The Link between Gulf Stream Precipitation Extremes and European Blocking in General Circulation Models and the Role of Horizontal Resolution

KRISTIAN STROMMEN,[a] SIMON L. L. MICHEL,[a] AND HANNAH M. CHRISTENSEN[a]

[a] *Department of Physics, University of Oxford, Oxford, United Kingdom*

(Manuscript received 4 June 2024, in final form 3 March 2025, accepted 22 April 2025)

ABSTRACT: Past studies show that coupled model biases in European blocking and North Atlantic eddy-driven jet variability decrease as one increases the horizontal resolution in the atmospheric and oceanic model components, but it remains unclear if atmospheric or oceanic resolution plays a greater role and why. Here, following recent work by Schemm et al., we leverage a large multimodel ensemble to show that a coupled model's ability to simulate extreme Gulf Stream precipitation is tightly linked to its simulated frequency of European blocking and northern jet excursions. Furthermore, the reduced biases in blocking and jet variability are consistent with better resolved precipitation extrema in high-resolution models. The analysis supports a hypothesis that models which simulate more extreme precipitation can generate more strongly poleward-propagating cyclones and more intense anticyclonic anomalies due to the stronger latent heat release occurring during extreme events. By contrast, typical North Atlantic SST biases are found to share only a weak or negligible relationship with blocking and jet biases. Finally, while previous studies have used a comparison between coupled models and models run with prescribed SSTs to argue for the role of ocean resolution, we emphasize here that models run with prescribed SSTs experience greatly reduced precipitation extremes due to their excessive thermal damping, making it unclear if such a comparison is meaningful. Instead, we speculate that most of the reduction in coupled model biases may actually be due to increased atmospheric resolution leading to better resolved convection.

KEYWORDS: Atmosphere-ocean interaction; Blocking; Extratropical cyclones; Precipitation; Sea surface temperature; Climate models

1. Introduction

Accurately simulating the variability of the North Atlantic eddy-driven jet and northern midlatitude blocking events has been a longstanding challenge in climate science and numerical weather prediction. Considerable progress has been made on this challenge in the last 30 years (Davini and D'Andrea 2016; Davini and d'Andrea 2020; Dorrington et al. 2022), at least partially as a result of the increased horizontal resolution of models (Davini and d'Andrea 2020; Schiemann et al. 2020). However, models still exhibit systematic biases. A notable and well-known example concerns so-called European blocking; following Athanasiadis et al. (2022), this refers to blocking in the region 30°W–15°E, 45°–65°N. Composites of the circulation during European blocking events using ERA5 reanalysis data (Hersbach et al. 2020) are shown in Fig. 1. These highlight the characteristic features of European blocking: a high-pressure anomaly over northwest Europe/Scandinavia and a northward displacement of the eddy-driven jet, corresponding to an anticyclonic Rossby wavebreaking event. Models systematically underestimate the occurrence of European blocking events (Davini and d'Andrea 2020; Schiemann et al. 2020), and related northern jet excursions (Dorrington et al. 2022), leading to uncertainty in future projections of how such events will change with global warming (Woollings et al. 2018a). Weather forecasts also still struggle to accurately predict European blocking events (Matsueda and Palmer 2018; Büeler et al. 2021). Since European blocking events are closely linked to extreme Euro-Atlantic surface weather (Kautz et al. 2022), a lot of effort has been devoted to understanding the origin of model biases and the exact role played by model resolution. The purpose of this paper is to contribute toward this effort, in particular the role of horizontal model resolution. The vertical resolution has also been shown to affect blocking and jet variability (Anstey et al. 2013), hypothetically due to enhanced stratosphere–troposphere coupling (Scaife and Knight 2008), but we do not discuss this important topic here. For pointers toward more theoretical work on blocking, the reader may refer to, e.g., Woollings et al. (2018a).

Much of the past literature on biases in European blocking and North Atlantic jet variability has focused on the role of North Atlantic sea surface temperature (SST) biases in models, especially in the Gulf Stream region, which for us will always mean the region 70°–50°W, 35°–45°N, highlighted in Fig. 1a. The separation of the Gulf Stream from the North American continent creates a sharp SST gradient that can affect downstream jet and blocking variability by modulating surface wind convergence (Minobe et al. 2008), near-surface baroclinicity (Brayshaw et al. 2011), and meridional heat transport (Ambaum and Novak 2014; O'Reilly et al. 2016;

Denotes content that is immediately available upon publication as open access.

Supplemental information related to this paper is available at the Journals Online website: https://doi.org/10.1175/JCLI-D-24-0311.s1.

Corresponding author: Kristian Strommen, kristian.strommen@physics.ox.ac.uk

DOI: 10.1175/JCLI-D-24-0311.1

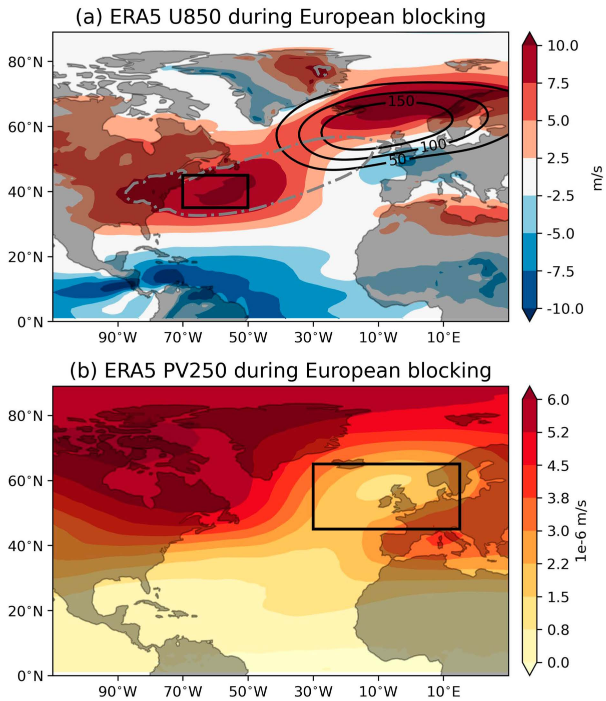

FIG. 1. (a) Zonal 850-hPa winds (filled contours), and 500-hPa geopotential height (black line contours), averaged over European blocking days taking place in winters 1979–2014. The gray dashed–dotted contour shows the 8–12 m s^{-1} contour of climatological zonal winds. (b) 250-hPa PV during European blocking. The black box in (a) highlights the Gulf Stream region 70°–50°W, 35°–45°N, while the box in (b) highlights the domain of European blocking: 30°W–15°E, 45°–65°N. See methods for details on how the European blocking index is defined. Data from ERA5.

Novak et al. 2017; O'Reilly et al. 2017). Following its separation from the continent, the Gulf Stream Current turns northward due to bottom topography, before veering east again toward Europe (the so-called northwest corner), with important consequences for the mean state of the SSTs in the central North Atlantic (Keeley et al. 2012; Drews et al. 2015). However, the Gulf Stream in models fails to separate early enough from the continent (Chassignet and Marshall 2008; Moreno-Chamarro et al. 2022; Tsartsali et al. 2022), leading to an overly smooth SST gradient (O'Reilly et al. 2016). Models also fail to simulate the "northwest corner," and instead simulate a Gulf Stream Current that extends zonally from Newfoundland to Europe, resulting in North Atlantic SSTs exhibiting a cold bias that can be as great as 10°C (Keeley et al. 2012; Wang et al. 2014; Drews et al. 2015). Recently, Athanasiadis et al. (2022) also argued that models show an overly *sharp* SST gradient in the central North Atlantic compared to observations, due to the failure of models to simulate the northwest corner, and that the associated bias in near-surface baroclinicity is partly responsible for European blocking biases.

Because it is known that correctly simulating the pathway of the Gulf Stream depends at least in part on the ocean model resolution (Chassignet and Marshall 2008; Tsartsali et al. 2022), it is natural to hypothesize that much of the bias in European blocking and jet variability is a result of insufficient horizontal resolution in models, especially in the ocean component. This hypothesis has been explicitly made and argued for in some select single-model studies, such as Scaife et al. (2011), and more recently in the three multimodel studies Davini and d'Andrea (2020), Athanasiadis et al. (2022), and Michel et al. (2023). Since the atmospheric and oceanic horizontal resolutions are typically increased in tandem, a clean attribution of their relative roles cannot be immediately made. Nevertheless, Scaife et al. (2011), Athanasiadis et al. (2022), and Michel et al. (2023) all argue that while atmospheric resolution plays a role, ocean resolution is more important, with a key piece of evidence in all cases being a comparison between coupled models and models forced with prescribed observed SSTs (which we henceforth refer to as AMIP models). Specifically, they find that the bias reduction seen with increased resolution is either absent or much smaller in AMIP models compared to coupled models and interpret this as evidence for the importance of ocean resolution. However, the evidence here does not seem completely conclusive. First, the classic work of Barsugli and Battisti (1998) on air–sea coupling suggests that differences between AMIP and coupled models could be hard to interpret, due to the excessive thermal damping experienced by AMIP models. More recently, Mathews and Czaja (2024) have emphasized the role of air–sea coupling to blocking maintenance, which, if correct, would result in further systematic differences between AMIP and coupled models.

Second, AMIP models broadly have similar levels of biases in blocking compared to coupled models (Schiemann et al. 2020), despite having no SST biases whatsoever. Third, some of the earlier studies arguing for the role of the ocean are based on very small ensembles, such as Scaife et al. (2011) (a single pair of low and high resolution simulations), which we now know to be inadequate in the face of the small signals and large internal variability characteristic of the North Atlantic circulation (Scaife and Smith 2018). In some cases, it is known that hundreds of ensemble members are needed to robustly detect forced signals in the North Atlantic (Ye et al. 2024), making it unclear if even the 17 ensemble members used in Athanasiadis et al. (2022) are sufficient.

In recent years, a growing body of literature has highlighted the crucial role played by *diabatic processes* in the Gulf Stream region in the formation and maintenance of blocking (Pfahl et al. 2015; Steinfeld and Pfahl 2019; Steinfeld et al. 2020; Yamamoto et al. 2021; Athanasiadis et al. 2022; Wenta et al. 2024; Mathews and Czaja 2024), adding to earlier literature on the importance of such processes on the North Atlantic storm track (Hoskins and Valdes 1990; Chang et al. 2002). The recent work of Mathews et al. (2024) adds compelling evidence toward this, by showing that forcibly suppressing the Gulf Stream moisture fluxes in a model leads to a large reduction in blocking frequencies across the entire Northern Hemisphere. In this vein, Schemm (2023) performed AMIP aquaplanet simulations with a regionally varying horizontal atmospheric resolution, where the resolution is around 10 km in the storm-track region and around 5 km in the Gulf Stream region. Schemm finds that this increased resolution leads to enhanced precipitation rates in the Gulf Stream region, greater northward propagation of cyclones, and an associated jet which is more tilted and northward. Schemm therefore hypothesizes that the "too zonal, too equatorward" bias in the North Atlantic jet seen in models is a direct result of inadequately resolved diabatic processes in the Gulf Stream region and that this can be alleviated with increased horizontal resolution. It is natural, based on the crucial role played by cyclones in creating and maintaining blocking anticyclones (Colucci 1985; Nakamura and Wallace 1993; Yamazaki and Itoh 2009; Steinfeld and Pfahl 2019; Wenta et al. 2024), to hypothesize that this could also explain model biases in European blocking. The follow-up study by De Luca et al. (2024) looked at the same AMIP simulations used in Schemm (2023) and indeed found that the increased resolution leads to enhanced blocking frequencies downstream. They hypothesize that improved diabatic processes in the Gulf Stream region are responsible.

Schemm's experiments in these two studies only used a single model over a small range of resolutions and did not include orography. A key motivation of the present study is to test the hypotheses from Schemm (2023) and De Luca et al. (2024) ("Schemm's hypothesis" for short) in a large ensemble of coupled models spanning a wide range of resolutions including orography. In fact, we will slightly sharpen Schemm's hypothesis by emphasizing the role of *extreme* precipitation in the Gulf Stream region. As will be seen, models do not show a systematic bias in the mean precipitation of the Gulf Stream region, and so such biases cannot clearly explain the systematic bias in blocking frequencies. However, as one might expect, models do systematically underestimate the magnitude of extreme precipitation. The hypothesis we will explore, therefore, is whether these two systematic underestimations, of extreme precipitation on the one hand and European blocking/northward jet excursions on the other hand, are related. One clue that this might be the case is that precipitation in the Gulf Stream region varies intermittently between periods of low rainfall (drizzle) followed by particularly intense precipitation associated with cyclone fronts (Catto et al. 2012; Messori and Czaja 2013). Another clue is already present in Schemm's work, which showed that the impact of increased resolution in his model was to enhance the northward propagation of only the very largest cyclones (those in the upper 99th percentile), with the mean propagation remaining unchanged. This suggests that when measuring the links between Gulf Stream precipitation, blocking/jet variability, and resolution, it may be important to measure higher-order moments of the relevant distributions in some way. Schemm did this by looking in the tails of the cyclone distribution, where precipitation might be expected to be particularly intense. We will test this hypothesis against a large coupled multimodel ensemble (135 members), spanning a wide range of model resolutions (from 400- to 25-km atmospheric grid spacing at the equator). Concretely, we will use the ensemble to address the following questions:

1) Is the ability to simulate extreme Gulf Stream precipitation clearly linked to the mean frequency of occurrence of European blocking events and northward eddy-driven jet excursions in models?
2) To what extent does atmospheric horizontal resolution affect Gulf Stream precipitation extremes, and can the improved representation of European blocking and northward jet excursions with model resolution be explained by improved precipitation extremes?

Additionally, drawing on Barsugli and Battisti (1998), we will examine a third point:

3) How does Gulf Stream precipitation variability compare between coupled and AMIP models, and what can we say about the relative role of atmospheric versus oceanic resolution in reducing biases in European blocking?

As part of our discussion on the role of atmospheric versus oceanic resolution, we also look at how SST biases in our ensemble are linked to blocking, including revisiting the hypothesis of Athanasiadis et al. (2022).

Finally, we emphasize that Davini et al. (2022) highlighted that improvements to orography can play a role in improving European blocking that goes beyond that obtained purely by increasing the atmospheric resolution. We will not examine the contribution of orography in our paper.

2. Data

We analyze daily data drawn from a total of 135 historical model simulations, consisting of multiple ensemble members drawn from a range of CMIP6 models (Eyring et al. 2016), CMIP5 models (Taylor et al. 2012), and High-Resolution Model Intercomparison Project (HighResMIP) models (Haarsma et al. 2016). The CMIP5 and CMIP6 historical experiments consist of coupled uninitialized climate runs forced with historical greenhouse gas and aerosol forcings over the twentieth century, after a spinup from a free-running preindustrial control run. The CMIP5 simulations span either 1900–2005 or 1950–2005, while the CMIP6 simulations span 1900–2015 or 1950–2015. The HighResMIP models are initialized in 1950, following a short 50-yr spinup, and they span the 65 years between 1950 and 2015 and use the same historical forcings as CMIP6. The data are in all cases restricted to the December–February seasons between 1979–01 and 2014–12. Thus, CMIP5 data cover 1979–2005 and CMIP6 and HighResMIP cover 1979–2014. The HighResMIP models are sometimes referred to as the "Process-Based Climate Simulation: Advances in High-Resolution Modeling and European Climate Risk Assessment (PRIMAVERA)" models, in reference to the European Union Project under which they were produced. Tables detailing the exact models and ensemble members used are given (Tables S1–S3 in the online supplemental material).

As our observational estimate, we use ERA5 (Hersbach et al. 2020) for SSTs, 850-hPa winds (U850), 500-hPa geopotential height (Z500), and 250-hPa potential vorticity (PV250). For precipitation, it is crucial to avoid using precipitation from ERA5 when considering moments beyond the mean. Reanalysis rainfall is based on model output, and thus suffers from the typical "drizzle bias" of models, whereby they overestimate low precipitation events and underestimate high precipitation events (Chen et al. 2021). Instead, we use the NASA/Goddard Space Flight Center satellite product IMERG version 7, hereafter IMERG (Huffman et al. 2020, 2023). This is a greatly updated version of the now defunct TRMM/3B42 satellite dataset (Kawanishi et al. 2000). IMERG has a native spatial resolution of 0.1° (10 km) and estimates the rainfall every 30 min by postprocessing the output of a number of satellites. Extensive discussion on the strengths and limitations of IMERG can be found at Huffman and Tan (2023). IMERG covers the period 2000 onward, so for consistency both IMERG and ERA5 data are restricted to the period 2000–15. The blocking and jet metrics we consider (see next section) are very similar in ERA5 whether computed over this period or the longer period 1979–2015, with differences on the order of 5%. The precipitation metrics in the model ensemble are also nearly identical if computed over the period 1979–2015 or 2000–15, owing to our use of daily data (not shown). Thus, while the short coverage of IMERG does introduce some unavoidable uncertainty, the restriction does not obviously affect our analysis.

