RESEARCH ARTICLE

Numerical impacts on tracer transport: A proposed intercomparison test of Atmospheric General Circulation Models

Aman Gupta[1] | Edwin P. Gerber[1] | Peter H. Lauritzen[2]

[1]Center for Atmosphere-Ocean Science, Courant Institute of Mathematical Sciences, New York University, New York

[2]National Center for Atmospheric Research, Boulder, Colorado

Correspondence
A. Gupta, Center for Atmosphere-Ocean Science, Courant Institute of Mathematical Sciences, 251 Mercer St #801, New York, NY 10012, USA.
Email: agupta@cims.nyu.edu

Abstract

The transport of trace gases by the atmospheric circulation plays an important role in the climate system and its response to external forcing. Transport presents a challenge for Atmospheric General Circulation Models (AGCMs), as errors in both the resolved circulation and the numerical representation of transport processes can bias their abundance. In this study, two tests are proposed to assess transport by the dynamical core of an AGCM. To separate transport from chemistry, the tests focus on the age-of-air, an estimate of the mean transport time by the circulation. The tests assess the coupled stratosphere–troposphere system, focusing on transport by the overturning circulation and isentropic mixing in the stratosphere, or Brewer–Dobson Circulation, where transport time-scales on the order of months to years provide a challenging test of model numerics. Four dynamical cores employing different numerical schemes (finite-volume, pseudo-spectral, and spectral-element) and discretizations (cubed sphere versus latitude–longitude) are compared across a range of resolutions. The subtle momentum balance of the tropical stratosphere is sensitive to model numerics, and the first intercomparison reveals stark differences in tropical stratospheric winds, particularly at high vertical resolution: some cores develop westerly jets and others easterly jets. This leads to substantial spread in transport, biasing the age-of-air by up to 25% relative to its climatological mean, making it difficult to assess the impact of the numerical representation of transport processes. This uncertainty is removed by constraining the tropical winds in the second intercomparison test, in a manner akin to specifying the Quasi-Biennial Oscillation in an AGCM. The dynamical cores exhibit qualitative agreement on the structure of atmospheric transport in the second test, with evidence of convergence as the horizontal and vertical resolution is increased in a given model. Significant quantitative differences remain, however, particularly between models employing spectral versus finite-volume numerics, even in state-of-the-art cores.

KEYWORDS

age of air, Brewer–Dobson circulation, dynamical cores, stratospheric dynamics, tracer transport.

1 | INTRODUCTION

It has been a quarter of a century since Held and Suarez (1994) proposed a test to compare the dynamical cores of atmospheric general circulation models, the numerical solvers that integrate the primitive equations on the sphere. The Held and Suarez, 1994 test, hereafter referred to as HS94, is a simple recipe for parametrizing all non-conservative processes for a dry (moisture-free) atmosphere, including diabatic processes, such as radiative transfer and convection, and frictional processes, specifically the atmospheric boundary layer. This allowed them to see if two dynamical cores with very different numerics – a pseudo-spectral versus a finite-difference scheme – would produce the same general circulation. It was designed to complement existing tests for dynamical cores, which examine the evolution of numerics on shorter time-scales in more controlled settings, for example, the reproduction of a single baroclinic life-cycle experiment (Williamson *et al.*, 1992). In the ensuing decades, however, the HS94 recipe for producing a fairly realistic climate with such a simple parametrization of diabatic processes has led to substantial amount of research on the dynamics and circulation of the atmosphere (e.g., Son and Lee, 2005; Lorenz and DeWeaver, 2007; Butler *et al.*, 2010).

In this paper, we propose an update of the HS94 benchmark, designed to take into account two advances in climate research over the past 25 years. First, there has been a growing awareness that the tropospheric circulation is tightly coupled with that of the stratosphere, such that climate and weather prediction is improved by accurately representing the stratosphere (e.g., Gerber *et al.*, 2012; Butler *et al.*, 2018; Ayarzagüena *et al.*, 2020). We update the HS94 recipe to allow one to explore the impact of numerics and resolution on the coupled stratosphere–troposphere circulation based on the work of Polvani and Kushner (2002). The test allows us to compare the large-scale circulation and variability of the stratosphere, in particular Sudden Stratospheric Warming (SSW) events.

Second, atmospheric chemistry has become an important element in climate prediction, as evidenced by the AerChemMIP (Collins *et al.*, 2017) within the Coupled Model Intercomparsion Project, Phase 6 (CMIP6; Eyring *et al.*, 2016). Our test seeks to ensure that dynamical cores consistently represent the transport of trace gases by the resolved atmospheric flow. Transport – the advection, mixing, and diffusion of constituents by the atmospheric circulation – plays a fundamental role in setting the distribution of short-lived species that impact the radiative balance of the planet, for example, aerosols and ozone. The distribution of trace constituents depends on both chemistry (their sources and sinks, including those of precursors) and how they are advected and diffused by the atmospheric flow. We focus exclusively on the latter process, isolating the role of model numerics on transport from questions of emissions and atmospheric chemistry.

Trace gases also play a vital role in our ability to observe the atmospheric circulation, particularly in the stratosphere (e.g., Plumb, 2002). *In situ* and satellite-based measurements of trace gases can be inverted to place constraints on dynamical quantities (in particular, vertical velocity) which cannot be observed directly (Linz *et al.*, 2017). Ensuring that AGCMs provide an accurate representation of transport can thus help to both understand and quantify the observed circulation, and predict its response to external forcings.

