2015). Elevation-dependent warming in mountain regions of the world. *Nature Climate Change*, *5*(5), 424. https://doi.org/10.1038/nclimate2563

Pepin, N., Duane, W., & Hardy, D. (2010). The montane circulation on Kilimanjaro, Tanzania and its relevance for the summit ice fields: Comparison of surface mountain climate with equivalent reanalysis parameters. *Global and Planetary Change*, *74*(2), 61–75. https://doi.org/10.1016/j.gloplacha.2010.08.001

Perkins, S., & Fischer, E. M. (2013). The usefulness of different realizations for the model evaluation of regional trends in heat waves. *Geophysical Research Letters*, *40*, 5793–5797. https://doi.org/10.1002/2013GL057833

Pierce, D. W., Barnett, T. P., Santer, B. D., & Gleckler, P. J. (2009). Selecting global climate models for regional climate change studies. *Proceedings of the National Academy of Sciences*, *106*(21), 8441–8446. https://doi.org/10.1073/pnas.0900094106

Pohl, B., & Camberlin, P. (2011). Intraseasonal and interannual zonal circulations over the equatorial Indian Ocean. *Theoretical and Applied Climatology*, *104*(1-2), 175–191. https://doi.org/10.1007/s00704-010-0336-1

Rodó, X., Pascual, M., Doblas-Reyes, F. J., Gershunov, A., Stone, D. A., Giorgi, F., et al. (2013). Climate change and infectious diseases: Can we meet the needs for better prediction? *Climatic Change*, *118*(3–4), 625–640. https://doi.org/10.1007/s10584-013-0744-1

Rupp, D. E., Abatzoglou, J. T., Hegewisch, K. C., & Mote, P. W. (2013). Evaluation of CMIP5 20th century climate simulations for the Pacific Northwest USA. *Journal of Geophysical Research: Atmospheres*, *118*, 10–884. https://doi.org/10.1002/jgrd.50843

Sakamoto, T. T., Komuro, Y., Nishimura, T., Ishii, M., Tatebe, H., Shiogama, H., et al. (2012). MIROC4h—A new high-resolution atmosphere–ocean coupled general circulation model. *Journal of the Meteorological Society of Japan. Series II*, *90*(3), 325–359. https://doi.org/10.2151/jmsj.2012-301

Schmidt, G. A., Kelley, M., Nazarenko, L., Ruedy, R., Russell, G. L., Aleinov, I., et al. (2014). Configuration and assessment of the GISS ModelE2 contributions to the CMIP5 archive. *Journal of Advances in Modeling Earth Systems*, *6*, 141–184. https://doi.org/10.1002/2013MS000265

Scoccimarro, E., Gualdi, S., Bellucci, A., Sanna, A., Giuseppe Fogli, P., Manzini, E., et al. (2011). Effects of tropical cyclones on ocean heat transport in a high-resolution coupled general circulation model. *Journal of Climate*, *24*(16), 4368–4384. https://doi.org/10.1175/2011JCLI4104.1

Shamir, E., Halper, E., Modrick, T., Georgakakos, K. P., Chang, H.-I., Lahmers, T. M., & Castro, C. (2019). Statistical and dynamical downscaling impact on projected hydrologic assessment in arid environment: A case study from Bill Williams River basin and Alamo Lake, Arizona. *Journal of Hydrology X*, *2*, 100019. https://doi.org/10.1016/j.hydroa.2019.100019

Simmons, A., Berrisford, P., Dee, D., Hersbach, H., Hirahara, S., & Thépaut, J.-N. (2017). A reassessment of temperature variations and trends from global reanalyses and monthly surface climatological datasets. *Quarterly Journal of the Royal Meteorological Society*, *143*(702), 101–119. https://doi.org/10.1002/qj.2949

Smith, T. M., Reynolds, R. W., Peterson, T. C., & Lawrimore, J. (2008). Improvements to NOAA's historical merged land–ocean surface temperature analysis (1880–2006). *Journal of Climate*, *21*(10), 2283–2296. https://doi.org/10.1175/2007JCLI2100.1

Sun, J., Fu, J. S., Huang, K., & Gao, Y. (2015). Estimation of future PM2.5- and ozone-related mortality over the continental United States in a changing climate: An application of high-resolution dynamical downscaling technique. *Journal of the Air & Waste Management Association*, *65*(5), 611–623. https://doi.org/10.1080/10962247.2015.1033068