The question of how to assess instrument and methodological uncertainty in a dataset like IMERG is challenging (Huffman and Tan 2023). Supporting Fig. S1 shows that the distributions of daily Gulf Stream precipitation differ notably between TRMM/3B42 and IMERG. This is especially true in the tails, where we found that the 99th percentile in TRMM is around 45% greater than in IMERG, a far greater difference than what can be explained by sampling variability alone. The same figure highlights the underestimation of extremes by ERA5. Figure S1 thus suggests that the uncertainty in IMERG could be large. To highlight this uncertainty, we will at times discuss the differences one would obtain in our analysis if we replaced IMERG with TRMM. TRMM covers 1998–2015, with a native spatial resolution of 0.25°. We restrict TRMM to 2000–15 also, though note that using 1998–2015 would give almost identical results.

Data are always restricted to the DJF season. All Z500, U850, and SST data are regridded to a regular 2.5° grid. Precipitation data are regridded to a regular 1° grid. When computing SST gradients, SST data are regridded to a regular 0.5° grid, to more accurately distinguish between low- and high-resolution models.

Finally, we note that for a few select metrics, we were unable to obtain the necessary data for one or two models. Thus, the sample size of models varies from 135 to 130 across the various figures. Due to the very high sample size, this effect is negligible, so we do not draw further attention to this.

3. Methods

a. *Definition of indices and metrics*

Our main goal is to compute correlations capturing the intermodel spread of climatological values. In other words, we will compute correlations in scatterplots where each point represents a single model (represented by its climate mean value). In this context, the presence of trends in the underlying time series is simply assumed to contribute to the climate mean values, and we make no attempt at untangling forced and unforced contributions to these or characterizing differences in trends across models. No trend removal of any kind is therefore performed for the indices/metrics that follow.

1) European blocking

For all datasets studied here, we use the same blocking index as Athanasiadis et al. (2022), which consists of spatiotemporally averaged instantaneous blocking conditions in the area 30°W–15°E, 45°–65°N over all DJF seasons in the historical period 1979–2014. This region will be highlighted in a relevant figure.

Instantaneous blocking conditions for each grid point are determined from 500-hPa geopotential height (Z500) fields where two conditions are checked to detect a strong reversal in the meridional gradient—a typical signature of blocking (Scherrer et al. 2006; Davini and d'Andrea 2020). Therefore, on any given day, we consider that a given grid point with longitude λ_0 and latitude ϕ_0 is blocked if two conditions C_1 and C_2 are met:

$$C_1 : \mathrm{GHGS}(\lambda_0, \phi_0) > 0$$
$$C_2 : \mathrm{GHGN}(\lambda_0, \phi_0) < -10.$$

Here, geopotential height-gradient south (GHGS) and geopotential height-gradient north (GHGN), expressed in meters per latitudinal degree, are defined by

$$\mathrm{GHGS}(\lambda_0, \phi_0) = \frac{Z500(\lambda_0, \phi_0) - Z500(\lambda_0, \phi_S)}{\phi_0 - \phi_S}$$
$$\mathrm{GHGN}(\lambda_0, \phi_0) = \frac{Z500(\lambda_0, \phi_{\mathrm{text}N}) - Z500(\lambda_0, \phi_0)}{\phi_N - \phi_0},$$

with $\phi_S = \phi_0 - 15°$ and $\phi_N = \phi_0 + 15°$. Consistent with the study area from Athanasiadis et al. (2022), λ_0 here ranges from 30°W to 15°E and ϕ_0 ranges from 45° to 65°N. The frequency of European blocking is the fraction of DJF days for which blocking occurs somewhere in this domain.

When creating composite maps based on European blocking, it is helpful to restrict to a more temporally and spatially coherent definition of such events, as opposed to just averaging across all days for which blocking occurs at some grid point in the domain. Therefore, to detect whether there is a spatially consistent blocking event in the area of interest on a given day, we apply k-means clustering to the blocking index maps for each time step of the dataset. In this case, k means is set with $k = 2$, clustering time steps based on the spatial distribution of blocked grid cells (ones for blocked and zeros for not blocked) according to the 2D blocking index definition described above.

Let $\mathbf{X} = (x_{ij})_{i,j} \in \mathbb{R}^{n\times p}$, where n is the number of time steps and p is the number of grid cells. The k-means clustering process follows these steps:

1) Randomly initialize k cluster centers, denoted as $\mu_1^{(1)}, \ldots, \mu_k^{(1)}$.

2) Repeat until convergence (i.e., when cluster assignments no longer change):

   ○ Assign each data point (row of the **X** matrix) to the nearest cluster c_j, $j = 1, \ldots, k$, based on Euclidean distance.
   ○ Update the cluster centers using the formula:

$$\mu_j^{(t+1)} = \frac{1}{\{c_j^{(t)}\}} \sum_{i \in c_j^{(t)}} x_i, \quad j = 1, \ldots, k.$$

This produces a time-series vector of zeros (no coherent block) and ones (coherent block) for the region of interest. To ensure the stability of the k-means clustering for each dataset (because outputs depend on random initialization), the process is repeated 100 times, and the result that minimizes the sum of distances to the cluster centers is selected. Taking composites of the original blocking index based on days when a coherent block is present gives the expected result, namely, a spatially coherent region of blocked grid points centered in the European domain (not shown).

2) Northward jet excursions

Northward jet excursions are defined using the eddy-driven jet latitude index first introduced by Woollings et al. (2010), which we compute using the simplified methodology of Parker et al. (2019). Daily mean zonal winds at 850 hPa in the DJF season are first regridded to a common 1° grid and then zonally averaged over the Atlantic domain 0°–60°W, 15°–75°N. A 5-day low-pass filter is then applied to remove synoptic variability. For each day, the jet latitude is now defined as the latitude at which the zonally averaged winds are maximum in these smoothed zonal averages. The histogram of the jet latitude is trimodal (Woollings et al. 2010), including in the majority of the CMIP6 generation models (Dorrington et al. 2022). A northward jet day is then defined, for all models and reanalysis, as one where the jet latitude exceeds 52°N, which spans the approximate range of the northernmost peak in a typical trimodal jet latitude histogram. From this, we can compute a climatological frequency of northern jet day occurrence over all DJF seasons in the historical period 1979–2014; this will sometimes be referred to as "NorthDays" for short.

It should be emphasized that this is a fairly crude definition, in that it fails to account for split jets, days with insignificant jet activity, or information about the meridional tilt of the jet. However, since the goal here is to understand the impact of Gulf Stream precipitation on "northward jet excursions" only in a broad, climate mean-state sense, ignoring such finer-scale structure is not unreasonable.

3) Gulf Stream precipitation extremes

We follow Schemm (2023) and define the Gulf Stream region to be the rectangular box 70°–50°W, 35°–45°N. This region will be highlighted in relevant figures. As in Schemm (2023), precipitation here is to be thought of as a proxy for latent heat release by moist diabatic processes. It is possible that more sophisticated measures would provide better results. For example, the amount of latent heating does not always match the total precipitation due to the possibility of reevaporation, something which could be accounted for by measuring precipitation efficiency rather than precipitation. However, using raw precipitation as a simple proxy already has great explanatory power, so we did not pursue any such increased sophistication.

In this work, we define our measure of Gulf Stream precipitation extremes (referred to as GS PR99 for short) as being the 99th percentile of daily precipitation, as computed across all grid points in the Gulf Stream region and all days in all DJF seasons in the historical period 1979–2014. In other words, we compute the 99th percentile of the Gulf Stream time series of length time × latitude × longitude obtained by collapsing the three-dimensional (time, latitude, longitude) array of daily DJF Gulf Stream precipitation to a one-dimensional array. Models which have higher values of this percentile are ones which can simulate more extreme daily rainfall in the Gulf Stream domain.

Qualitatively, very similar results were obtained if we rather measured extremes using (i) the 90th or 95th percentile; (ii) the maximum monthly precipitation obtained at any of the Gulf Stream grid points in any of the December–February months (PRmax); or (iii) the mean number of "extreme" precipitation days (days for which any of the grid points in the Gulf Stream region experiences daily precipitation exceeding 40 mm) in the DJF season. Similar results can also be obtained by using the standard deviation of daily precipitation. We also obtained almost identical results if the percentiles were computed using the daily, area-averaged precipitation. The 99th percentile was chosen because it results in strong signals when plotting composite anomalies. Future studies might clarify if more principled choices are possible. We have also tested the effect of using the mean Gulf Stream precipitation instead, and we discuss the differences in the results.

Finally, we found that computing the percentiles using data on the native model grid, as opposed to regridded data, has a negligible impact. Our results are therefore not sensitive to this choice. This is plausibly related to the fact that the size of the Gulf Stream domain is small relative to the typical grid spacing of the models considered.

4) Rossby wavebreaking data

To support our proposed mechanism, we will examine how the intensity of anticyclonic wavebreaking in ERA5 depends on the intensity of Gulf Stream precipitation. To do this, we make use of a wavebreaking dataset for ERA5 generated using software developed by Kaderli (2023). The identification of a wavebreaking event is based on analyzing the dynamical tropopause [defined as the 2 PV unit contour; 1 PV unit (PVU) = 10^{-6} K kg^{-1} m^2 s^{-1}] and tracking PV streamers (Wernli and Sprenger 2007; Sprenger et al. 2017). A streamer is an elongated feature of the 2-PV contour which extends away from the main contour: In Fig. 1b, this is the high-PV band sitting under the low-PV "crest" of the wave. More precisely, it is characterized in terms of two contour points that are close geographically speaking but far apart as measured along the 2-PV contour. The intensity of a wavebreaking event is defined as the average momentum flux in the streamer associated with the event.

5) Atmospheric horizontal resolution

To measure model resolution, we focus on atmospheric horizontal resolution, because oceanic resolution is strongly correlated with atmospheric resolution across models (i.e., modeling centers tend to increase both in tandem) and atmospheric resolution varies more finely across models. We use the nominal grid spacing at the equator as our measure of nominal atmospheric resolution. The nominal resolutions were obtained by reading the relevant model implementation paper for all the models, converting from spectral resolution to grid spacing where necessary. The results are summarized in Supporting Tables S1–S3. We note that in a few limited cases we were unable to find a written quote about a specific model version. In these cases, we assumed the resolution matched that of the most closely matching model version.

b. Estimating internal variability of blocking and jet variability

The North Atlantic jet and European blocking are known to exhibit notable decadal variability in both observations and reanalysis (Hurrell and Van Loon 1997; Woollings et al. 2018b; Dorrington et al. 2022). This can complicate questions around model biases, since models may show discrepancies from observations due to inhabiting different modes of unforced ("internal") decadal variability from the real world, independently of any failure to resolve the physical processes correctly. For example, the apparent underestimation of blocking frequencies in models could, in principle, simply be a result of the real world randomly undergoing a period of higher-than-average blocking in recent decades, which models would not on the whole be expected to capture. It is thus important to include some measure of this internal variability in observational estimates of climatological European blocking frequencies and northward jet excursions. We do this here by a simple bootstrap procedure, whereby we resample with replacement blocking/northward jet frequencies from the 15 winters spanning 2000—in order 15, to generate 1000 random time series (interpreted as random variations of the period 2000–15), and define the magnitude of internal variability to be twice the standard deviation of the resulting distribution. Estimating internal variability by rather selecting random chunks of years (of the correct size) from 1900 to 2010 using the twentieth-century reanalysis product ERA20C (Poli et al. 2013) yielded similar results (not shown).

Note that the analysis of Woollings et al. (2018b) suggests that recent real-world blocking frequencies are if anything on the lower end of what might be expected from real-world decadal variability (cf. Fig. 11 therein). Thus, the fact that models seem to underestimate European blocking in the recent period is unlikely to be an artifact of decadal variability.

c. Statistical analysis

Because of the non-Gaussianity of the distribution of precipitation, correlations are always computed as Spearman's rank correlations, using the spearmanr function from the Python package SciPy (Virtanen et al. 2020). However, despite the non-Gaussianity of some of the individual distributions considered, the observed relationships between variables are in almost all cases approximately linear, and so using Pearson's correlations gave virtually identical results in almost all cases. The one exception here is the observed link between precipitation extremes and atmospheric horizontal resolution, which shows evidence of nonlinearity. In this case, the Pearson correlation appears to underestimate the observed link, with the Spearman correlation being higher.

To estimate the statistical significance of any computed correlations, we almost always use bootstrap resampling, to avoid the use of standard tests that assume Gaussianity. Concretely, given samples x and y and a null hypothesis of there being no relationship, we resample x and y randomly with replacement 10 000 times and correlate the resulting random samples to obtain a distribution of correlations for the null hypothesis. Given our typical sample size of $N = 135$, this method always produced a 95% confidence interval for the null hypothesis of around ± 0.19 and a 99% interval of around ± 0.22. We thus interpret correlations exceeding 0.2 in magnitude as unlikely to be random. The exception to our use of bootstrap resampling is when testing for the effect of mediation, where we use an explicit Sobel test (Baron and Kenny 1986).

A weakness of our statistical analysis is that we deliberately do not distinguish between simulations performed by different models and different ensemble members simulated using a single model. This is done to increase the sample size as much as possible and to ensure any links uncovered are robust to the large internal variability ensembles using single models can exhibit (Scaife and Smith 2018). The drawback is that ensemble size varies a lot between the models, meaning that estimates may sometimes be affected by, for example, the presence of a particularly large single-model ensemble at a particular horizontal resolution. Visual inspection of the relevant scatterplots suggests that this effect is probably small, and we confirmed that correlations do not substantially change if only a single ensemble member is retained for each model or if the ensemble mean averages are used for models with multiple realizations. It is worth remarking that different climate models are not generally independent of each other either, and this can inherently bias any multimodel analysis, irrespective of the number of ensemble members retained (Knutti et al. 2013; Annan and Hargreaves 2017; Boé 2018).

4. Results

a. Model biases in the mean and extremes of key variables

For context, we start by visualizing biases of key quantities in the multimodel ensemble. Note that similar multimodel analysis has been carried out in several previous papers, including for precipitation (Yazdandoost et al. 2021; Abdelmoaty et al. 2021), SSTs (Zhang et al. 2023), and zonal winds (Harvey et al. 2020; Athanasiadis et al. 2022). Our analysis is included here anyway to make the present paper self-contained.

Figures 2a–c show the mean monthly precipitation for the multimodel mean, IMERG, and the multimodel bias (multimodel mean minus IMERG). The SST contours for the

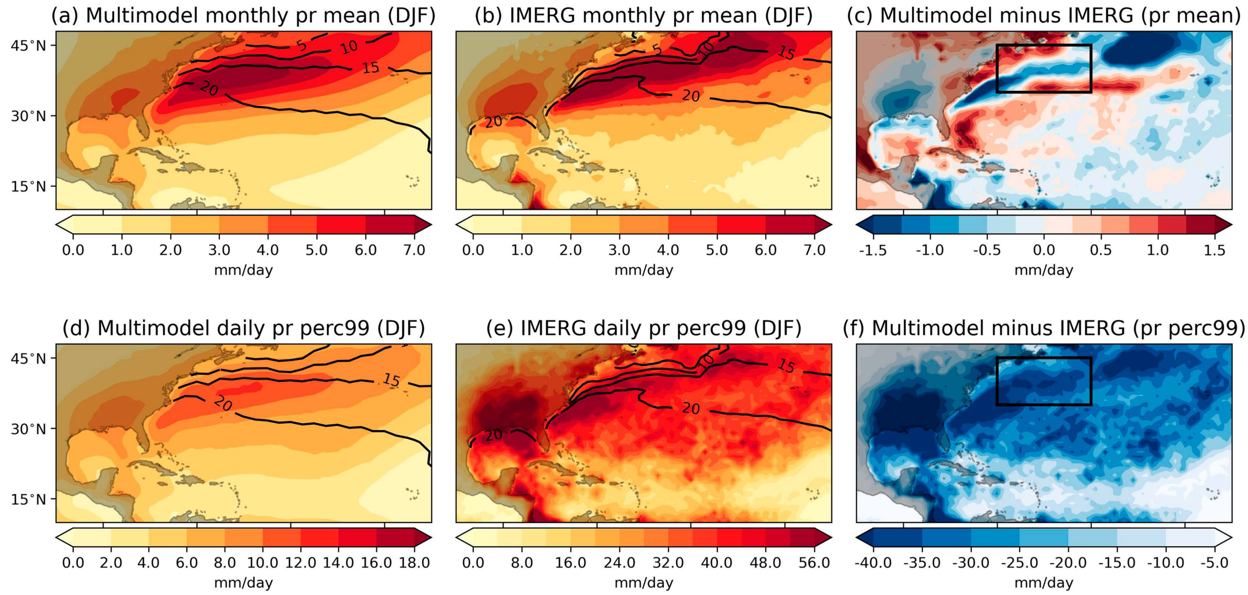

FIG. 2. (a) The multimodel mean monthly precipitation (filled contours), along with the multimodel mean SSTs (line contours), at each grid point. (b) The mean monthly IMERG precipitation (filled contours), along with the ERA5 mean SSTs (line contours). (c) The difference (a) minus (b). (d) The multimodel mean daily 99th percentile of precipitation (filled contours) and multimodel mean SSTs (line contours). (e) The 99th percentile of daily IMERG precipitation. (f) The difference (d) minus (e). The season is always restricted to DJF. The period covered is 1979–2014 for model data and 2000–15 for observations/reanalysis. (c) and (f) The Gulf Stream region is highlighted with a black box.

multimodel mean and ERA5 are shown as well. In both models and observations, there is a strong band of precipitation following the Gulf Stream, as expected (Minobe et al. 2008). However, the band is generally more diffuse in models. While some of this "blurring" is likely due to different models placing the front at slightly different latitudes, it is also consistent with the general failure of models to reproduce the early Gulf Stream separation, and consequent sharp SST gradients, observed in the real world: We confirm this in section 5b. Figures 2d–f show the analogous plots but for the 99th percentile of daily precipitation. Comparing Figs. 2d and 2e suggests the underestimation of precipitation extremes in models is often as high as 50% (note the different ranges of the color bars). The equivalent plot using TRMM rather than IMERG is given in Fig. S2. TRMM places the peak precipitation more clearly in the Gulf Stream box, unlike IMERG, which produces peaks further upstream and downstream. TRMM also generally produces even more intense precipitation than IMERG: See also Fig. S1. The model bias is therefore even greater when measured against TRMM. It is unclear to the authors whether TRMM or IMERG gives the more realistic estimate.