It has long been established that model numerics play a critical role in transport (e.g., Rood, 1987; Hall *et al.*, 1999). Kent *et al.* (2014) constructed a series of tests to compare transport in dry dynamical cores on short time-scales. To the authors' knowledge, however, there does not exist an equivalent to the Held and Suarez (1994) benchmark test to determine how the underlying numerics and resolution of an AGCM impact the climatological transport properties. Building on the work of Orbe *et al.* (2012), we introduce a tracer to the HS94 set-up which allows one to compute the "age-of-air". As further detailed in Section 2, the age-of-air provides an integrated measure of transit through the free atmosphere, quantifying the average time it takes air to travel from the surface boundary layer to a given point in the atmosphere (Hall *et al.*, 1994).

We focus on the circulation and transport through the stratosphere for two reasons. First, differences in the numerical representation of stratospheric dynamics between dynamical cores has motivated several past studies (e.g., Yao and Jablonowski, 2015; 2016). Subtle differences in momentum transport by waves and the mean flow, and the ability of schemes to conserve angular momentum, can lead to substantial differences in the circulation of the tropical stratosphere related to the Quasi-Bienniel Oscillation (QBO). Second, the absence of small-scale convection enables one to exclusively focus on the large-scale transport by the model's numerical schemes. In the troposphere, the transport depends critically on both the resolved circulation and transport associated with parametrized moist convection (Orbe *et al.*, 2017a; 2017b; Wu *et al.*, 2018).

We review the stratospheric circulation and transport in Section 2, chiefly to justify our selection of the age-of-air as a metric to assess transport. Section 3 then introduces four dynamical cores developed by the Geophysical Fluid Dynamics Laboratory (GFDL) and the National Center for Atmospheric Research (NCAR), which we use to establish a benchmark for the stratosphere–tropospheric circulation and transport. We are particularly interested in the behaviour of the cubed-sphere finite-volume core

developed at GFDL (Putman and Lin, 2007), which recently became the core of fvGFS, the US National Weather Service forecasting system, and the cubed-sphere spectral-element core developed at NCAR (Lauritzen *et al.*, 2018), the newly added dynamical core to the Community Atmosphere Model 5 (CAM5), which is being prepared to replace the current finite-volume-based dynamical core for future CMIP experiments.

Two closely related tests for atmospheric models are proposed in Section 4. As with the HS94 intercomparison test, the goal is to remove the influence of differences in the treatment of unresolved processes, for instance the gravity wave drag and convective parametrizations, on the resolved dynamics and transport, allowing us to focus exclusively on the impact of a model's numerical formulation on the circulation and transport. However, the subtle momentum balance of the tropical stratosphere is a challenge for dynamical cores. In our first test, which is most similar in spirit to the original HS94 recipe, state-of-the-art numerical cores fail to generate a consistent representation of the wind fields in the tropical stratosphere. Differences in the large-scale circulation make it difficult to focus on the role of model numerics on transport. This inspired a second test, where the tropical winds in the stratosphere are specified in a manner akin to specifying the Quasi-Biennial Oscillation, which allows one to focus in on the impact of model numerics on tracer transport.

Section 5 documents the general circulation and transport of the four models under our two tests. A key result is the different behaviour between dynamical cores using spectral and finite-volume numerics. Particularly striking differences in the resolved dynamics between the models emerge as the vertical resolution is increased. This is explored in greater detail in Section 6. We also show that, once the tropical stratospheric circulation is constrained, the models show evidence of a more consistent representation of both the residual, overturning circulation and transport throughout the stratosphere. Finally, we consider the sensitivity of the four cores to vertical and horizontal resolution in Section 7. While explicit convergence of the large-scale circulation is not well defined, we show evidence that modern cores provide fundamentally different answers to stratospheric transport at high resolution. We conclude that trace gas transport still presents a challenge for model development in Section 8.

2 | STRATOSPHERIC TRANSPORT AND THE AGE-OF-AIR

The meridional circulation of the stratosphere draws air up from the tropical troposphere, transporting it poleward through both hemispheres before returning it back to the troposphere in the Extratropics. It is known as the Brewer–Dobson Circulation, as its existence was first inferred by Brewer (1949) and Dobson (1956) based on measurements of trace gases, water vapour and ozone, respectively. The poleward transport is stronger in the winter hemisphere and the asymmetry between the summer and winter hemispheres increases with height. In the upper stratosphere and mesosphere, the circulation becomes a single cell, transporting air from the summer to winter hemisphere; the two cells appear only in the annual mean at these levels. However, this overturning of mass from the Tropics to Extratropics exists only in a Lagrangian sense. The Eulerian mean circulation of the stratosphere mirrors the multicell structure of the troposphere, for example, with a tropical Hadley cell, but Ferrel cells in the midlatitudes, with zonal mean equatorward flow (Murgatroyd and Singleton, 1961). Outside the Tropics, eddies are the primary movers of mass, and are responsible for the poleward transport of air. The overturning of the stratosphere is thus characterized by the residual mean circulation (Andrews and McIntyre, 1976; 1978; Boyd, 1976).

The mean overturning of mass through the stratosphere is often referred to as a "diabatic circulation", as it requires air to be radiatively warmed for it to rise through isentropic (constant potential temperature) surfaces in the Tropics, and radiatively cooled to return downward through them in the higher latitudes. These diabatic processes are not the driver of the circulation, but rather a response to the mechanical forcing that draws the air poleward into both hemispheres (Haynes *et al.*, 1991). Rossby waves, which originate from the Extratropics, exert a negative torque on the stratosphere when they break, slowing the flow down so that it can move across angular momentum surfaces as it travels poleward. As described by Holton *et al.* (1995), these wave torques provide an "extratropical pump" that drives the circulation. The circulation is slow, with a characteristic time-scale of months in the lower stratosphere, extending to years at upper levels.