Sun, L., Semazzi, F. H., Giorgi, F., & Ogallo, L. (1999). Application of the NCAR regional climate model to eastern Africa: 1. Simulation of the short rains of 1988. *Journal of Geophysical Research*, *104*(D6), 6529–6548. https://doi.org/10.1029/1998JD200051

Taylor, K. E., Stouffer, R. J., & Meehl, G. A. (2012). An overview of CMIP5 and the experiment design. *Bulletin of the American Meteorological Society*, *93*(4), 485. https://doi.org/10.1175/BAMS-D-11-00094.1

Thielke, A., & Mölg, T. (2019). Observed and simulated Indian Ocean Dipole activity since the mid-19th century and its relation to East African short rains. *International Journal of Climatology*, *39*(11), 4467–4478. https://doi.org/10.1002/joc.6085

Voldoire, A., Sanchez-Gomez, E., y Mélia, D. S., Decharme, B., Cassou, C., Sénési, S., et al. (2013). The CNRM-CM5.1 global climate model: Description and basic evaluation. *Climate Dynamics*, *40*(9–10), 2091–2121. https://doi.org/10.1007/s00382-011-1259-y

von Salzen, K., Scinocca, J. F., McFarlane, N. A., Li, J., Cole, J. N. S., Plummer, D., et al. (2013). The Canadian Fourth Generation Atmospheric Global Climate Model (CanAM4). Part I: Representation of physical processes. *Atmosphere-Ocean*, *51*(1), 104–125. https://doi.org/10.1080/07055900.2012.755610

Washington, R., Harrison, M., Conway, D., Black, E., Challinor, A., Grimes, D., et al. (2006). African climate change: Taking the shorter route. *Bulletin of the American Meteorological Society*, *87*(10), 1355–1366. https://doi.org/10.1175/BAMS-87-10-1355

Watanabe, M., Suzuki, T., O'ishi, R., Komuro, Y., Watanabe, S., Emori, S., et al. (2010). Improved climate simulation by MIROC5: Mean states, variability, and climate sensitivity. *Journal of Climate*, *23*(23), 6312–6335. https://doi.org/10.1175/2010JCLI3679.1

Watanabe, S., Hajima, T., Sudo, K., Nagashima, T., Takemura, T., Okajima, H., et al. (2011). MIROC-ESM 2010: Model description and basic results of CMIP5-20c3m experiments. *Geoscientific Model Development*, *4*(4), 845. https://doi.org/10.5194/gmd-4-845-2011

Wilks, D. (2006). *Statistical methods in the atmospheric sciences* (2nd ed.). Academic Press.

Winkler, M., Kaser, G., Cullen, N. J., Mölg, T., Hardy, D. R., & Pfeffer, W. T. (2010). Land-based marginal ice cliffs: Focus on Kilimanjaro. *Erdkunde*, *64*(2), 179–193. https://doi.org/10.3112/erdkunde.2010.02.05

Wu, T., Song, L., Li, W., Zhang, H., Xin, X., Zhang, Y., et al. (2014). An overview of BCC climate system model development and application for climate change studies. *Journal of Meteorological Research*, *28*(1), 34–56. https://doi.org/10.1007/s13351-014-3041-7

Xue, Y., Smith, T. M., & Reynolds, R. W. (2003). Interdecadal changes of 30-yr SST normals during 1871–2000. *Journal of Climate*, *16*(10), 1601–1612. https://doi.org/10.1175/1520-0442-16.10.1601

Yang, W., Seager, R., Cane, M. A., & Lyon, B. (2015). The rainfall annual cycle bias over East Africa in CMIP5 coupled climate models. *Journal of Climate*, *28*(24), 9789–9802. https://doi.org/10.1175/JCLI-D-15-0323.1

Yukimoto, S., Adachi, Y., Hosaka, M., Sakami, T., Yoshimura, H., Hirabara, M., et al. (2012). A new global climate model of the Meteorological Research Institute: MRI-CGCM3—Model description and basic performance. *Journal of the Meteorological Society of Japan*, *90A*, 23–64. https://doi.org/10.2151/jmsj.2012-A02