Figures 3a–c (top row) show the mean 850-hPa zonal winds (U850) for models, ERA5, and the model bias, while Figs. 3d–f (bottom row) show the analogous plot but for SSTs. The zonal wind biases show the too zonal, too equatorward bias emphasized in Schemm (2023). Figure 3c also highlights a strong negative bias emerging from the southern tip of Greenland. This is the region highlighted when doing composite analysis of the northern jet latitude events (Woollings et al. 2010), and this negative zonal wind bias is likely therefore partially reflecting model biases in such northern jet events. Some of this bias is likely also related to biases in Greenland tip jet events (Doyle and Shapiro 1999), which make up some proportion of the northern jet events (White et al. 2019). This region is also highlighted in Fig. 1a and thus is likely also reflecting European blocking biases (O'Reilly et al. 2017; Dorrington et al. 2022). Figure 3f clearly highlights two model biases: the reduced SST gradient in the Gulf Stream separation region and the large cold bias in the central subpolar North Atlantic. As explained in the introduction, these biases are closely linked to the failure of models to simulate the early Gulf Stream separation and northwest corner.

It is clear from Fig. 2 that the model bias in precipitation extremes is substantial and systematic, something which is further confirmed in the next section. To better understand how systematic the mean biases in precipitation and SSTs in the Gulf Stream region are, we show in Fig. 4 histograms of the Gulf Stream area-averaged precipitation and SSTs across all models, along with the observational estimates. These make it clear that there is no significant mean bias in the ensemble as a whole, despite differences in the Gulf Stream separation and the more diffuse precipitation band in the models. For precipitation, this is likely because models are often tuned to achieve a realistic energy budget (Hourdin et al. 2017), and the mean precipitation in models is tightly governed by energetic constraints, with the total amount of precipitation and evaporation being approximately balanced in observations and models on long time scales and large spatial scales (Dagan et al. 2019). Note that the mean Gulf Stream precipitation value in ERA5 is almost identical to that of IMERG.

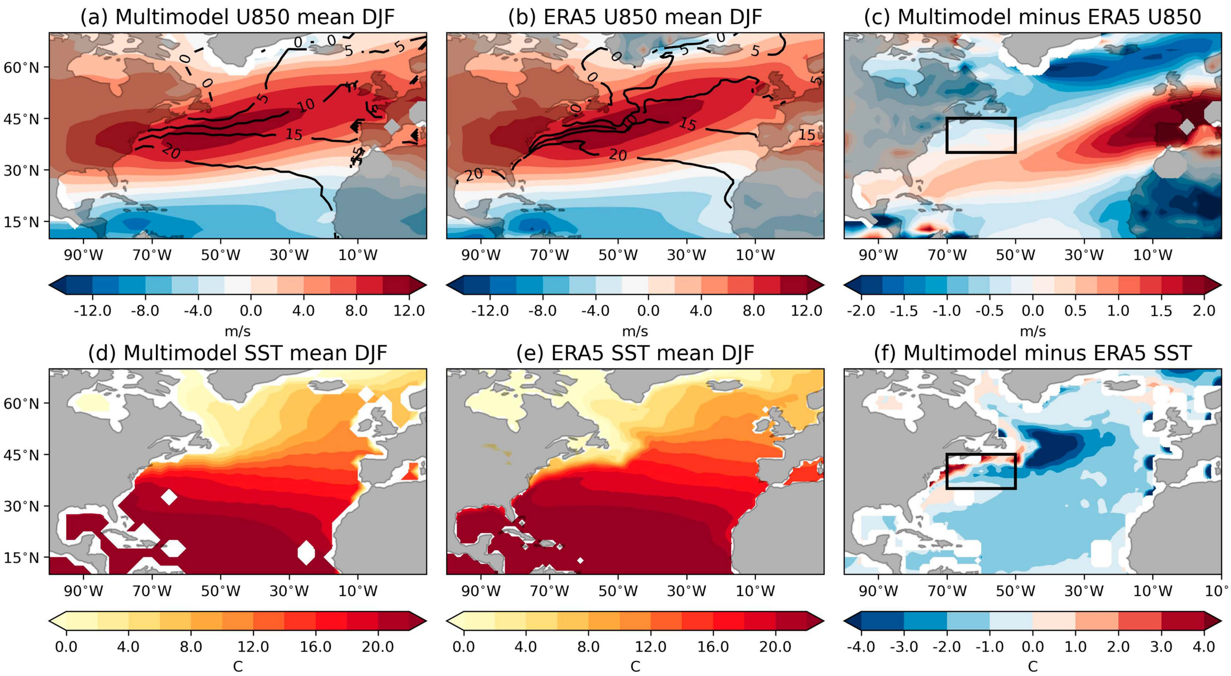

FIG. 3. (a) The multimodel mean monthly 850-hPa zonal winds (filled contours), along with the multimodel mean SSTs (line contours), at each grid point. (b) The mean monthly ERA5 850-hPa zonal winds (filled contours), along with the ERA5 mean SSTs (line contours). (c) The difference (a) minus (b). (d) The multimodel mean SSTs (filled contours). (e) The mean ERA5 SSTs. (f) The difference (d) minus (e). The season is always restricted to DJF. The period covered is 1979–2014 in all cases. (c) and (f) The Gulf Stream region is highlighted with a black box. For a discussion on the occasional missing values in model data, see methods.

b. The link between Gulf Stream precipitation extremes, European blocking, and northward jet excursions

We now examine the link between Gulf Stream precipitation extremes and European blocking/northern jets. We compute correlations across the multimodel ensemble between a model's GS PR99 and the climatological mean of PRmax, U850, and Z500 at each spatial location. This is shown in Fig. 5. Thus, if a correlation at a particular grid point is positive (negative), it means that models with greater Gulf Stream precipitation variability tend to have a greater (lower) climatological mean PRmax, U850, or Z500 at that grid point. Figure 5a shows that models with greater GS PR99 also experience enhanced precipitation maxima broadly throughout the North Atlantic domain, consistent with both the model fundamentally simulating precipitation better and dynamical downstream effects via the storm track. Figure 5b shows that lower GS PR99 is also associated with a too zonal, too equatorward jet, consistent with the bias seen in Fig. 3c. Finally, Fig. 5c shows that greater GS PR99 is associated with a positive Z500 anomaly over the subpolar North Atlantic, stretching into Europe. Comparing this figure with Fig. 1 suggests a link between GS PR99 and blocking.

These impressions are confirmed in the scatterplots shown in Fig. 6. Figure 6a shows a visually clear link between GS PR99 and European blocking in models: The correlation is 0.63, with approximately 40% of the spread in European blocking frequencies being accounted for by GS PR99. For NorthDays, the correlation is 0.55, corresponding to around 30% of the variance explained. The somewhat smaller correlation for NorthDays may be because the impact of GS PR99 on the jet consists of both an increase in the northern jet latitude regime *and* a reduction in the central (and possibly southern) jet regime (see Fig. 5b), meaning the full impact on the jet latitude variability is not visible solely in the NorthDays metric. It should also be recalled that while European blocking events are most typically associated with a northern jet, the two metrics are relatively decoupled from each other on daily time scales, with the jet often found at central and southern latitudes when European blocking is ongoing (Woollings et al. 2010).

While Fig. 6 gives the impression that some models are able to achieve realistic levels of precipitation variability, a very different impression would have been given if we had used TRMM in place of IMERG: We comment more on this in the following subsection. Figure S3 shows the equivalent of Fig. 6 if mean precipitation is used, showing almost identical correlations to those in Fig. 6. As previously mentioned, this is a result of the fact that the precipitation distribution is bounded below by zero and fat tailed, which results in the mean precipitation being strongly correlated with higher-order moments. However, as emphasized in Fig. 4, there is no systematic bias in the mean precipitation in the models, unlike the systematic underestimation of blocking. Thus, the mean Gulf Stream precipitation is unlikely to play a major role in driving model biases in

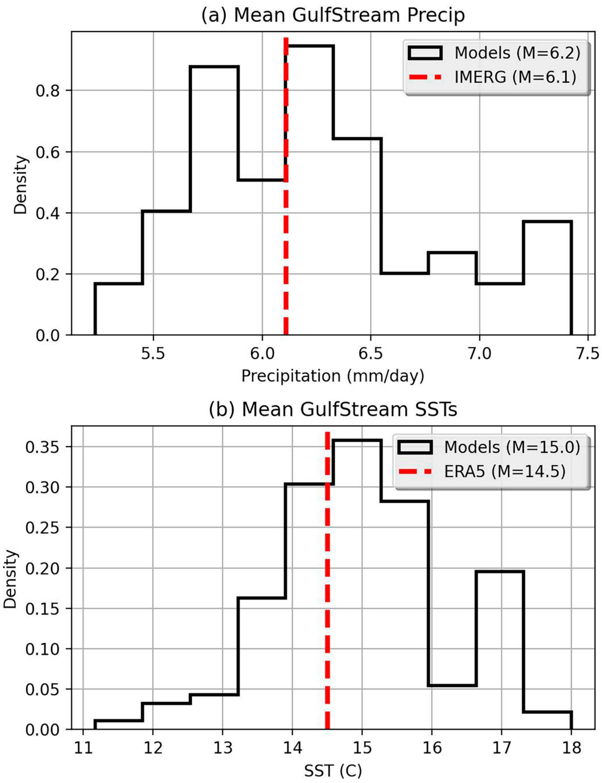

FIG. 4. (a) Histogram of model mean daily precipitation averaged over the Gulf Stream domain. (b) Histogram of model monthly mean SSTs averaged over the Gulf Stream region. The value for observations/reanalysis is marked with a red line in each, with the legends indicating also the multimodel means. The period covered is 1979–2014 except for IMERG, which covers 2000–15.

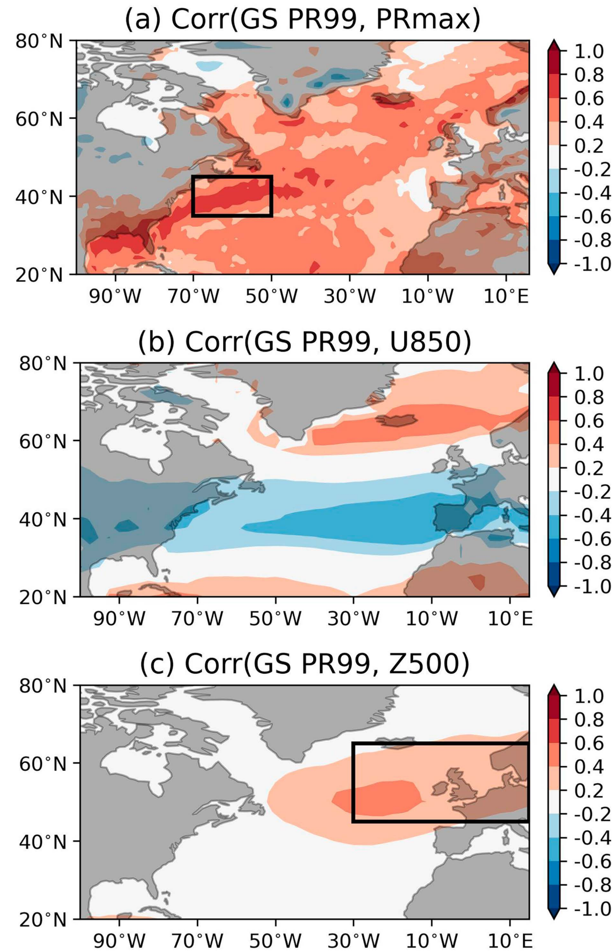

FIG. 5. Correlations between GS PR99 and the local (a) climatological maximum monthly precipitation PRmax obtained across any of the winter months; (b) climatological DJF U850; (c) climatological DJF Z500. Correlations are computed across the multimodel ensemble. The period considered is always 1979–2014. Correlations outside the zero contour (± 0.2) are statistically significant ($p < 0.05$). (a) The Gulf Stream region is highlighted with a black box, and (c) the region used to define European blocking is similarly highlighted.

blocking. On the other hand, if any measure of variability or extremes is used, IMERG and TRMM will both appear at the upper end of what models are able to simulate, in better agreement with the situation for European blocking.

Finally, we note that if we only retain a single ensemble member for every model, the correlations between GS PR99 and European blocking/NorthDays are around 0.5 for both. Thus, the link highlighted here is not an artifact of a few individual models with many ensemble members.

c. *The effect of atmospheric horizontal resolution*

In the previous section, we have seen that Gulf Stream precipitation extremes are strongly linked to European blocking frequency and northern jet excursions. We now examine how all these three factors are linked to atmospheric horizontal resolution.

Figure 7 shows scatterplots of horizontal resolution versus (i) GS PR99, (ii) European blocking frequencies, and (iii) northern jet days. Significant correlations are found in all cases, with the highest correlation being that of -0.58 between resolution and blocking. This correlation is identical to that reported in Davini and d'Andrea (2020) between resolution and a closely related blocking metric[1] using a similarly large multimodel dataset: See their Fig. 3b. The link between resolution and European blocking therefore appears to be particularly robust. As in the previous section, the correlation for NorthDays is smaller.

Figure 7a confirms the expected association between enhanced precipitation extremes and higher atmospheric resolution, noted in a variety of past modeling studies (Kopparla et al. 2013; Bador et al. 2020; Strandberg and Lind 2021), including ones focused on the Gulf Stream region in particular

[1] Davini and d'Andrea (2020) compute DJFM blocking frequencies over a central/eastern European region which partially overlaps with the region we consider, over the period 1961–2000.

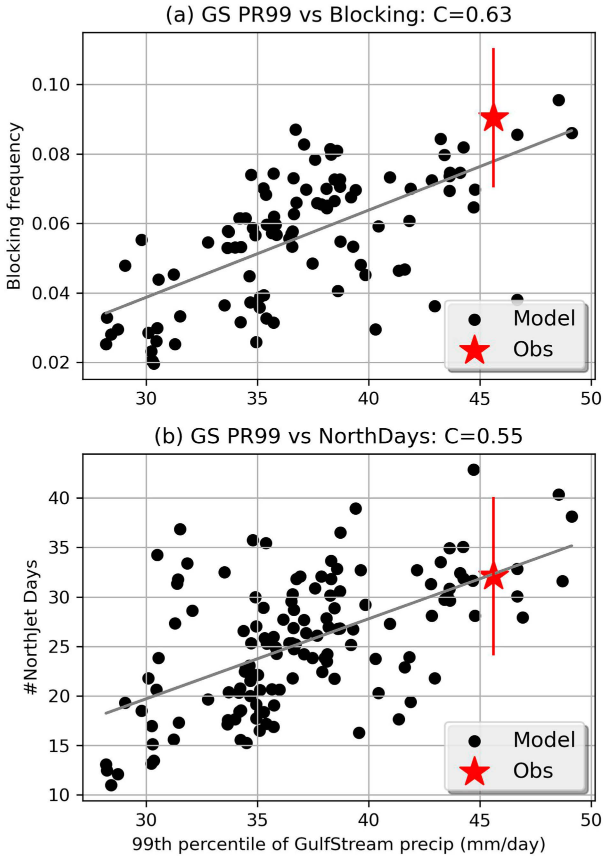

FIG. 6. Scatterplots of GS PR99 vs (a) European blocking frequency and (b) northern jet days. Black dots are models, while the red star is observations (IMERG/ERA5). The red lines indicate our measure of internal variability in observations. The gray lines in each panel are a linear fit to the black dots. The value of C in each title is the correlation measured using model data. Model data cover the period 1979–2014, while observations cover 2000–15.