In addition to driving the slow cross-isentropic advection of air, Rossby wave breaking plays a key role in transporting trace gases on faster time-scales, on the order of days (McIntyre and Palmer, 1983). As Rossby waves break, air is rapidly transported along isentropic surfaces. While this "isentropic mixing" leads to no net transport of mass, it can transport chemical species which exhibit a climatological gradient. Such gradients arise in response to differences in sinks and sources. Ozone, for example, is primarily produced in the tropical stratosphere, but transported poleward by isentropic mixing, and ultimately concentrated at lower levels in the Extratropics by the diabatic circulation (e.g., Butchart, 2014). The transport of trace gases by an atmospheric model is sensitive to the

representation of both the mean overturning circulation and isentropic mixing. An accurate representation of the mean overturning circulation requires a model to capture both the Rossby wave breaking and the heat and momentum fluxes that mix potential vorticity and drive the extratropical pump. Furthermore, given the slow time-scale of this overturning circulation, quasi-vertical diffusion across isentropic surfaces by the numerical scheme can greatly accelerate the slow transport of trace gases upward in the Tropics.

The isentropic mixing of trace gases by Rossby waves in atmospheric models is also sensitive to the details of the numerics. Breaking Rossby waves stretch and shear out filaments of air from the Tropics and Extratropics, continually increasing the gradient between air masses until the point that diffusion can finally smear them out (McIntyre and Palmer, 1984). As diffusive processes in the atmosphere operate on scales well below that of a model's grid of 10 to 100 km, diffusion in "model world" is primarily numerical, not physical. Schemes that are more diffusive will artificially increase the impact of isentropic mixing, particularly across regions of the atmosphere that are isolated from Rossby wave breaking, chiefly the polar vortex of the winter hemisphere and the tropical stratosphere.

To study transport of tracers by both the mean overturning circulation and isentropic mixing, an idealized tracer called *age-of-air* has been used extensively in the literature (e.g., Kida, 1983; Hall and Plumb, 1994; Engel *et al.*, 2009; Hang *et al.*, 2009; Linz *et al.*, 2017; Chabrillat *et al.*, 2018). Waugh and Hall (2002) provide a comprehensive review.

The age-of-air, hereafter "age", quantifies the mean time that has elapsed since air was last in contact with the surface. Depending on the specific formulation, the age is defined relative to the atmospheric boundary layer at the surface, or relative to the time it enters the stratosphere; in this latter case, the entire troposphere is viewed as a boundary layer in contact with the surface. For an infinitesimal, undiluted parcel of air, the age would simply be the time since it left the boundary layer. But for any finite amount of air (or for our purposes, a grid box in a numerical model), the age is the mean over the full spectrum of transit times by all the different infinitesimal parcels of air within it (Hall and Plumb, 1994; Holzer and Hall, 2000). Our focus on age, an integrative measure of tracer transport, thus complements the work of Kent *et al.* (2014), who analyze short-term transport of passive tracers in response to prescribed analytical wind fields.

The age can be represented by a passive tracer with a constant production rate, that is, the "age" increases one unit per unit of time, with a sink only near the surface, where it is reset to zero when air enters the boundary layer. In practice, however, we find it simpler to compute it using a "clock tracer", a tracer that is specified to be linearly increasing at the surface (increasing 1 unit for every second of integration, so that the concentration at the surface equals the current model time) and otherwise passively advected by the model through the free atmosphere (Hall and Plumb, 1994). The age at any grid box in the model is then simply the difference between the model time and concentration at that point; Section 4 gives more details on the implementation in the dynamical cores.

The key is that the clock tracer is transported by the model's advection scheme. Differences in age profiles across models provide information about differences in the representation of transport processes, both resolved (mean Lagrangian transport and isentropic mixing) and diffusive, explicit or numerical.

Comprehensive atmospheric models have been shown to struggle with transport processes. Hall *et al.* (1999) computed age among 20 two- and three-dimensional climate models and inferred age from *in situ* observations of selected trace gases in the stratosphere. The study found large differences in age between models and observations, and substantial spread across the models as well, allowing them to conclude that there were large biases in model transport schemes. More recently, Karpechko *et al.* (2013) found evidence that transport remains a problem in chemistry–climate model forecasts of ozone recovery in the Chemistry Climate Model Validation activities, phases 1 and 2 (Eyring *et al.*, 2005). Based on correlations with passive tracers, they concluded that the speed of ozone recovery in the models was very sensitive to the degree to which the stratospheric polar vortex remained isolated from the midlatitudes.

Age has also been used to investigate regional- and planetary-scale transport in idealised dry dynamical cores, similar to the models employed in this study. Orbe *et al.* (2012) computed the mean age of air using the transit-time distribution, a more detailed description of atmospheric transport pathways (Hall and Plumb, 1994), and Orbe *et al.* (2013) studied the effect of a changing climate on tropospheric transport through an analysis of air mass origins. Chen *et al.* (2017) used the age to study the role of monsoon-induced eddy circulation in inter-hemispheric transport of tracers in a HS94 dry atmospheric model and Waugh *et al.* (2019) even used age to estimate the transport time-scales in the Martian atmosphere.

3 | THE DYNAMICAL CORES

We compare four dynamical cores (hereafter, dycores) in this study, as summarized in Table 1. The cubed-sphere finite-volume (GFDL-FV3; Putman and Lin, 2007) and pseudospectral (GFDL-PS) dycores were developed at the

Geophysical Fluid Dynamics Laboratory (GFDL). The Community Atmosphere Model (CAM) finite-volume dycore (CAM-FV; Lin, 2004)) was developed at NASA/GFDL, and the CAM spectral element dycore (CAM-SE; Lauritzen *et al.*, 2018) was developed at the National Center for Atmospheric Research (NCAR). We briefly highlight the key features of the four dycores here. Details on their tracer advection schemes are included in Appendix A.

The GFDL-PS dycore is the oldest of the cores, and a direct descendent of the pseudospectral model analyzed in HS94. While it is no longer used by GFDL for climate prediction modeling, it is still used extensively in the idealized modelling community, for example serving as the dynamical core of the Isca modelling framework (Vallis *et al.*, 2018). It uses spherical harmonics in the horizontal coupled with a finite-difference discretization in the vertical. It uses the vorticity-divergence formulation of the equations and an explicitly imposed hyperdiffusion to represent subgrid-scale diffusivity.

The GFDL-FV3 dycore is GFDL's latest dynamical core, developed for weather and climate applications at GFDL and NASA Goddard Space Flight Center, and the National Centers for Environmental Prediction. It forms the dynamical core for the GFDL-AM4 atmospheric model (Zhao *et al.*, 2018a; 2018b) and also serves as the core of the fvGFS weather forecasting system. It employs finite-volume flux-limiting schemes in the horizontal coupled with a purely Lagrangian finite-volume transport scheme in the vertical. It uses a cubed-sphere grid to discretize the horizontal which implements a gnomonic projection of the sphere on to the six sides of a cube. Each face of the cube uses a local coordinate system native to the tile to separately solve the equations. Such a discretization provides a quasi-uniform Cartesian-like grid which helps circumvent the pole problem (Williamson, 2007) and allows for a higher CFL time step and provides higher scalability than the other grids. The solution from the six slices are then "stitched" together and interpolated onto an equi-spaced latitude–longitude grid for output using a fourth-order mass- and energy- conserving interpolation scheme (Putman and Lin 2007 give details).

The CAM-FV dycore uses the same class of numerical methods as the GFDL-FV3 dycore, using finite-volume schemes to integrate the equations. The two finite-volume-based models differ in that the CAM-FV uses a traditional latitude–longitude grid for horizontal discretization and an Eulerian finite-volume scheme in the vertical. CAM-FV is the principal core in NCAR's comprehensive climate models CESM1, Whole Atmosphere Community Climate Model (WACCM; Gettelman *et al.*, 2019), and CESM2.0 (Danabasoglu *et al.*, 2020).

The CAM-SE dycore is the newest addition to the dynamical core suite offered by the Community Atmosphere Model (Lauritzen *et al.*, 2018) and is being prepared to succeed CAM-FV as the principal dynamical core for future CMIP studies. It uses a continuous Galerkin spectral finite-element method to solve the primitive equations. It bears resemblance in formulation to both the GFDL-FV3 and GFDL-PS cores. It uses a cubed-sphere horizontal discretization and a vertical Lagrangian advection scheme like the GFDL-FV3 core, and hyperdiffusion as the sub-grid horizontal dissipation mechanism like the GFDL-PS core. CAM-SE differs from FV3 in applying a 4×4 spectral element within each of its cubed-sphere grid cells (Dennis *et al.*, 2012; Lauritzen *et al.*, 2018 provide details), but still interpolates the final output to a latitude–longitude grid.

To study how model resolution impacts transport, we compare integrations at four different resolutions. We first individually refine (double) the resolution in the horizontal and the vertical, respectively, and then we collectively refine the resolution in both the horizontal and the vertical. Table 1 summarizes the grid sizes and effective resolutions used. The model grid nomenclature (i.e., the meaning of T42 or NE16) is described in more detail in Appendix A. It is not trivial to compare the resolution of fundamentally different numerical schemes, but we have sought a fair comparison between the dycores. In particular, results from the NE16 and NE30 integrations of CAM-SE are interpolated onto default 256×129 (1.4°) and 512×257 (0.7°) latitude–longitude grid, which are finer than the effective model resolution, as detailed in Appendix A. For brevity, we will refer to all the lower and higher horizontal resolution integrations as 2° or 1° runs, respectively. Higher-resolution integrations with CAM-SE model are also considered in Section 7.

4 | TWO INTERCOMPARISON TESTS FOR THE DYNAMICAL CORES OF AGCMS

We propose two closely related tests to assess the impact of model numerics on circulation and transport in dynamical cores, as summarized in Table 2. We refer to them as the "free running" (FR) and "specified tropical winds" (SP) tests. They differ only in one aspect: the treatment of the tropical stratospheric winds (bottom row of Table 1). Our tests build on the the work of HS94, Polvani and Kushner (2002, hereafter PK02) and Orbe *et al.* (2012). The new contributions of this paper to these configurations are only in the bottom two rows, where we have added a clock tracer to the model and the ability to specify the tropical winds.