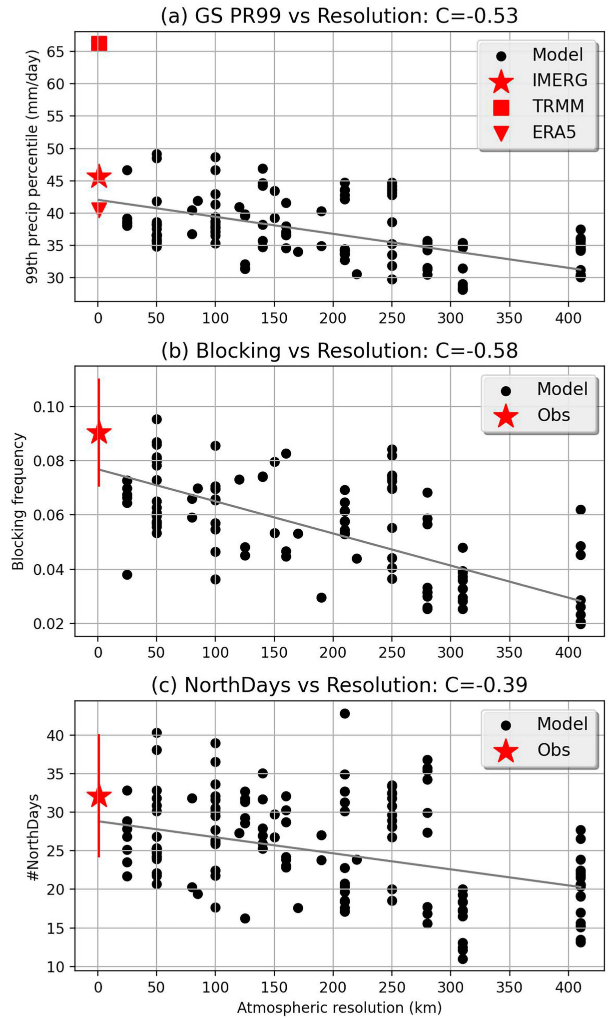

FIG. 7. Scatterplots of atmospheric horizontal resolution [grid spacing at the equator (km)] vs (a) GS PR99, (b) European blocking frequency, and (c) northern jet day occurrence. Black dots are models, while the red symbols are observations (IMERG/TRMM/ERA5: See legends). The red lines in (b) and (c) indicate our measure of internal variability in observations. The gray lines in each panel are a linear fit to the black dots. The value of C in each title is the correlation measured using model data. Model data cover the period 1979–2014, while observations cover 2000–15.

(Scher et al. 2017). IMERG gives the clear impression that a nominal atmospheric resolution of around 25 km is sufficient to simulate realistic precipitation variability. Strikingly, however, TRMM is considerably higher than the 25-km models, and a major step change would need to take place at higher resolutions to bring model variability in line with TRMM. In fact, the possibility that such a major step change may occur once a threshold resolution has been exceeded has been hypothesized before (Slingo et al. 2022). Previous studies, like Scher et al. (2017) and Strandberg and Lind (2021), have shown that the impact of atmospheric resolution on the rainfall distribution is asymmetric, with little/no change to low-magnitude events and larger changes to extreme events. Such asymmetries, likely related to the increasingly accurate ability of high-resolution models to resolve individual storm systems, are often associated with nonlinearities, which a step change would be an example of. A nonlinear relationship can also be motivated by a simple conceptual model. Precipitation simulated by a climate model in a grid box is, roughly speaking, obtained as an average over individual systems taking place therein. One can therefore imagine a model's precipitation as being akin to a spatially smoothed version of real-world precipitation. To assess how such smoothing affects the 99th percentile, one can, for example, apply a uniform or Gaussian filter to fixed samples from a Pareto distribution, which is commonly used as a statistical model for precipitation (Cavanaugh et al. 2015). This shows that the value of the 99th percentile drops off sharply as the window size of the filter increases (i.e., as the samples are smoothed more), before flatlining as the effect of additional smoothing begins to saturate: See Fig. S4. Thus, a steep rise in the magnitude of extremes as models exceed 25-km resolution cannot be ruled out.

Interestingly, there is some indication of nonlinearity in Fig. 7a, since the Spearman correlation of -0.55 is greater in magnitude than the corresponding Pearson correlation of -0.45, indicative of a monotonic relationship beyond the purely linear one. However, we do not pursue this line of thought further and do not make any robust claims. We only discuss these potential nonlinearities because of the relevance these would have for model development. Our actual results are not affected either way.

Based on the results seen so far, we hypothesize that the pathway from "increased horizontal resolution" to "reduction in blocking/jet biases" is mediated, at least in part, by "improved Gulf Stream precipitation extremes." The standard test for mediation (Baron and Kenny 1986), in the context of European blocking specifically, is therefore as follows (the procedure is analogous for NorthDays):

1) Perform a linear regression of horizontal resolution against European blocking to obtain **blocking** $= A \times$ **resolution** $+ C_1 +$ **noise** for some constants A and C_1, and check that there is a significant correlation.
2) Perform a linear regression of GS PR99 against European blocking to obtain **blocking** $= B \times$ **GS PR99** $+ C_2 +$ **noise** for some constants B and C_2, and check that there is a significant correlation.
3) Perform a multiple linear regression of resolution *and* precipitation against European blocking to obtain **blocking** $= A' \times$ **resolution** $+ B' \times$ **GS PR99** $+ C_3 +$ **noise** for some constants A', B', and C_3, and assess if $A' < A$.

The interpretation here is that if the impact of resolution is not through precipitation, then the coefficients A and A' should be similar, while if the impact is mostly through precipitation then $A' < A$, because this mediating effect will be subsumed by the B' coefficient. Steps 1 and 2 here have been affirmed in the positive with Figs. 6 and 7. We carried out step 3 separately: The regression coefficient associated with resolution in step 1 was estimated as -0.73 (using time series that are normalized to have mean zero and standard deviation 1), and this halves to -0.36 in step 3. This reduction was estimated as statistically significant ($p < 0.05$) using the Aroian version of the Sobel test, described in Baron and Kenny (1986). Thus, the data we have analyzed here support our hypothesis.

d. *Precipitation variability in AMIP models*

We will now look briefly at how Gulf Stream precipitation extrema differ between AMIP and coupled models. We do this here by referring to the classic work of Barsugli and Battisti (1998) (hereafter BB98). The implications this analysis has for assessing the relative roles of atmospheric and oceanic resolutions will be considered in the discussion.

BB98 constructed a simple pair of stochastic ordinary differential equations to model air–sea coupling and used it to explain observed differences between coupled models and AMIP-style models. It is well known that heat fluxes are unrealistic in AMIP models, due to the ocean being an infinite source/sink of energy, and this is explained by the BB98

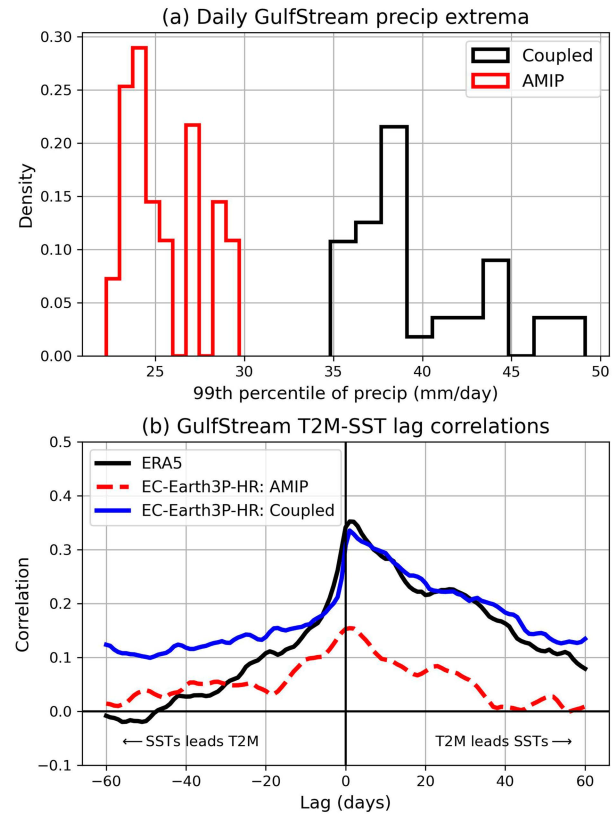

FIG. 8. (a) Histograms of GS PR99 for the coupled PRIMAVERA/HighResMIP models (black) and the counterpart AMIP models (red). (b) Lead–lag correlations of daily Gulf Stream SST vs daily Gulf Stream T2M for ERA5 (black), the coupled version of EC-EARTH3P-HR (blue), and the AMIP version of EC-EARTH3P-HR (red stippled). Negative lags correspond to SSTs leading T2M, and positive lags correspond to T2M leading SSTs.

model. What seems to be less well known is that AMIP models have systematically reduced daily surface temperature variability compared to coupled models, due to what BB98 describe as the excessive thermal damping taking place in AMIP models. A self-contained presentation of the BB98 model and how it accounts for these differences between AMIP and coupled models is given in the appendix.

The reason this is relevant is because the reduced variability of daily surface temperatures results in a similarly reduced variability of daily precipitation, due to the effect of temperature on convective processes, such as evaporation and condensation, as we show next. This reduction in variability comes hand in hand with a reduction in the tails of the precipitation distribution, due to the asymmetric nature of the distribution (bounded below and fat tailed). Figure 8a shows the 99th percentile of daily Gulf Stream precipitation for the AMIP simulations of HighResMIP (red histogram) and the corresponding set of coupled simulations (black histogram). It can be seen that even the highest resolution AMIP model experiences extremes significantly lower than the lowest-resolution coupled model, with the magnitude being around 40% lower on average in the AMIP models. BB98 demonstrate that this is due to the missing coupling, as measured by the lead–lag correlations of SSTs versus surface temperature (T2M). An example of this using the EC-EARTH3P-HR model is shown in Fig. 8b, reproducing what is expected from the analysis of BB98.

5. Discussion

a. Mechanisms

There are at least three challenges to clarifying the mechanisms involved in the link between extreme precipitation and European blocking. First, "there is no comprehensive theory capturing the different processes acting at all stages of the blocking life cycle" (Woollings et al. 2018a), with some theories emphasizing more local processes (like explosive cyclogenesis) and others emphasizing more global processes (like the interaction between planetary waves). Second, our measure of frequency of occurrence mixes up blocking onset events with the persistence of already-formed blocks, and the processes involved could differ between the two. Third, explicitly testing Schemm's suggestion that more intense precipitation produces more poleward-propagating cyclones would require us to perform Lagrangian cyclone tracking in the 135 model simulations we have made use of. While we believe much could be learned by carrying out such tracking in model data, we did not do this here due to the intense computational cost involved.

However, existing literature suggests a plausible mechanism explaining our results. Preliminary analysis suggests that increases in blocking occurrence in models are most clearly related to an increased number of onsets, as opposed to increases in the persistence of blocks (see Supporting Fig. S5). This is also consistent with the results of Schiemann et al. (2020). Untangling the difference between onset and persistence more comprehensively goes beyond the scope of this paper, but this preliminary result prompted us to focus our mechanistic hypothesis on the link between extreme precipitation and blocking onset. In the following discussion, we draw in particular from the summary given in Woollings et al. (2018a).

Cyclones undergoing intensification experience a strong ascent of air in the warm sector of the cyclone, the so-called warm conveyor belt. Such warm conveyor belts effectively act as conduits for negative PV anomalies that are deposited where the ascent peaks. Latent heating amplifies this ascent, allowing the PV outflow to happen at higher levels, with the net effect being to generate more intense anticyclonic, negative PV anomalies, precisely the features characterizing a blocking event (Madonna et al. 2014; Methven 2015). Pfahl et al. (2015) showed that this contribution from latent heating is of primary importance in the development of blocks. Given these points, it is natural to assume that more intense precipitation is associated with more intense latent heat release and, consequently, even more intense anticyclonic anomalies. We thus hypothesize that models which can simulate more intense Gulf Stream precipitation are able to more readily generate intense anticyclonic wavebreaking, and hence more frequently trigger the onset of European blocking episode, because the stronger latent heat release occurring during intense precipitation allows the outflow of negative PV along a cyclone warm conveyor belt to occur at higher levels. For the same reason, we hypothesize that models which simulate more intense precipitation can more readily simulate explosive cyclogenesis, and hence produce more poleward-propagating cyclones, as Schemm suggests. This hypothesis would thus naturally explain both our results and those of Schemm (2023) and De Luca et al. (2024).

In lieu of explicit cyclone tracking, we offer two pieces of analysis supporting this hypothesis. First, we examine composite anomaly maps in our multimodel ensemble. Figure 9 shows sea level pressure (SLP) and precipitation anomalies in the aftermath of a day where the area-averaged Gulf Stream precipitation was low/moderate (between 5 and 10 mm accumulated over the day) or intense (more than 15 mm, the low end of the 99th percentiles occurring in the ensemble). It can be seen that the rainfall anomalies coincide with a low-pressure anomaly suggestive of a cyclone in the Gulf Stream domain. Following low/moderate rainfall, the pressure anomaly moves almost directly eastward, with the minimum SLP anomaly shifting northward by around 2°. The anomaly has effectively vanished by day 3, with the anomalous precipitation gone already on day 2. Following intense rain, the low-pressure anomaly is considerably larger in both magnitude and spatial extent, as is the spatial extent of the rainfall. On day 2, the pressure anomaly has moved visibly north–east out of the Gulf Stream region, with the minimum SLP anomaly shifting north by around 5°; the anomalous rainfall now extends to the coast of Greenland. The pressure anomaly is still visible on day 3. Figure 10 shows similar composites for 250-hPa geopotential height anomalies (Z250). Both low/moderate and intense rainfall are followed by a Rossby wave train propagating eastward and equatorward, but the wave train is considerably greater in magnitude and spatial extent following intense rainfall, with positive pressure anomalies extending into the European domain on day 3. These figures are thus consistent with the intense rainfall being associated with more explosive cyclogenesis and further poleward propagation of the resulting cyclone. We remark without further comment that the wave train seen in Fig. 10 is reminiscent of an Atlantic Ridge weather regime, which is more likely to transition to a European blocking regime than other Euro-Atlantic weather regimes (Matsueda and Palmer 2018).

Figure 11 furthermore shows model composite anomalies of SLP and precipitation in the three days leading up to the onset of a European blocking episode (where an episode is here defined using the clustering method described in the methods). These show a strong anomaly in rainfall in the Gulf Stream region 2 and 1 days prior to onset, with the rainfall extending all the way up to the coast of Greenland. There are also pressure anomalies suggesting the presence of a preexisting ridge and an upstream cyclone, as expected based on prior literature (see, e.g., Steinfeld and Pfahl 2019). These composite maps thus jointly suggest that (i) anomalously high Gulf Stream precipitation associated with an extratropical cyclone is a typical precursor for European blocking in the models, and (ii) more intense precipitation is associated with larger cyclones that propagate further poleward and persist for longer. Similar composites are found when using ERA5 (not shown).

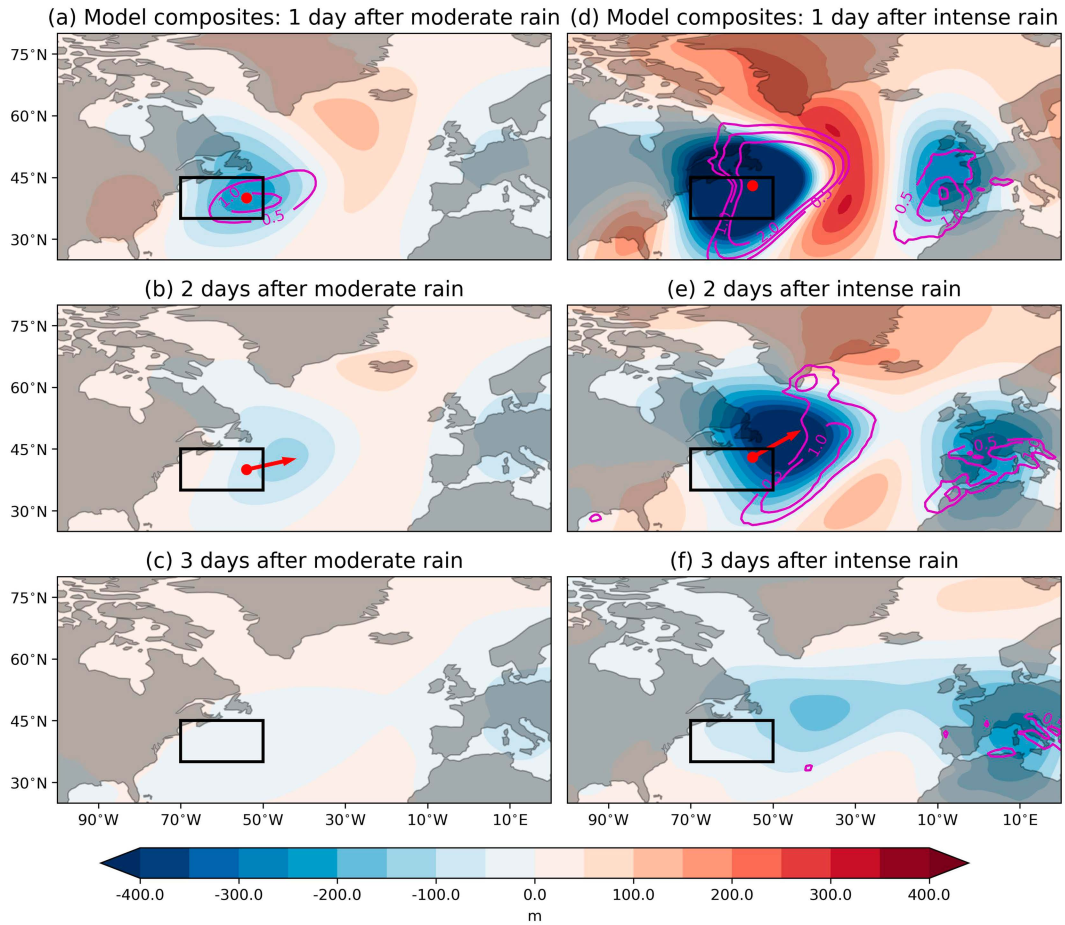

FIG. 9. (left) Anomalous SLP composites (filled contours), across all model data, in the 3 days following a day with moderate Gulf Stream precipitation, and (right) extreme Gulf Stream precipitation. The mauve line contours are composites of anomalous precipitation. Moderate precipitation days are defined as days where the area-averaged Gulf Stream precipitation is between 5 and 10 mm, and extreme precipitation days are defined as ones where the averaged precipitation exceeds 15 mm. The Gulf Stream box is highlighted in black. (a),(d) The red dot shows the location of the minimum SLP anomaly. (b),(e) The arrow shows where the minimum SLP anomaly moves to on the next day. Model data are restricted to 1979–2014.