TABLE 1   A summary of the dynamical cores and their resolutions compared in this study

Numerics	Model grid	Effective resolution	Output grid
GFDL – pseudospectral (GFDL-PS)	T42L40	$2.8° \times 2.8°$	$128 \times 64 \times 40$
	T42L80		$128 \times 64 \times 80$
	T85L40	$1.4° \times 1.4°$	$256 \times 128 \times 40$
	T85L80		$256 \times 128 \times 80$
GFDL – cubed-sphere finite-volume (GFDL-FV3)	C48L40	$2° \times 2°$	$192 \times 96 \times 40$
	C48L80		$192 \times 96 \times 80$
	C90L40	$1° \times 1°$	$360 \times 96 \times 40$
	C90L80		$360 \times 180 \times 80$
CAM – spectral element (CAM-SE)	NE16np4L40	$2° \times 2°$	$256 \times 129 \times 40$
	NE16np4L80		$256 \times 129 \times 80$
	NE30np4L40	$1° \times 1°$	$512 \times 257 \times 40$
	NE30np4L80		$512 \times 257 \times 80$
CAM – finite-volume (CAM-FV)	F19L40	$2.5° \times 1.9°$	$144 \times 96 \times 40$
	F19L80		$144 \times 96 \times 80$
	F09L40	$1.25° \times 0.9°$	$288 \times 192 \times 40$
	F09L80		$288 \times 192 \times 80$

Note: Details on the "resolution nomenclature" (e.g., T42 versus C48) are provided in Appendix A. Identical 40 (80) vertical levels, as specified in Polvani and Kushner (2002), were chosen for all cores. For the GFDL-PS and CAM-SE, the output grid only reflects the grid to which the output was interpolated, and does not provide an accurate representation of the model's effective resolution. For the GFDL-PS core, the model resolution is more accurately represented by the number of resolved spherical harmonics (up to wavenumber 42 or 85). For GFDL-FV3 and CAM-SE, the resolution is determined by the effective number of points along each face of the cube. Thus, the NE16np4 (NE30np4) grid is comparable in resolution to the C48 (C90) grid.

TABLE 2   A summary of the model set-up and configuration for our free running (FR) and specified tropical wind (SP) dynamical core intercomparison tests

Forcing	Test 1: Free running (FR)	Test 2: Specified tropical winds (SP)
Diabatic forcing	Equations (1), (A1) and (A2) in Polvani and Kushner (2002)	
Planetary boundary layer	Rayleigh friction as given in Held and Suarez (1994), p. 1826, first equation in the box	
Topography	3 km amplitude sinusoidal wave 2 mountain given by Equation (1) in Gerber and Polvani (2009)	
Model top	Stratospheric sponge as prescribed in Polvani and Kushner (2002), para. [24]	
Tracer forcing	Linearly increasing clock tracer with source region $p \geq 700$ hPa, Equation (1)	
Tropical stratosphere	Undamped	Zonal winds damped as specified in Equation (B1)

Note: The two protocols are identical except for the treatment of the tropical stratosphere (final row).

4.1 | The free-running (FR) intercomparison test

The key to the HS94 test is to replace all diabatic forcings (radiation, convection, etc.) with a simple temperature tendency, where the temperature is relaxed linearly toward an analytic equilibrium profile, T_{eq}, a state approximating a radiative-convective equilibrium. The diabatic forcing in the troposphere is as prescribed in HS94, except for the addition of a north–south asymmetry to induce a perpetual January climatology, as detailed in PK02. The solstice climatology, as opposed to the equinox in HS94, was used because seasonality is critical for the stratosphere.

In order to drive a more realistic stratospheric circulation, we follow the PK02 modification of the HS94 equilibrium temperature profile above 100 hPa. This forcing imposes a fixed vertical lapse rate of -4 K·km^{-1} to T_{eq} in the boreal high latitudes to approximate radiative cooling

in the polar night. Away from the pole, the T_{eq} follows the US Standard Atmosphere (1976) temperature profile, with a positive lapse rate in the stratosphere. This forcing results in a cold pole in the winter (Northern) hemisphere, associated with strong cyclonic winds, or polar vortex. The summer (Southern) hemisphere is quiescent, with weak easterlies, as observed.

A smooth wave 2 topography of amplitude 3 km, as set up by Gerber and Polvani (2009; their equation 1) is added in the northern midlatitudes. The topography generates large-scale stationary waves in the troposphere, providing sufficient planetary wave forcing in the stratosphere. The planetary-scale waves strengthen the diabatic circulation in the winter stratosphere, due to enhanced wave-breaking, and drive intermittent breakdowns of the polar vortex, or Stratospheric Sudden Warmings (SSWs). The -4 K·km^{-1} lapse rate of T_{eq} in the winter stratosphere, paired with the 3 km topography, were found by Gerber and Polvani (2009) to produce the most realistic frequency of SSWs.

No other parametrizations are employed in the set-up except near the surface and the model top. For $p/p_{surf} \geq 0.7$, boundary-layer friction is approximated with a simple Rayleigh drag to damp the low-level winds. For $p \leq 0.5$ hPa, Rayleigh friction is added to damp the winds to zero as a crude parametrization of gravity wave drag. The sponge layer also absorbs any residual momentum due to upward propagating Rossby or large-amplitude gravity waves. The details of the stratospheric sponge layer are provided in paragraph [24] of PK02.

The age-of-air is computed by introducing a *clock tracer* near the surface (Hall and Plumb, 1994). While many studies have used the tropopause as the source region for computing the age, for instance Waugh and Hall (2002), we use a thick, near-surface layer as the source region for the clock tracer to avoid any ambiguity in defining the tropopause, or issues of resolution near the surface. In the source region, defined as $p \geq 700$ hPa to be consistent with the height of surface friction layer, the clock tracer concentration χ is specified to be $\chi(\lambda, \phi, p, t) = t$ for $p \geq 700$ hPa, where t is the model integration time, λ is the longitude and ϕ is the latitude. In dycores where only the tendency of the tracer can be specified, for example, GFDL-PS, we strongly damp χ to t with a time-scale of 1/10th of a day. Hence, the clock tracer value in the boundary region is forced to be equal to the present model time. Above 700 hPa, the clock tracer is passively advected with no sinks or sources, such that in total,