Second, we have analyzed how the intensities of anticyclonic wavebreaking (AWB) events are affected by Gulf Stream precipitation in observations. Here, we make use of a wavebreaking dataset for ERA5 generated using software developed by Kaderli (2023): see methods. While it would be of interest to compute wavebreaking intensities also in model data, this would, like explicit cyclone tracking, be computationally unfeasible for this study. We restrict ourselves therefore to using precomputed data for ERA5; as always, we use IMERG for precipitation. Figure 12 shows composites of anticyclonic wavebreaking intensity and 250-hPa PV two days after low/moderate or extreme Gulf Stream precipitation. The choice of 2 days is motivated by the peak Gulf Stream precipitation seen 2 days prior to blocking onset in Fig. 11b. Figure 12 shows that 2 days after low/moderate rain there is no significant anomalous AWB taking place anywhere in the Euro-Atlantic sector. However, following extreme rain, there is a significant increase in AWB intensity in the southern part of the European domain, indicating a wavebreaking event precisely of the kind associated with European blocking (Fig. 1b). Consistent with this, there is a strong PV dipole spanning the domain. Figure 12 thus suggests that strong AWB in the European domain (i.e., European blocking) is far more likely to take place following extreme rainfall. We note that if ERA5 rainfall had been used in place of IMERG to make these composites, then one can still see a marked increase in AWB intensity following intense precipitation in roughly the same region as Fig. 12b, but the change is not statistically significant and there is no clear PV dipole. This suggests that ERA5 does not place its own extremes on the same days as IMERG.

While these two pieces of analysis offer some evidence toward our suggested mechanism, they are by no means conclusive. In particular, it may be that compositing on heavy precipitation is simply tantamount to compositing on the presence of a stronger/more well-defined cyclone in the Gulf Stream. Future work should look more closely at the transient

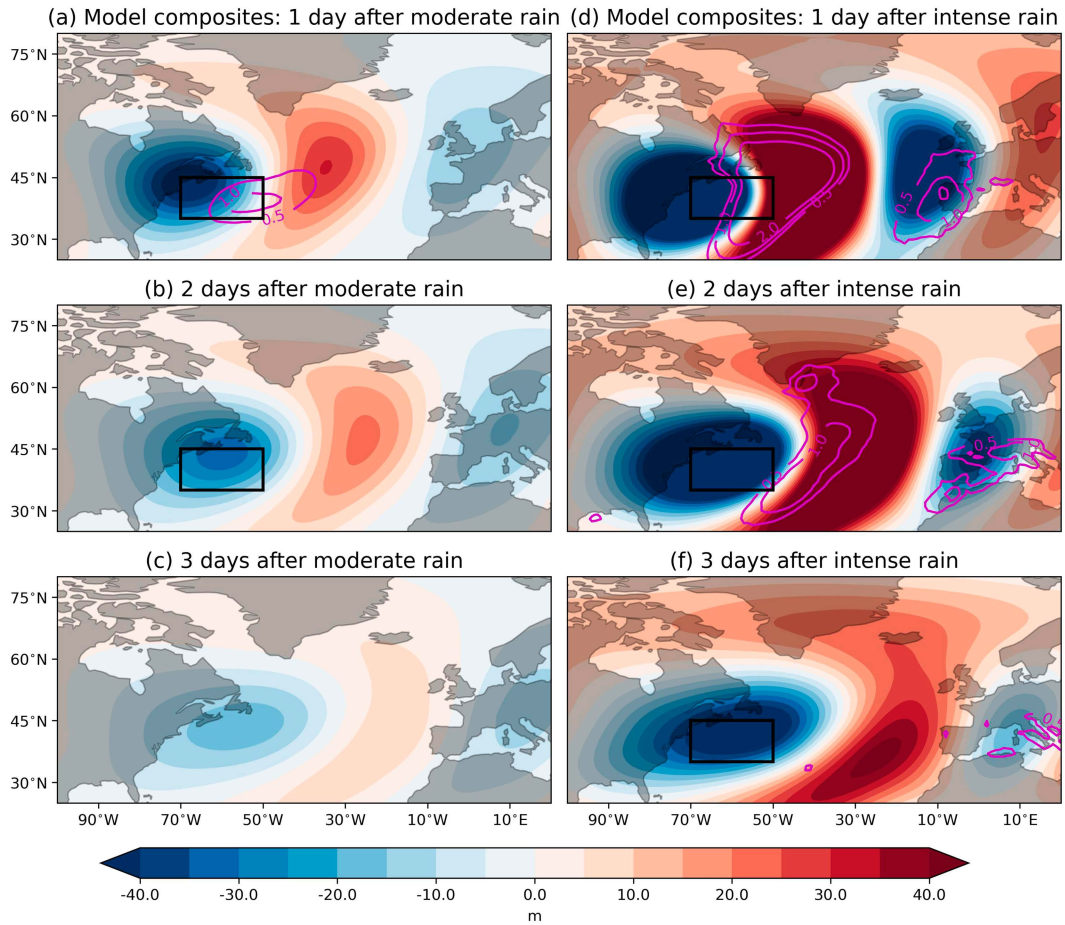

FIG. 10. (left) Anomalous Z250 composites, across all model data, in the 3 days following a day with moderate Gulf Stream precipitation, and (right) extreme Gulf Stream precipitation. The mauve line contours are composites of anomalous precipitation. Moderate precipitation days are defined as days where the area-averaged Gulf Stream precipitation is between 5 and 10 mm, and extreme precipitation days are defined as ones where the averaged precipitation exceeds 15 mm. (a),(d) The Gulf Stream box is highlighted. Model data are restricted to 1979–2014.

evolution of blocks and cyclone propagation in multimodel ensembles and the contribution of extreme precipitation, as has been done for reanalysis data (Wenta et al. 2024; Mathews and Czaja 2024). Additional idealized studies aimed at probing the role of dry versus moist effects, as in Mathews et al. (2024), would also be invaluable.

b. Alternative hypotheses

One possible alternative explanation for the correlations identified in Figs. 6 and 7 is that the causality goes the other way. That is, perhaps European blocking triggers more intense precipitation, and hence, models which simulate more frequent blocking also simulate more frequent precipitation extremes. However, Fig. 13 shows that Gulf Stream precipitation only visibly peaks before, and not after, the onset of European blocking. It therefore seems unlikely that the causality could be reversed here.

Another alternative explanation for the correlations in Figs. 6 and 7 is that there is some other process or component of the climate system which modulates both blocking/jet and precipitation extremes. In terms of processes, one natural possibility would be transient eddy feedbacks, since one could imagine stronger feedback (i) invigorating storm systems in the Gulf Stream, thereby enhancing precipitation variability, and (ii) also allowing cyclones to persist for longer and thereby amplify blocking anticyclones. However, Dorrington et al. (2022) showed that eddy feedbacks do not systematically increase with horizontal resolution in the multimodel ensemble considered here, unlike blocking/NorthDay frequency and precipitation extremes. Variations in eddy feedback strength therefore would not be able to explain the correlations we find here. Another possibility is that atmospheric resolution improves the model's ability to simulate the strong updrafts that occur as weather fronts move across Gulf Stream SST anomalies, which could also affect jet variability, as suggested in Wills et al. (2024). We consider this an important avenue for future work which could clearly complement the analysis of our paper, given that the most intense precipitation in the storm-track region is associated with fronts and their associated cyclones (Catto et al. 2012; Hawcroft et al. 2012).

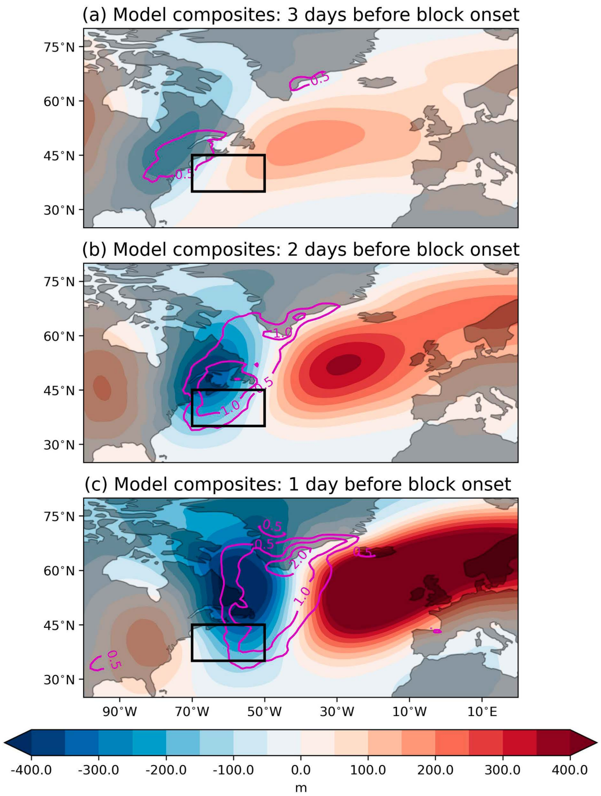

FIG. 11. Anomalous SLP composites (filled contours), across all model data, in the 3 days prior to the onset of a European blocking episode. The mauve line contours are composites of anomalous precipitation. The Gulf Stream box is highlighted in black. Model data are restricted to 1979–2014.

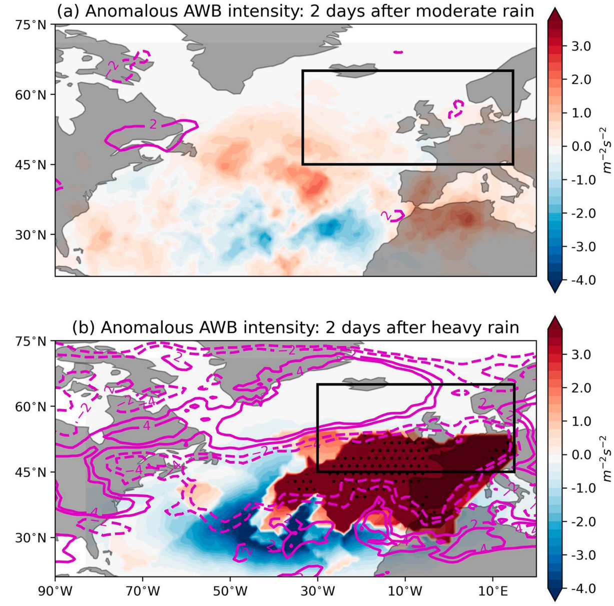

FIG. 12. Anomalous AWB intensity (filled contours) and 250-hPa PV anomalies (PV250; mauve contours: Stippled are negative and solid are positive) in ERA5 following (a) days with moderate Gulf Stream precipitation and (b) extreme Gulf Stream precipitation. Moderate precipitation days are defined as days where the area-averaged Gulf Stream precipitation is between 5 and 10 mm, and extreme precipitation days are defined as ones where the averaged precipitation exceeds 15 mm. ERA5 is used for AWB and IMERG for precipitation. The data cover 2000–15. Stippling indicates statistical significance ($p < 0.05$). The European blocking domain is highlighted in black.

In terms of other components of the climate system, the most obvious ones are SST biases. To test if SST biases can explain our results, we tested three different SST metrics. First, the SST gradient in the central North Atlantic region (45°–15°W, 40°–50°N): This is the bias that Athanasiadis et al. (2022) argued is responsible for biases in European blocking/NorthDays.[2] Second, the SST gradient in the Gulf Stream region we have worked with here (70°–50°W, 35°–45°N): Gradients in this region have been argued to influence blocking/NorthDays in O'Reilly et al. (2016) and O'Reilly et al. (2017). Third, a measure of the North Atlantic "cold SST bias" is computed as the mean SST over the box 40°–50°N, 45°–30°W. This box roughly encloses the large cold bias visible in Fig. 3f. This bias has been argued to influence blocking/jet variability by Scaife et al. (2011) and Keeley et al. (2012): It is sometimes colloquially referred to as the "blue blob of death." Our choices are motivated by the desire to test metrics that are easily interpretable and corroborated by prior literature. Thus, we do not test metrics such as the basinwide root-mean-square error (RMSE) of North Atlantic SSTs, as done in Davini and d'Andrea (2020), as these cannot be clearly interpreted: The biases contributing to the error could in principle differ widely from model to model. Note that both SST

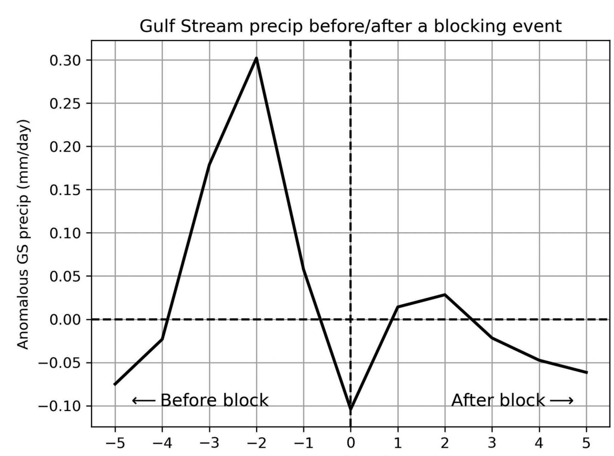

FIG. 13. Lagged composites of area-averaged Gulf Stream precipitation in 5 days before and after the onset of a European blocking event, using the model ensemble. Blocking begins at day 0 with negative (positive) lags indicating days before (after) onset.

[2] Note that Athanasiadis et al. (2022) do also suggest that biases in diabatic processes may be playing a role. However, they do not test this quantitatively. Only the link between SST gradients and blocking is tested quantitatively.

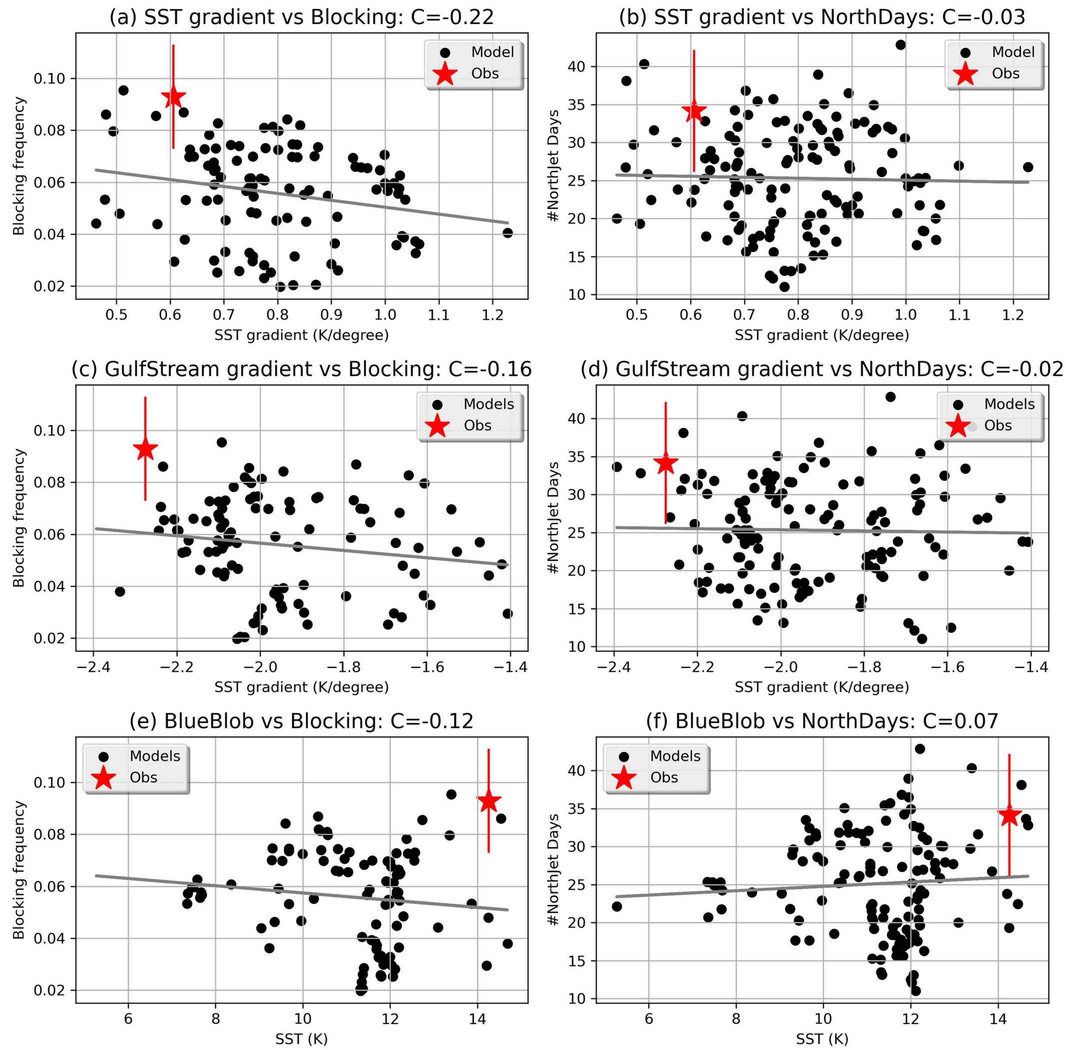

FIG. 14. Scatterplots of the SST gradient in the central North Atlantic vs (a) European blocking frequency and (b) northern jet day occurrences (NorthDay). (c),(d) As in (a) and (b), but for the Gulf Stream SST gradient. (e),(f) As in (c) and (d), but for the North Atlantic cold SST bias (blue blob of death). Black dots are models, while the red star is observations (ERA5). The red lines indicate our measure of internal variability in observations. The value of C in each title is the correlation measured using all the model data. Model data cover the period 1979–2014, while observations cover 2000–15. The SST gradient metric is directly adapted from that used in Athanasiadis et al. (2022).

gradient metrics are computed by taking zonal averages of the SSTs over the respective region and regressing the zonally averaged SSTs against latitude: The regression coefficient is defined to be the SST gradient.

The resulting scatterplots are shown in Fig. 14. In all three cases, correlations between the SST metric and European blocking/NorthDay frequencies are found to be low. For European blocking, the correlations, in order, are −0.22, −0.16, and −0.12, while for NorthDays, they are −0.03, −0.02, and 0.07. The only significant correlation is the one between the central North Atlantic SST gradient and European blocking. This suggests SST gradients in this region are robustly associated with blocking frequencies, consistent with the hypothesis of Athanasiadis et al. (2022). However, the SST gradient correlation of −0.22 is considerably lower in magnitude than the one we report between Gulf Stream precipitation variability and blocking of 0.63, and the SST gradient correlation of −0.03 is negligible in the case of NorthDays, unlike the correlation of 0.55 reported using precipitation. It is worth emphasizing that the SST gradient correlations are notably higher if one restricts to only HighResMIP models, in line with the consistency of the relationship reported by Athanasiadis et al. (2022): See Fig. S6. There is some evidence to suggest that HighResMIP may to some extent be systematically different from the wider multimodel ensemble, at least in the case of European blocking. It cannot therefore be ruled out that SST gradients play a negligible role in coarse-resolution models but a larger role for higher-resolution models. However, it also seems plausible that the stronger link between SST gradients and blocking frequencies in HighResMIP is inflated by sampling variability, and that the true link is lower, as in our multimodel ensemble. See Text S1 for further discussion. We also note that Athanasiadis et al. (2022) use a more strict definition of "northern jet days" as ones where the jet latitude exceeds 60°N, as opposed to our threshold of 52°N. We found that if the stricter threshold is used, then there is a stronger correlation of −0.37 between NorthDays and the SST gradient from Athanasiadis et al. (2022), as opposed to the correlation of −0.03 shown in Fig. 14b: See Fig. S7. However, if the stricter threshold is used, then we found no impact of model resolution on NorthDays: see Fig. S8. Thus, while SST gradients may be more strongly linked to the most extreme northern jet days, this link appears to be unrelated to model resolution (in the ocean or atmosphere). The link is also still smaller than that found using Gulf Stream precipitation.

Thus, based on the three metrics we considered, it seems that North Atlantic SST biases are at best weakly associated with biases in European blocking and northern jet excursions. This is not to say SST biases are unimportant! For any *specific* model, its biggest SST bias could contribute more strongly, as suggested by the correlation between RMSE and blocking reported by Davini and d'Andrea (2020). Our analysis suggests, however, that there may not be any systematic SST bias across all models that explains biases in blocking/jet variability. In our view, a more likely explanation is that reductions in SST biases in coupled models may so far have been too modest to have a big effect. While Athanasiadis et al. (2022) show that there have been some significant reductions in SST biases with resolution, the most dominant bias of the North Atlantic cold spot still remains, with the vast majority of models exhibiting a cold-spot bias of around 2 or more degrees (see Fig. 14f). Note that this cold-spot bias was not found to be significantly linked to either atmospheric or oceanic resolution in our ensemble, with correlations in both cases of 0.14 and −0.03, respectively.

The possibility that reductions in North Atlantic SST biases have so far been very modest in coupled models is arguably not surprising. The key source of ocean-driven biases in the North Atlantic arises from a failure to simulate the path of the Gulf Stream correctly. While Michel et al. (2023) argue there are some improvements to the simulated trajectory with increased ocean resolution, Tsartsali et al. (2022) conclude that, of all the HighResMIP models they consider, only the one model using a 1/12° ocean is able to simulate a more realistic Gulf Stream separation. The evidence that biases in CMIP models have been significantly reduced due to a more realistic Gulf Stream therefore seems unclear given that virtually no models to date have run at such a high ocean resolution. The low/negligible correlations we find here (Fig. 14) corroborate this ambiguity.

c. *Coupled versus AMIP models and the relative roles of atmospheric and oceanic resolution*

Both Athanasiadis et al. (2022) and Michel et al. (2023) argue that the improvements seen in blocking/jet variability in coupled models with increased model resolution are primarily a result of increased ocean resolution. An important quantitative argument rests in both cases on the observation that the bias reduction associated with increased resolution appears to be smaller in AMIP models than in coupled models, which has been interpreted as evidence for the dominant role of the ocean resolution. On the other hand, it is noteworthy that AMIP models by and large have similar biases in blocking to coupled models (Schiemann et al. 2020), meaning that artificially reducing the SST biases to zero does not actually eliminate blocking biases. This strongly suggests that high oceanic resolution alone is not sufficient, and increased atmospheric resolution must also be playing some role in alleviating biases. Indeed, Schiemann et al. (2020) suggest that the similar biases in AMIP and coupled models may indicate the importance of air–sea coupling which requires high oceanic *and* atmospheric resolution (Moreton et al. 2021).

If atmospheric resolution plays any role whatsoever at alleviating biases, then it follows that the improvement with resolution must be smaller in AMIP models than in coupled models. This is because in coupled models the resolution increase happens almost always in both the ocean and the atmosphere, and so the increased resolution in coupled models will be associated with an alleviation of both the atmospheric *and* oceanic resolution-related biases, while in AMIP models the only biases alleviated are ones related to the atmospheric resolution. This difference is likely exacerbated by the suppressed precipitation extremes in AMIP models that we reported in section 4d, especially given the potentially nonlinear relationship between atmospheric resolution and precipitation extremes. In other words, increasing the atmospheric resolution in an AMIP model may have a much more modest impact on precipitation variability compared to the same increase in a coupled model. Even ignoring any potential nonlinearity, AMIP models at 25-km atmospheric resolution still have precipitation extremes well below that of 100-km resolution coupled models (Fig. 8), meaning high-resolution AMIP models still vastly underestimate the real-world extremes. Consequently, the resolution-driven bias reduction in European blocking and northern jet excursions may be more modest in AMIP models. For the same reason, AMIP experiments may also overemphasize the importance of SSTs and underemphasize the importance of diabatic processes.

The fact that AMIP models have lower precipitation extremes but similar blocking frequencies seems at first glance to contradict the importance of Gulf Stream precipitation. Instead, we argue that this may actually help explain the puzzle about why AMIP and coupled models have similar biases. In AMIP models, the SST biases have been eliminated, but the lack of coupling suppresses the precipitation extremes. These factors may cancel each other out and result in no overall change in blocking biases. Another important factor may be the broader role played by air–sea coupling in blocking formation and maintenance, as laid out in Mathews and Czaja (2024), though it is not clear to us if the lack of coupling in AMIP models would enhance or suppress the mechanisms they describe. What is in any case clear is that a direct comparison between AMIP and coupled models is not straightforward.

We thus argue that it is hard to conclude that oceanic resolution has played a more important role than atmospheric resolution in models to date. It does not even seem easy to rule out the possibility that the effect of increased ocean resolution has so far been minimal. After all, increased atmospheric resolution by itself will enhance precipitation variability (Scher et al. 2017; Schemm 2023), and improvements to blocking/jet variability would themselves be expected to reduce SST biases, via air–sea coupling. Previously reported correlations between blocking/jet variability and SST biases, such as those in Athanasiadis et al. (2022), could therefore be mediated by Gulf Stream precipitation variability which in turn may be set primarily by the atmospheric resolution.

It is important to add that Athanasiadis et al. (2022) do provide additional evidence for the role of SST biases beyond the comparison between AMIP and coupled models. In particular, they compared pairs of simulations for which either only the atmospheric resolution or only the oceanic resolution has been increased and found that the latter pairs showed a greater bias reduction in blocking. However, there are only four such pairs each, and there is substantial internal variability in blocking frequencies across multiple realizations from the same model, even when using 65-yr climatologies: See Fig. S9. It is therefore not clear if strong conclusions can be drawn from such a small number of pairs.

More detailed analysis on the relative role of atmospheric versus oceanic resolution here goes beyond the scope of the present paper but is of clear interest for future work.

d. Shortcomings of our work

A key shortcoming of our analysis is the lack of process-based analysis. In particular, we have not attempted to further our detailed understanding of the relevant physical processes involved in translating Gulf Stream precipitation to European blocking and northern jets and how these processes are simulated in models. One prominent way this shortcoming affects our results is the uncertainty introduced by the effect of different convection parameterizations in models. Figure 7a shows that, for a fixed horizontal resolution, different subsets of the ensemble sometimes appear to cluster around very different values of GS PR99. While we cannot rule out a potential influence of decadal modes of variability, which will vary across the models, a pragmatic explanation is that this clustering is highlighting different realizations from the same model or models with similar convection parameterization. A sensitivity to the convection parameterization was also reported in Strandberg and Lind (2021). If Gulf Stream precipitation extremes really are a primary source of intermodel spread in blocking/northern jets, then the similar "clustering" behavior seen in Figs. 7b and 7c at different fixed resolutions may also be partially due to differences in the convection parameterizations. Accounting for these differences would probably require the kind of careful process-based analysis that our paper omits. However, our work here offers a complementary view to the increasing body of such detailed studies, by showing that even quite crude measures suffice to establish robust correlations in multimodel ensembles. It is therefore possible that a more carefully chosen metric of the relevant diabatic processes would result in even higher correlations than those seen in Fig. 6.

Another clear weakness of our analysis is that our choice to pool all models together might obscure genuine differences across, e.g., successive generations of climate models. There is, for example, no reason a priori for the biases in high-resolution HighResMIP models and coarse-resolution CMIP5 models to share the same underlying cause. Our results show that Gulf Stream precipitation variability could be such a common cause, but the correlations we computed would, viewed in isolation, not be sufficient evidence to draw a strong conclusion. However, when viewed in combination with the several targeted experiments and case studies assessing the role of diabatic processes in the Gulf Stream on blocking, it becomes more plausible that our analysis is highlighting a genuine relationship.

It should also be noted that increased resolution (in both atmosphere and ocean) is typically associated with many changes to the large-scale circulation beyond blocking or northern jets. Some of these are discussed in Athanasiadis et al. (2022) and others in Dorrington et al. (2022). One relevant such change is the overall improved representation of the entire jet latitude distribution, not just the northern mode, as highlighted in Dorrington et al. (2022). Such changes could also play a role in the improvements seen here and are unlikely to be attributable purely to changes in precipitation, but we have made no attempt to account for all of these.

Finally, when assessing the role of horizontal resolution, we used the nominal grid resolution as our index, which tends to be a systematic overestimate of the effective resolution (Klaver et al. 2020). This could potentially bias our analysis. It would be valuable if modeling centers offered an explicit estimate of the effective resolution of all their CMIP-related simulations as a matter of course.

6. Summary and conclusions

We present a brief synopsis of our results:

1) A simple measure of climatological Gulf Stream precipitation extremes is shown to be strongly associated with both climatological European blocking frequency (correlation of 0.63) and northern jet days (correlation of 0.55) in a large coupled multimodel ensemble (Fig. 6). In the case of European blocking, as much as 40% of the intermodel spread is associated with variations in precipitation extremes. Models experience particularly severe biases in precipitation extremes.
2) Gulf Stream precipitation extremes go up with atmospheric horizontal resolution (Fig. 7), possibly in a superlinear manner. Statistical tests for mediation corroborate a hypothesis that the reduced biases in blocking and northern jet days seen with increased model resolution are mediated by enhanced precipitation extremes.

3) In the model ensemble, composite analysis suggests days with extreme Gulf Stream rainfall are associated with larger cyclones that experience greater poleward propagation, compared to days with low/moderate rainfall (Fig. 9). In ERA5, days following extreme rainfall are far more strongly associated with intense anticyclonic wavebreaking compared to days with low/moderate rainfall (Fig. 12).
4) AMIP models exhibit systematically reduced precipitation variability compared to coupled models, due to the "excess thermal damping" mechanism described by BB98 (Fig. 8). This complicates comparisons between AMIP and coupled models.
5) Important SST biases, such as Gulf Stream SST gradients or the North Atlantic cold SST bias, are found to share only weak or negligible correlations with European blocking frequencies and northern jet days (Fig. 14).

We interpret our results as corroborating the hypothesis, put forth in Schemm (2023), that biases in diabatic processes in the Gulf Stream region are a key source of biases in the North Atlantic eddy-driven jet and furthermore that increasing the atmospheric horizontal resolution of models reduces jet biases by better simulating these diabatic processes. We add to Schemm's work, and that of De Luca et al. (2024), by showing that biases in European blocking are also closely linked to Gulf Stream precipitation biases. We have also emphasized the role of extreme precipitation. In particular, we hypothesize that models which can simulate more extreme Gulf Stream precipitation are able to more readily generate intense anticyclonic wavebreaking, and hence more frequently trigger the onset of European blocking episode, because the stronger latent heat release occurring during intense precipitation allows the outflow of negative PV along a cyclone warm conveyor belt to occur at higher levels. For the same reason, we hypothesize that models which simulate more intense precipitation can more readily simulate explosive cyclogenesis and hence produce more poleward-propagating cyclones, as Schemm suggests.

The most important limitation of our analysis is that we made no attempt at robustly isolating the causal influence of precipitation in our multimodel ensemble, and the apparent link we find between precipitation extremes and blocking could be mediated by some other resolution-dependent physical process, such as the vertical motions highlighted in Wills et al. (2024). Our results must therefore be viewed in combination with evidence from idealized experiments (Mathews et al. 2024; De Luca et al. 2024) and process-oriented case studies (Wenta et al. 2024). Our work adds to these by showing that diabatic processes in the Gulf Stream region are not just important for understanding the variability of particular events or in particular models but also for the question of understanding intermodel spread and the impact of model resolution.

While previous studies have tended to focus on the role of SST biases and oceanic horizontal resolution, studies such as Scher et al. (2017), Schemm (2023), and De Luca et al. (2024) show that atmospheric resolution alone can improve both precipitation variability, the jet and blocking. Given that (i) we find a weak or negligible link between SST biases and jet variability in our ensemble and (ii) previous arguments based on a comparison of AMIP and coupled models are problematic in light of the systematically suppressed precipitation variability experienced by AMIP models, we speculate that improvements to metrics like European blocking frequency with model resolution, reported in Davini and d'Andrea (2020) and reproduced here, are primarily due to increased atmospheric resolution. We do expect that reducing the SST biases also helps, but our analysis suggests that coupled models have to date only experienced a modest reduction in SST biases, likely owing to the apparent need for an oceanic resolution of 1/12th of a degree or higher to simulate a genuinely realistic Gulf Stream separation (Tsartsali et al. 2022). A crucial caveat to this speculation is that the role of oceanic resolution and SST biases could be greater for higher-resolution models and lower for coarse-resolution models. Some evidence is found to support this (see section 5c), consistent with the evidence put forth in Athanasiadis et al. (2022). Further analysis would be required to unravel these points.