$$\frac{D\chi}{Dt} = \begin{cases} 0 & p < 700 \text{ hPa}, \\ 1 & p \geq 700 \text{ hPa}, \end{cases} \quad (1)$$
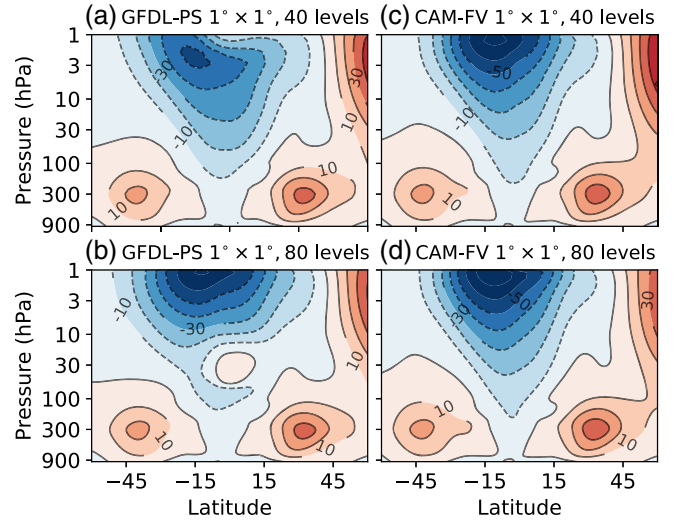

FIGURE 1 The climatological and zonal mean zonal wind $\bar{u}$ (m·s^{-1}), as simulated by two different dynamical cores, (a, b) pseudospectral (GFDL-PS) and (c, d) finite-volume (CAM-FV), with (a, c) 40 vertical levels and (b, d) 80 vertical levels. With higher vertical resolution, the pseudospectral core develops westerlies in the tropical stratosphere between 20 and 80 hPa, while the finite-volume core consistently simulates easterlies at both vertical resolutions. Both cores have comparable horizontal resolution. The contour intervals is 10 m·s^{-1}

where D/Dt is the material derivative following a parcel. As discussed in Section 2, the age is then simply

$$\Gamma(\lambda, \phi, p, t) = \chi(\lambda, \phi, p \geq 700 \text{ hPa}, t) -$$
$$\chi(\lambda, \phi, p, t) = t - \chi(\lambda, \phi, p, t). \quad (2)$$

4.2 | The specified tropical winds (SP) intercomparison tests

Initially, we had hoped that these idealized FR forcing could serve as our intercomparison test for tracer transport in the dynamical cores. However, we found that this proved too much of a test: the dynamical cores introduced in Section 3 failed to produce a comparable circulation, making it nearly impossible to assess the impact of numerics on transport. As shown in Figures 1 and 2, significant differences appeared in the tropical zonal wind structure, which resulted in differences in the diabatic circulation and isentropic mixing in the middle stratosphere, as discussed in detail in Section 6.

Figure 1 shows the behaviour of two different dynamical cores in response to changes in vertical resolution when allowed to evolve freely under the FR protocol. Figure 1(c, d) show the climatological zonal mean zonal winds in two integrations of CAM-FV. With this dynamical
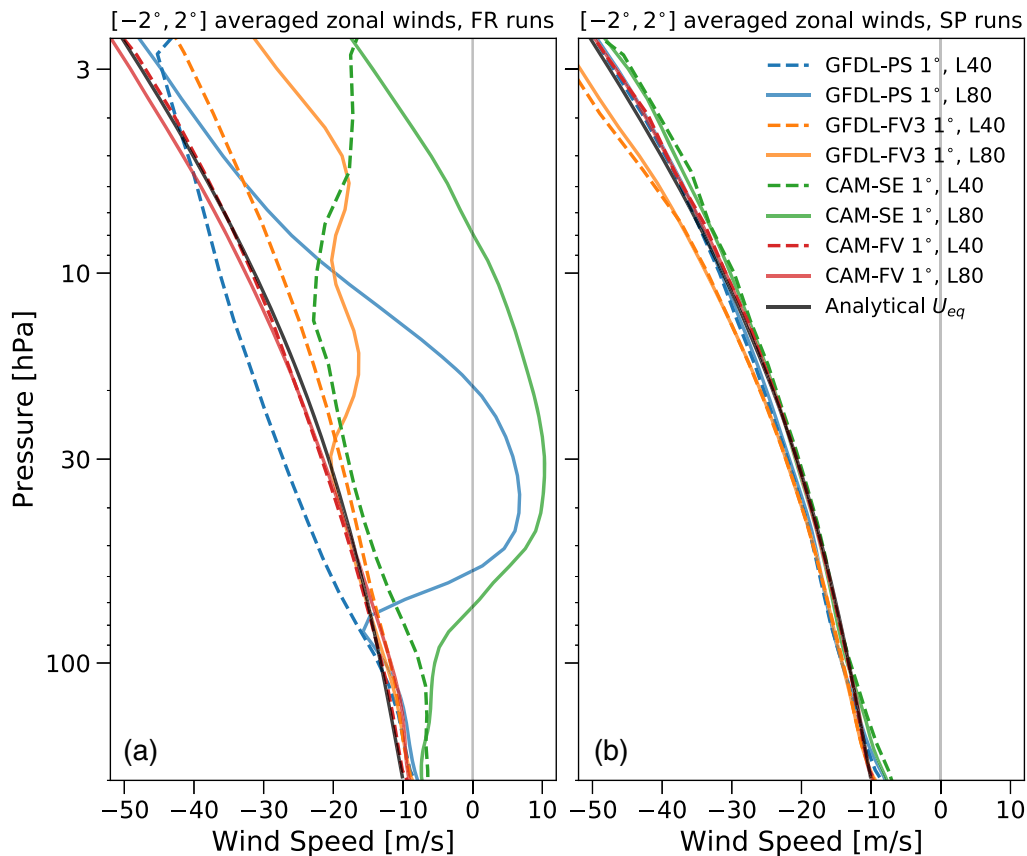
FIGURE 2 Climatological zonal wind $\bar{u}$ averaged over $[-2°,2°]$ for the four dynamical cores: GFDL-PS (blue), GFDL-FV3 (orange), CAM-SE (green) and CAM-FV (red), as obtained from the (a) FR integrations and (b) SP integrations. The solid curves and dashed curves show resolved winds at low (40) and high (80) vertical resolution, respectively. The black line shows the analytical wind profile specified in the Tropics in the SP experiments (Section 4 gives more details.)

core, the zonal mean winds appear insensitive to the vertical resolution. In contrast, integrations with the GFDL-PS dycore (Figure 1a, c) reveal a pronounced sensitivity of the climatological winds to changes in vertical resolution. While the differences appear only in the tropical stratosphere, the winds here have a pronounced impact on transport throughout the stratosphere. As shown in Section 5, the age increases by as much as 25% over much of the lower and mid-stratosphere in response to the formation of westerly jets in the tropical stratosphere.

While the CAM-FV model appears more "converged", that is, insensitive to changes in the resolution, we do not believe that it is necessarily providing a more accurate representation of the circulation. The angular momentum balance in the tropical stratosphere is very sensitive, with an adjustment time-scale on the order of years near the Equator (Holton *et al.*, 1995; Lauritzen *et al.*, 2011). In Earth's atmosphere, gravity waves drive an oscillation between westerly and easterly jets over a period of approximately 28 months, the QBO. Comprehensive models must be run at high vertical resolution with a parametrization of gravity wave momentum deposition to capture this variability (Giorgetta *et al.*, 2002).

Figure 2a shows that this sensitivity of the tropical winds to model numerics and resolution extends to all four of the dynamical cores. All of the models exhibit an easterly regime at moderate vertical resolution (dashed lines), but the two dynamical cores with pseudospectral and spectral element numerics shift to a westerly regime at higher vertical resolution (solid lines). While the CAM-FV model shown in Figure 1 appears remarkably insensitive to vertical resolution, there is evidence of a slight westerly shift at upper levels when the finite-volume model developed by GFDL is run at higher vertical resolution.

To allow a meaningful comparison of the numerical transport properties of different dynamical cores, we remove this source of uncertainty for the analysis by weakly damping the tropical stratospheric winds to a steady easterly profile in the specified tropical winds (SP) test. The damping results in very similar tropical winds across all cores (at least to 1 hPa; issues remain above due to differences in dissipation at the model top) allowing a more consistent comparison of their transport properties. While we would have preferred an unconstrained test, as in the FR experiments, there is precedent for specifying the QBO winds in comprehensive atmospheric models (e.g.,

Matthes *et al.*, 2010). Except for this damping, the FR and SP tests are identical.

Damping of tropical stratospheric winds begins at 200 hPa and is strongest at the Equator, gradually tapering off to zero at 15°N/S. An analytical expression and detailed description of the damping is provided in Appendix B, and the analytical wind profile is shown in Figure 2 (black lines). The easterly profile was chosen as a neutral background state, similar to that produced by all models at lower vertical resolution. The damping is applied to each grid cell independently, such that waves are also damped; damping only the zonal mean winds would have minimized the impact on resolved waves, but this is not trivial with cubed-sphere grids. The damping time-scale, 40 days in the lower stratosphere, rising to 10 days above 3 hPa (Equation (B5)), is long compared to the Rossby waves, which break on synoptic time-scales of a few days. Figure 2b shows that the damping brings all models in to the same tropical wind regime.

We stress that we only weakly damp the tropical stratospheric winds to a prescribed wind profile, otherwise allowing the model to freely determine the circulation and transport. This differs considerably from cases where stratospheric winds are strongly nudged towards a specified structure throughout the stratosphere. For instance, Froidevaux *et al.* (2019) used the FR and the specified dynamics (SD) versions of the CESM1 Whole Atmosphere Community Climate Model (i.e. FR-WACCM and SD-WACCM) to evaluate model transport and to study the climatological distribution of stratospheric trace gases (including ozone, water vapour and nitrous oxide) over the 2005–2014 period. Similarly, Chrysanthou *et al.* (2019) used nudging to investigate the spread in residual circulation in an ensemble of Chemistry Climate Model Initiative (CCMI) simulations, with winds nudged towards a reanalysis dataset over the 1980–2009 period. Ongoing research suggests that specifying the dynamics does not always improve transport as the model numerics are no longer consistent with the circulation (Linz 2020, personal communication).

4.3 | Experimental Details

The stratospheric circulation and transport evolves on time-scales of months to years, such that a long integration time is needed. Therefore, to test a dynamical core, we recommend an integration of 10,000 days (~27 years). While the winds reach a statistical steady state relatively quickly (within a few hundred days), it takes about 6,000 days for the clock tracer distribution (and hence the age) to converge to a statistically steady state at upper levels. We use the last 8000 days of the integration to analyze dynamical quantities (e.g., the zonal winds and temperature variance), and the last 3300 days to compute the age. These long analysis periods allow us to average out internal variability, which is particular large in the winter polar vortex.

5 | RESULTS: CIRCULATION AND TRANSPORT BENCHMARKS

We now compare the circulation and age-of-air distributions that result from our new intercomparison tests, considering the effects of model numerics and resolution. Circulation and age-of-air benchmarks for dynamical cores are provided, which can be used to assess newly developed dynamical cores and alternative tracer advection schemes.