It must be emphasized at this point that the Gulf Stream per se is conspicuously absent in our analysis. That is, we have looked at the precipitation variability in the Gulf Stream region, but we have not linked this variability to, e.g., oceanic heat transport or the SST gradients created by the Gulf Stream. Consideration of lagged influences of precipitation with such quantities could help clarify the role played by the Gulf Stream itself, beyond just being a source of warm water for the atmosphere to feed on. The work of Mathews and Czaja (2024) strongly suggests that one should view the system as a fundamentally coupled one, and a natural avenue for future work is to look at how this coupling is simulated in models.

We end with some thoughts about what might happen when we reach the kilometer-scale models championed by Slingo et al. (2022) and others. This will at least partially depend on whether IMERG or TRMM is a better estimate of reality. If TRMM is the better estimate, then a considerable step change increase in precipitation extremes may well take place. One interesting possibility here is that resolving mesoscale eddies in the ocean may profoundly impact air–sea fluxes in models, as hinted at by recent work (Yang et al. 2024). This could end up playing an important role in any such step changes. Less clear is what the impact will be of models simulating a realistic Gulf Stream (including the northwest corner) *en masse*, since this could expose the effect of compensating biases or tuning efforts made to suppress the effect of North Atlantic SST biases. Here too, though earlier work suggests that very high-resolution models may behave quite differently, including the simulated circulation response to diabatic heating anomalies in the Gulf Stream (Famooss Paolini et al. 2022). Coupled model experiments aimed at understanding the impact of resolving mesoscale eddies on the climate system are the goal of the "EERIE" Project (https://doi.org/10.3030/101081383). One may therefore hope that some of these questions will begin to be resolved over the coming years.

Acknowledgments. This publication is part of the EERIE project (Grant Agreement 101081383) funded by the European

Union. Views and opinions expressed are, however, those of the author(s) only and do not necessarily reflect those of the European Union or the European Climate Infrastructure and Environment Executive Agency (CINEA). Neither the European Union nor the granting authority can be held responsible for them. University of Oxford's contribution to EERIE is funded by U.K. Research and Innovation (UKRI) under the U.K. government's Horizon Europe funding guarantee, Grant 10049639. All three authors therefore acknowledge funding from this grant. HMC was further funded by Natural Environment Research Council Grant NE/P018238/1. K. S. thanks Fenwick Cooper for several helpful conversations about precipitation in models versus reality, as well as Matthew Patterson for helpful feedback on an initial draft. We thank two anonymous reviewers and Akira Kuwano-Yoshida for invaluable feedback and suggestions. We are indebted to Severin Kaderli for producing the ERA5 wavebreaking dataset and to Hugo Banderier for assistance in obtaining it. We acknowledge the World Climate Research Programme, which, through its Working Group on Coupled Modelling, coordinated and promoted CMIP6. We thank the climate modeling groups for producing and making available their model output, the Earth System Grid Federation (ESGF) for archiving the data and providing access, and the multiple funding agencies that support CMIP6 and ESGF.

Data availability statement. ERA5 data are publicly available via the Copernicus Data Store (Hersbach et al. 2023). IMERG version 7 data are available courtesy of NASA. The IMERG version 7 data were provided by the NASA/Goddard Space Flight Center's precipitation processing center, which develops and computes IMERG as a contribution to GPM, and archived at the NASA GES DISC (Huffman et al. 2024). CMIP6 data (Eyring et al. 2016) can be freely downloaded from the ESGF at https://esgf-node.llnl.gov/projects/cmip6/. CMIP5 and HighResMIP data can similarly be downloaded from the ESGF.

APPENDIX

A Simple Model Explaining the Effect of Air–Sea Coupling

We give an informal and brief description of the statistical model of BB98 used to represent a coupled atmosphere–ocean system and an AMIP-style counterpart. We do this in order to keep the present paper self-contained and to draw renewed attention to the importance of BB98 for studies comparing coupled and AMIP models. Readers should refer to BB98 for more details, including how the parameter values were chosen.

Let T_a denote anomalous near-surface air temperatures and T_o anomalous sea surface temperatures, both thought of as being averaged over a generic North Atlantic domain. Then, BB98 model the evolution of T_a and T_o as a system of coupled stochastic differential equations:

$$\frac{d}{dt}T_a = AT_a + BT_o + \epsilon, \tag{A1}$$

$$\frac{d}{dt}T_o = CT_a + DT_o. \tag{A2}$$

Here A, B, C, and D are the constant coefficients and ε is a stochastic noise term assumed to be normally distributed with mean zero. Such models have been used extensively in the literature to capture atmosphere–ocean coupling (see, e.g., Penland and Sardeshmukh 1995; Penland and Magorian 1993; Alexander et al. 2008; Hawkins and Sutton 2009; Newman et al. 2009). The coefficients A and D, which are necessarily negative in a stable system, represent damping terms which relax T_a and T_o back to their mean (i.e., zero), while the coefficients B and C account for coupling.

Using the notation from BB98, we have $A = -(\lambda_{sa}c + \lambda_a)/\gamma_a$, $B = b\lambda_{sa}/\gamma_a$, $C = c\lambda_{so}/\gamma_o$, and $D = -(\lambda_o + \lambda_{so})/\gamma_o$. The unknown parameters here are given in Table 1 of BB98 as follows: $\gamma_a = 1 \times 10^{-7}$, $\gamma_o = 2 \times 10^{-8}$, $\lambda_{sa} = 23.9$, $\lambda_{so} = 23.4$, $\lambda_a = 2.8$, $\lambda_o = 1.9$, $c = 1$, and $b = 0.5$. After rescaling the entire system by 10^6, this gives $A = -2.67$, $B = 1.195$, $C = 0.117$, and $D = -0.126$. The standard deviation of the noise in BB98 is given as λ_{sa}/γ_a, which after rescaling by 10^6 is 2.39.

One may now readily simulate a long time series of the fully coupled system by integrating Eqs. (A1) and (A2), yielding in particular a time series of simulated SSTs T_o^{CPL}. To generate an AMIP-style counterpart, following BB98, we simply integrate Eq. (A1) where at every time step t we put $T_o(t) = T_o^{\mathrm{CPL}}(t)$. We carried this out, integrating the system for 50000 time steps, and then computed lag correlations of SSTs (T_o) and T2M (T_a) for the coupled and AMIP-style systems. The result is shown in Fig. A1, reproducing one of the key results of BB98. Comparison with Fig. 8b confirms the conclusions drawn already in BB98, that lack of coupling in a model has a profound impact on air–sea interactions, not just when T2M leads SSTs (which is a priori obvious) but also when SSTs lead T2M (see the discrepancy between the blue and red lines for lags between −10 and 0).

BB98 furthermore show that the lack of coupling leads to a systematic reduction in daily T2M variability. BB98

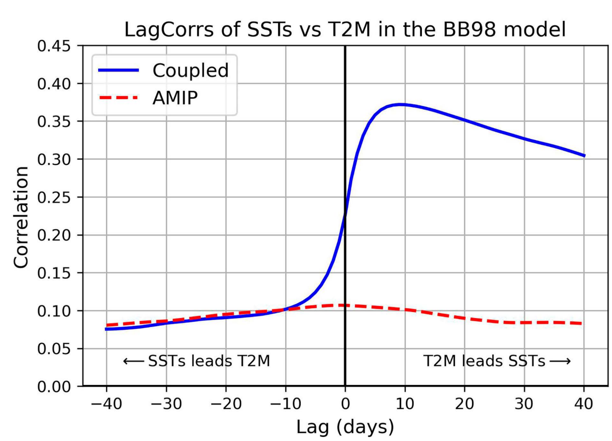

FIG. A1. Lead–lag correlations of daily SST vs daily T2M based on the BB98 model; the coupled version is plotted in solid blue, and the AMIP-style version is plotted in stippled red. Negative lags correspond to SSTs leading T2M, and positive lags correspond to T2M leading SSTs.

explain this result, as well as the change in lag correlations, as being due to the excessive thermal damping experienced by AMIP models. In the coupled system, if T_a and T_o are positive (negative) on a given time step, the coupling terms BT_o and CT_a will act to try to persist the positive (negative) anomalies on the next time step as well, opposing the effect of the damping terms (which act to remove the anomaly), thereby increasing the likelihood of very persistent anomalies occurring. The coupling terms thus serve to enhance the low-frequency variability of both T_a and T_o; when this coupling is lost, such as in AMIP-style simulations, this additional variability is lost.

A heuristic "physical" explanation for this mathematical effect comes from the consideration of heat fluxes. The thermal component of the anomalous heat flux is proportional to $T_a - T_o$, up to some sign convention. If one assumes that T_o is constant and zero (as in a simulation using fixed climatological SSTs), this implies that the sign of the anomalous thermal heat flux is determined purely by the sign of T_a and will thus always act to oppose the anomaly. If $T_a > 0$, the anomalous heat flux will point down (i.e., the atmosphere will, in an anomalous sense, donate extra heat to the ocean), while if $T_a < 0$, the anomalous heat flux will point up (the atmosphere will extract extra heat from the ocean). In AMIP-style simulations, the SSTs are not fixed, but because the SSTs are varying independently of T2M, the same heuristic essentially applies *on average*. This is the sense in which BB98 refer to AMIP models as experiencing excessive thermal damping.

REFERENCES

Abdelmoaty, H. M., S. M. Papalexiou, C. R. Rajulapati, and A. AghaKouchak, 2021: Biases beyond the mean in CMIP6 extreme precipitation: A global investigation. *Earth's Future*, **9**, e2021EF002196, https://doi.org/10.1029/2021EF002196.

Alexander, M. A., L. Matrosova, C. Penland, J. D. Scott, and P. Chang, 2008: Forecasting Pacific SSTs: Linear inverse model predictions of the PDO. *J. Climate*, **21**, 385–402, https://doi.org/10.1175/2007JCLI1849.1.

Ambaum, M. H., and L. Novak, 2014: A nonlinear oscillator describing storm track variability. *Quart. J. Roy. Meteor. Soc.*, **140**, 2680–2684, https://doi.org/10.1002/qj.2352.

Annan, J. D., and J. C. Hargreaves, 2017: On the meaning of independence in climate science. *Earth Syst. Dyn.*, **8**, 211–224, https://doi.org/10.5194/esd-8-211-2017.

Anstey, J. A., and Coauthors, 2013: Multi-model analysis of Northern Hemisphere winter blocking: Model biases and the role of resolution. *J. Geophys. Res. Atmos.*, **118**, 3956–3971, https://doi.org/10.1002/jgrd.50231.

Athanasiadis, P. J., and Coauthors, 2022: Mitigating climate biases in the midlatitude North Atlantic by increasing model resolution: SST gradients and their relation to blocking and the jet. *J. Climate*, **35**, 6985–7006, https://doi.org/10.1175/JCLI-D-21-0515.1.

Bador, M., and Coauthors, 2020: Impact of higher spatial atmospheric resolution on precipitation extremes over land in global climate models. *J. Geophys. Res. Atmos.*, **125**, e2019JD032184, https://doi.org/10.1029/2019JD032184.

Baron, R. M., and D. A. Kenny, 1986: The moderator–mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *J. Pers. Soc. Psychol.*, **51**, 1173–1182, https://doi.org/10.1037//0022-3514.51.6.1173.

Barsugli, J. J., and D. S. Battisti, 1998: The basic effects of atmosphere–ocean thermal coupling on midlatitude variability. *J. Atmos. Sci.*, **55**, 477–493, https://doi.org/10.1175/1520-0469(1998)055<0477:TBEOAO>2.0.CO;2.

Boé, J., 2018: Interdependency in multimodel climate projections: Component replication and result similarity. *Geophys. Res. Lett.*, **45**, 2771–2779, https://doi.org/10.1002/2017GL076829.

Brayshaw, D. J., B. Hoskins, and M. Blackburn, 2011: The basic ingredients of the North Atlantic storm track. Part II: Sea surface temperatures. *J. Atmos. Sci.*, **68**, 1784–1805, https://doi.org/10.1175/2011JAS3674.1.

Büeler, D., L. Ferranti, L. Magnusson, J. F. Quinting, and C. M. Grams, 2021: Year-round sub-seasonal forecast skill for Atlantic–European weather regimes. *Quart. J. Roy. Meteor. Soc.*, **147**, 4283–4309, https://doi.org/10.1002/qj.4178.

Catto, J. L., C. Jakob, G. Berry, and N. Nicholls, 2012: Relating global precipitation to atmospheric fronts. *Geophys. Res. Lett.*, **39**, L10805, https://doi.org/10.1029/2012GL051736.

Cavanaugh, N. R., A. Gershunov, A. K. Panorska, and T. J. Kozubowski, 2015: The probability distribution of intense daily precipitation. *Geophys. Res. Lett.*, **42**, 1560–1567, https://doi.org/10.1002/2015GL063238.

Chang, E. K. M., S. Lee, and K. L. Swanson, 2002: Storm track dynamics. *J. Climate*, **15**, 2163–2183, https://doi.org/10.1175/1520-0442(2002)015<02163:STD>2.0.CO;2.

Chassignet, E. P., and D. P. Marshall, 2008: Gulf Stream separation in numerical ocean models. *Ocean Modeling in an Eddying Regime*, Geophysical Monograph Series, Vol. **177**, Amer. Geophys. Union, 39–61, https://doi.org/10.1029/177GM05.

Chen, D., A. Dai, and A. Hall, 2021: The convective-to-total precipitation ratio and the "drizzling" bias in climate models. *J. Geophys. Res. Atmos.*, **126**, e2020JD034198, https://doi.org/10.1029/2020JD034198.

Colucci, S. J., 1985: Explosive cyclogenesis and large-scale circulation changes: Implications for atmospheric blocking. *J. Atmos. Sci.*, **42**, 2701–2717, https://doi.org/10.1175/1520-0469(1985)042<2701:ECALSC>2.0.CO;2.

Dagan, G., P. Stier, and D. Watson-Parris, 2019: Analysis of the atmospheric water budget for elucidating the spatial scale of precipitation changes under climate change. *Geophys. Res. Lett.*, **46**, 10 504–10 511, https://doi.org/10.1029/2019GL084173.

Davini, P., and F. D'Andrea, 2016: Northern Hemisphere atmospheric blocking representation in global climate models: Twenty years of improvements? *J. Climate*, **29**, 8823–8840, https://doi.org/10.1175/JCLI-D-16-0242.1.

——, and ——, 2020: From CMIP3 to CMIP6: Northern Hemisphere atmospheric blocking simulation in present and future climate. *J. Climate*, **33**, 10 021–10 038, https://doi.org/10.1175/JCLI-D-19-0862.1.

——, F. Fabiano, and I. Sandu, 2022: Orographic resolution driving the improvements associated with horizontal resolution increase in the Northern Hemisphere winter mid-latitudes. *Wea. Climate Dyn.*, **3**, 535–553, https://doi.org/10.5194/wcd-3-535-2022.

De Luca, P., B. Jiménez-Esteve, L. Degenhardt, S. Schemm, and S. Pfahl, 2024: Enhanced blocking frequencies in very-high resolution idealized climate model simulations. *Geophys. Res. Lett.*, **51**, e2024GL111016, https://doi.org/10.1029/2024GL111016.

Dorrington, J., K. Strommen, and F. Fabiano, 2022: Quantifying climate model representation of the wintertime Euro-Atlantic circulation using geopotential-jet regimes. *Wea. Climate Dyn.*, **3**, 505–533, https://doi.org/10.5194/wcd-3-505-2022.

Doyle, J. D., and M. A. Shapiro, 1999: Flow response to large-scale topography: The Greenland tip jet. *Tellus*, **51A**, 728–748, https://doi.org/10.3402/tellusa.v51i5.14471.

Drews, A., R. J. Greatbatch, H. Ding, M. Latif, and W. Park, 2015: The use of a flow field correction technique for alleviating the North Atlantic cold bias with application to the Kiel Climate Model. *Ocean Dyn.*, **65**, 1079–1093, https://doi.org/10.1007/s10236-015-0853-7.

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

Famooss Paolini, L., P. J. Athanasiadis, P. Ruggieri, and A. Bellucci, 2022: The atmospheric response to meridional shifts of the Gulf Stream SST front and its dependence on model resolution. *J. Climate*, **35**, 6007–6030, https://doi.org/10.1175/JCLI-D-21-0530.1.

Haarsma, R. J., and Coauthors, 2016: High Resolution Model Intercomparison Project (HighResMIP v1.0) for CMIP6. *Geosci. Model Dev.*, **9**, 4185–4208, https://doi.org/10.5194/gmd-9-4185-2016.

Harvey, B. J., P. Cook, L. C. Shaffrey, and R. Schiemann, 2020: The response of the Northern Hemisphere storm tracks and jet streams to climate change in the CMIP3, CMIP5, and CMIP6 climate models. *J. Geophys. Res. Atmos.*, **125**, e2020JD032701, https://doi.org/10.1029/2020JD032701.

Hawcroft, M. K., L. C. Shaffrey, K. I. Hodges, and H. F. Dacre, 2012: How much Northern Hemisphere precipitation is associated with extratropical cyclones? *Geophys. Res. Lett.*, **39**, L24809, https://doi.org/10.1029/2012GL053866.

Hawkins, E., and R. Sutton, 2009: Decadal predictability of the Atlantic Ocean in a coupled GCM: Forecast skill and optimal perturbations using linear inverse modeling. *J. Climate*, **22**, 3960–3978, https://doi.org/10.1175/2009JCLI2720.1.

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

——, and Coauthors, 2023: ERA5 monthly averaged data on single levels from 1940 to present. Copernicus Climate Change Service (C3S) Climate Data Store (CDS), accessed 1 June 2024, https://doi.org/10.24381/cds.f17050d7.

Hoskins, B. J., and P. J. Valdes, 1990: On the existence of storm-tracks. *J. Atmos. Sci.*, **47**, 1854–1864, https://doi.org/10.1175/1520-0469(1990)047<1854:OTEOST>2.0.CO;2.

Hourdin, F., and Coauthors, 2017: The art and science of climate model tuning. *Bull. Amer. Meteor. Soc.*, **98**, 589–602, https://doi.org/10.1175/BAMS-D-15-00135.1.

Huffman, G. J., and J. Tan, 2023: The climate data guide: IMERG precipitation algorithm and the Global Precipitation Measurement (GPM) mission. NCAR, accessed 12 September 2024, https://climatedataguide.ucar.edu/climate-data/gpm-global-precipitation-measurement-mission.

——, and Coauthors, 2020: Integrated Multi-Satellite Retrievals for the Global Precipitation Measurement (GPM) Mission (IMERG). *Satellite Precipitation Measurement: Volume 1*, V. Levizzani et al., Eds., Springer, 343–353.

——, and Coauthors, 2023: IMERG V07 release notes. 11 pp., https://gpm.nasa.gov/sites/default/files/2023-07/IMERG_V07_ReleaseNotes_final_230713.pdf.

——, and Coauthors, 2024: Integrated Multi-Satellite Retrievals for GPM (IMERG), version 0.7. NASA's Precipitation Processing Center, accessed 12 September 2024, https://doi.org/10.25966/gn7e-5b32.

Hurrell, J. W., and H. Van Loon, 1997: Decadal variations in climate associated with the North Atlantic Oscillation. *Climatic Change*, **36**, 301–326, https://doi.org/10.1023/A:1005314315270.

Kaderli, S., 2023: skaderli/WaveBreaking: WaveBreaking v0.3.7. Zenodo, https://doi.org/10.5281/zenodo.8123188.

Kautz, L.-A., O. Martius, S. Pfahl, J. G. Pinto, A. M. Ramos, P. M. Sousa, and T. Woollings, 2022: Atmospheric blocking and weather extremes over the Euro-Atlantic sector—A review. *Wea. Climate Dyn.*, **3**, 305–336, https://doi.org/10.5194/wcd-3-305-2022.

Kawanishi, T., and Coauthors, 2000: TRMM precipitation radar. *Adv. Space Res.*, **25**, 969–972, https://doi.org/10.1016/S0273-1177(99)00932-1.

Keeley, S. P., R. T. Sutton, and L. C. Shaffrey, 2012: The impact of North Atlantic sea surface temperature errors on the simulation of North Atlantic European region climate. *Quart. J. Roy. Meteor. Soc.*, **138**, 1774–1783, https://doi.org/10.1002/qj.1912.

Klaver, R., R. Haarsma, P. L. Vidale, and W. Hazeleger, 2020: Effective resolution in high resolution global atmospheric models for climate studies. *Atmos. Sci. Lett.*, **21**, e952, https://doi.org/10.1002/asl.952.

Knutti, R., D. Masson, and A. Gettelman, 2013: Climate model genealogy: Generation CMIP5 and how we got there. *Geophys. Res. Lett.*, **40**, 1194–1199, https://doi.org/10.1002/grl.50256.

Kopparla, P., E. M. Fischer, C. Hannay, and R. Knutti, 2013: Improved simulation of extreme precipitation in a high-resolution atmosphere model. *Geophys. Res. Lett.*, **40**, 5803–5808, https://doi.org/10.1002/2013GL057866.

Madonna, E., H. Wernli, H. Joos, and O. Martius, 2014: Warm conveyor belts in the ERA-Interim dataset (1979–2010). Part I: Climatology and potential vorticity evolution. *J. Climate*, **27**, 3–26, https://doi.org/10.1175/JCLI-D-12-00720.1.

Mathews, J., and A. Czaja, 2024: Oceanic maintenance of atmospheric blocking in wintertime in the North Atlantic. *Climate Dyn.*, **62**, 6159–6172, https://doi.org/10.1007/s00382-024-07196-0.

Mathews, J. P., A. Czaja, F. Vitart, and C. Roberts, 2024: Gulf Stream moisture fluxes impact atmospheric blocks throughout the Northern Hemisphere. *Geophys. Res. Lett.*, **51**, e2024GL108826, https://doi.org/10.1029/2024GL108826.

Matsueda, M., and T. N. Palmer, 2018: Estimates of flow-dependent predictability of wintertime Euro-Atlantic weather regimes in medium-range forecasts. *Quart. J. Roy. Meteor. Soc.*, **144**, 1012–1027, https://doi.org/10.1002/qj.3265.

Messori, G., and A. Czaja, 2013: On the sporadic nature of meridional heat transport by transient eddies. *Quart. J. Roy. Meteor. Soc.*, **139**, 999–1008, https://doi.org/10.1002/qj.2011.

Methven, J., 2015: Potential vorticity in warm conveyor belt outflow. *Quart. J. Roy. Meteor. Soc.*, **141**, 1065–1071, https://doi.org/10.1002/qj.2393.

Michel, S. L. L., A. S. von der Heydt, R. M. van Westen, M. L. Baatsen, and H. A. Dijkstra, 2023: Increased wintertime European atmospheric blocking frequencies in General Circulation Models with an eddy-permitting ocean. *npj Climate Atmos. Sci.*, **6**, 50, https://doi.org/10.1038/s41612-023-00372-9.

Minobe, S., A. Kuwano-Yoshida, N. Komori, S.-P. Xie, and R. J. Small, 2008: Influence of the Gulf Stream on the troposphere. *Nature*, **452**, 206–209, https://doi.org/10.1038/nature06690.

Moreno-Chamarro, E., and Coauthors, 2022: Impact of increased resolution on long-standing biases in HighResMIP-PRIMAVERA climate models. *Geosci. Model Dev.*, **15**, 269–289, https://doi.org/10.5194/gmd-15-269-2022.

Moreton, S., D. Ferreira, M. Roberts, and H. Hewitt, 2021: Air-sea turbulent heat flux feedback over mesoscale eddies. *Geophys. Res. Lett.*, **48**, e2021GL095407, https://doi.org/10.1029/2021GL095407.

Nakamura, H., and J. M. Wallace, 1993: Synoptic behavior of baroclinic eddies during the blocking onset. *Mon. Wea. Rev.*, **121**, 1892–1903, https://doi.org/10.1175/1520-0493(1993)121<1892:SBOBED>2.0.CO;2.

Newman, M., P. D. Sardeshmukh, and C. Penland, 2009: How important is air–sea coupling in ENSO and MJO evolution? *J. Climate*, **22**, 2958–2977, https://doi.org/10.1175/2008JCLI2659.1.

Novak, L., M. H. P. Ambaum, and R. Tailleux, 2017: Marginal stability and predator–prey behaviour within storm tracks. *Quart. J. Roy. Meteor. Soc.*, **143**, 1421–1433, https://doi.org/10.1002/qj.3014.

O'Reilly, C. H., S. Minobe, and A. Kuwano-Yoshida, 2016: The influence of the Gulf Stream on wintertime European blocking. *Climate Dyn.*, **47**, 1545–1567, https://doi.org/10.1007/s00382-015-2919-0.

——, ——, A. Kuwano-Yoshida, and T. Woollings, 2017: The Gulf Stream influence on wintertime North Atlantic jet variability. *Quart. J. Roy. Meteor. Soc.*, **143**, 173–183, https://doi.org/10.1002/qj.2907.

Parker, T., T. Woollings, A. Weisheimer, C. O'Reilly, L. Baker, and L. Shaffrey, 2019: Seasonal predictability of the winter North Atlantic Oscillation from a jet stream perspective. *Geophys. Res. Lett.*, **46**, 10 159–10 167, https://doi.org/10.1029/2019GL084402.

Penland, C., and T. Magorian, 1993: Prediction of Niño 3 sea surface temperatures using linear inverse modeling. *J. Climate*, **6**, 1067–1076, https://doi.org/10.1175/1520-0442(1993)006<1067:PONSST>2.0.CO;2.

——, and P. D. Sardeshmukh, 1995: The optimal growth of tropical sea surface temperature anomalies. *J. Climate*, **8**, 1999–2024, https://doi.org/10.1175/1520-0442(1995)008<1999:TOGOTS>2.0.CO;2.

Pfahl, S., C. Schwierz, M. Croci-Maspoli, C. M. Grams, and H. Wernli, 2015: Importance of latent heat release in ascending air streams for atmospheric blocking. *Nat. Geosci.*, **8**, 610–614, https://doi.org/10.1038/ngeo2487.

Poli, P., and Coauthors, 2013: The data assimilation system and initial performance evaluation of the ECMWF pilot reanalysis of the 20th-century assimilating surface observations only (ERA-20C). ERA Report Series 14, 62 pp., https://www.ecmwf.int/en/elibrary/76007-data-assimilation-system-and-initial-performance-evaluation-ecmwf-pilot.

Scaife, A. A., and J. R. Knight, 2008: Ensemble simulations of the cold European winter of 2005–2006. *Quart. J. Roy. Meteor. Soc.*, **134**, 1647–1659, https://doi.org/10.1002/qj.312.

——, and D. Smith, 2018: A signal-to-noise paradox in climate science. *npj Climate Atmos. Sci.*, **1**, 28, https://doi.org/10.1038/s41612-018-0038-4.

——, and Coauthors, 2011: Improved Atlantic winter blocking in a climate model. *Geophys. Res. Lett.*, **38**, L23703, https://doi.org/10.1029/2011GL049573.

Schemm, S., 2023: Toward eliminating the decades-old "too zonal and too equatorward" storm-track bias in climate models. *J. Adv. Model. Earth Syst.*, **15**, e2022MS003482, https://doi.org/10.1029/2022MS003482.

Scher, S., R. J. Haarsma, H. de Vries, S. S. Drijfhout, and A. J. van Delden, 2017: Resolution dependence of extreme precipitation and deep convection over the Gulf Stream. *J. Adv. Model. Earth Syst.*, **9**, 1186–1194, https://doi.org/10.1002/2016MS000903.

Scherrer, S. C., M. Croci-Maspoli, C. Schwierz, and C. Appenzeller, 2006: Two-dimensional indices of atmospheric blocking and their statistical relationship with winter climate patterns in the Euro-Atlantic region. *Int. J. Climatol.*, **26**, 233–249, https://doi.org/10.1002/joc.1250.

Schiemann, R., and Coauthors, 2020: Northern Hemisphere blocking simulation in current climate models: Evaluating progress from the Climate Model Intercomparison Project Phase 5 to 6 and sensitivity to resolution. *Wea. Climate Dyn.*, **1**, 277–292, https://doi.org/10.5194/wcd-1-277-2020.

Slingo, J., and Coauthors, 2022: Ambitious partnership needed for reliable climate prediction. *Nat. Climate Change*, **12**, 499–503, https://doi.org/10.1038/s41558-022-01384-8.

Sprenger, M., and Coauthors, 2017: Global climatologies of Eulerian and Lagrangian flow features based on ERA-Interim. *Bull. Amer. Meteor. Soc.*, **98**, 1739–1748, https://doi.org/10.1175/BAMS-D-15-00299.1.

Steinfeld, D., and S. Pfahl, 2019: The role of latent heating in atmospheric blocking dynamics: A global climatology. *Climate Dyn.*, **53**, 6159–6180, https://doi.org/10.1007/s00382-019-04919-6.

——, M. Boettcher, R. Forbes, and S. Pfahl, 2020: The sensitivity of atmospheric blocking to upstream latent heating—Numerical experiments. *Wea. Climate Dyn.*, **1**, 405–426, https://doi.org/10.5194/wcd-1-405-2020.

Strandberg, G., and P. Lind, 2021: The importance of horizontal model resolution on simulated precipitation in Europe—From global to regional models. *Wea. Climate Dyn.*, **2**, 181–204, https://doi.org/10.5194/wcd-2-181-2021.

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

Tsartsali, E. E., and Coauthors, 2022: Impact of resolution on the atmosphere–ocean coupling along the Gulf Stream in global high resolution models. *Climate Dyn.*, **58**, 3317–3333, https://doi.org/10.1007/s00382-021-06098-9.

Virtanen, P., and Coauthors, 2020: SciPy 1.0: Fundamental algorithms for scientific computing in Python. *Nat. Methods*, **17**, 261–272, https://doi.org/10.1038/s41592-019-0686-2.

Wang, C., L. Zhang, S.-K. Lee, L. Wu, and C. R. Mechoso, 2014: A global perspective on CMIP5 climate model biases. *Nat. Climate Change*, **4**, 201–205, https://doi.org/10.1038/nclimate2118.

Wenta, M., C. M. Grams, L. Papritz, and M. Federer, 2024: Linking Gulf Stream air–sea interactions to the exceptional blocking episode in February 2019: A Lagrangian perspective. *Wea. Climate Dyn.*, **5**, 181–209, https://doi.org/10.5194/wcd-5-181-2024.

Wernli, H., and M. Sprenger, 2007: Identification and ERA-15 climatology of potential vorticity streamers and cutoffs near the extratropical tropopause. *J. Atmos. Sci.*, **64**, 1569–1586, https://doi.org/10.1175/JAS3912.1.

White, R. H., C. Hilgenbrink, and A. Sheshadri, 2019: The importance of Greenland in setting the northern preferred position of the North Atlantic eddy-driven jet. *Geophys. Res. Lett.*, **46**, 14 126–14 134, https://doi.org/10.1029/2019GL084780.

Wills, R. C. J., A. R. Herrington, I. R. Simpson, and D. S. Battisti, 2024: Resolving weather fronts increases the large-scale circulation response to Gulf Stream SST anomalies in variable-resolution CESM2 simulations. *J. Adv. Model. Earth Syst.*, **16**, e2023MS004123, https://doi.org/10.1029/2023MS004123.

Woollings, T., A. Hannachi, and B. Hoskins, 2010: Variability of the North Atlantic eddy-driven jet stream. *Quart. J. Roy. Meteor. Soc.*, **136**, 856–868, https://doi.org/10.1002/qj.625.

——, and Coauthors, 2018a: Blocking and its response to climate change. *Curr. Climate Change Rep.*, **4**, 287–300, https://doi.org/10.1007/s40641-018-0108-z.

——, and Coauthors, 2018b: Daily to decadal modulation of jet variability. *J. Climate*, **31**, 1297–1314, https://doi.org/10.1175/JCLI-D-17-0286.1.

Yamamoto, A., M. Nonaka, P. Martineau, A. Yamazaki, Y.-O. Kwon, H. Nakamura, and B. Taguchi, 2021: Oceanic moisture sources contributing to wintertime Euro-Atlantic blocking. *Wea. Climate Dyn.*, **2**, 819–840, https://doi.org/10.5194/wcd-2-819-2021.

Yamazaki, A., and H. Itoh, 2009: Selective absorption mechanism for the maintenance of blocking. *Geophys. Res. Lett.*, **36**, L05803, https://doi.org/10.1029/2008GL036770.

Yang, H., and Coauthors, 2024: Observations reveal intense air-sea exchanges over submesoscale ocean front. *Geophys. Res. Lett.*, **51**, e2023GL106840, https://doi.org/10.1029/2023GL106840.

Yazdandoost, F., S. Moradian, A. Izadi, and A. Aghakouchak, 2021: Evaluation of CMIP6 precipitation simulations across different climatic zones: Uncertainty and model intercomparison. *Atmos. Res.*, **250**, 105369, https://doi.org/10.1016/j.atmosres.2020.105369.

Ye, K., T. Woollings, S. N. Sparrow, P. A. G. Watson, and J. A. Screen, 2024: Response of winter climate and extreme weather to projected Arctic sea-ice loss in very large-ensemble climate model simulations. *npj Climate Atmos. Sci.*, **7**, 20, https://doi.org/10.1038/s41612-023-00562-5.

Zhang, Q., B. Liu, S. Li, and T. Zhou, 2023: Understanding models' global sea surface temperature bias in mean state: From CMIP5 to CMIP6. *Geophys. Res. Lett.*, **50**, e2022GL100888, https://doi.org/10.1029/2022GL100888.