5.1 | Circulation benchmarks

We begin with the metrics proposed in HS94 to assess the general circulation, as shown in Figure 3. Only results based on the highest resolution runs (1° L80 runs: T85L80, C90L80, NE30L80) are shown, with the exception of CAM-FV, where the F09L40 integrations were used. This exception was required due to anomalous behaviour in the L80 layer integrations with CAM-FV, which we believe is associated with an issue at the model top, in the Rayleigh sponge layer where the circulation should be strongly damped. Inclusion of F09L80 statistics would greatly enhance the uncertainty of the ensemble result. The figure shows ensemble means of the zonal mean zonal winds, eddy temperature variance, and the zonal spectra of eddy kinetic energy. The same metrics at select pressure levels are shown in Figure 4, which allows a comparison between the individual models. In all line plots throughout the section, the dashed curves represent the FR integrations, and the solid curves represent the runs with SP tropical winds.

The contours in Figure 3a show the ensemble mean of the zonal mean wind patterns from FR runs, and the red shading shows the standard deviation across the four run ensemble. Outside the tropical stratosphere – differences which we highlighted already in Section 4 – all integrations capture nearly the same large-scale jet structure. The intermodel spread is largest in the extratropical stratosphere where the strength of the winter polar vortex (the westerly jet with peak wind speeds $\approx 50\,\mathrm{m\cdot s^{-1}}$ near 70°N and 3 hPa) varies between models. This can be seen more clearly in Figure 4b. There are also subtle differences in the extratropical troposphere, namely in the position of the midlatitude westerlies, which barely appear at the contour interval in Figure 3, but can be seen more clearly
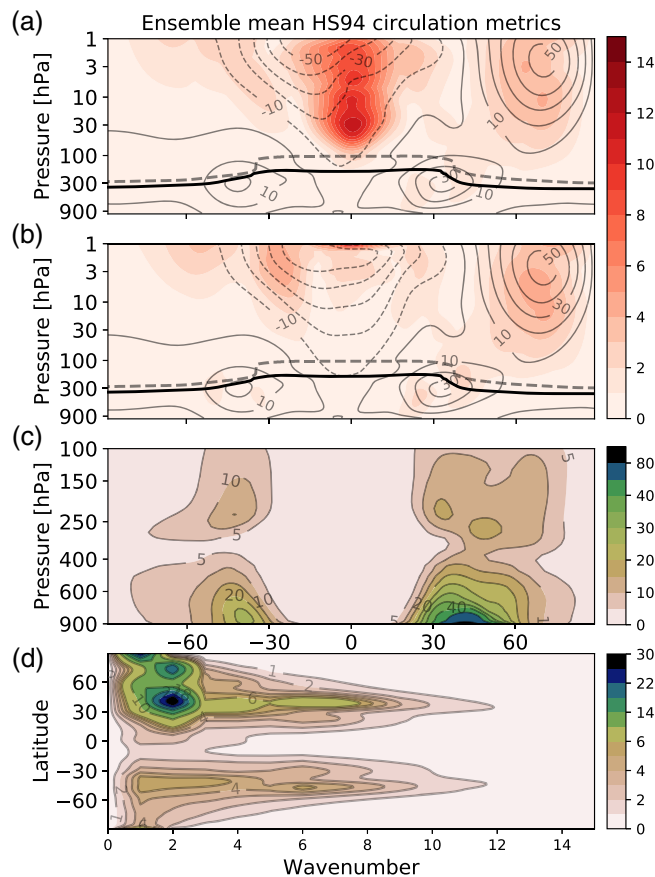
FIGURE 3 Ensemble mean of the zonal mean zonal wind for the (a) free running and (b) specified tropical wind integrations. The grey contours show the ensemble mean computed from the four 1° runs (T85L80, C90L80, NE30L80, F09L40) and the colours show the standard deviation of the winds across the ensemble. (c) The ensemble mean eddy temperature variance (K^2) and (d) the ensemble mean of the vertically integrated zonal spectra of eddy kinetic energy, both for the SP experiments. The solid and dashed bold lines in (a) and (b) show the thermal tropopause inferred from the models (ensemble mean) and ERA5 data (1990–2019 DJF), respectively

FIGURE 4 (a, b) Climatological zonal mean zonal wind and (c, d) eddy temperature variance, as shown in Figure 3(a, b), but at two different vertical levels in the 1° integrations of the four dycores. (a, c) compare the two quantities at 300 hPa, the level at which the tropospheric jet streams reach peak amplitude, and (b, d) compare them at 30 hPa, chosen to characterize behaviour in the stratosphere. The dashed and solid curves show the quantities for the free-running (FR) and the specified tropical winds (SP) integrations, respectively. The GFDL-PS core is shown in blue, GFDL-FV3 core in orange, CAM-SE in green, and CAM-FV in red

in Figure 4a. The subtropical and extratropical jets are merged in this climatology, more so than in the original HS94 test, but there are hints of greater separation between them in the finite-volume-based cores, particularly in the summer hemisphere (Figure 4a around 20°S).

In the ensemble mean of the FR experiments, the tropical stratosphere exhibits easterly winds, but the variance is large. Figure 4b shows the split between the spectral-based cores, which develop westerly jets, and the finite-volume cores, which maintain easterlies. This split in behaviour is eliminated in the SP integrations shown in Figure 3b. Variance between the tropical stratospheric winds in this test is mostly removed by construction, although there are still differences at upper levels, particularly near the model top (not shown).
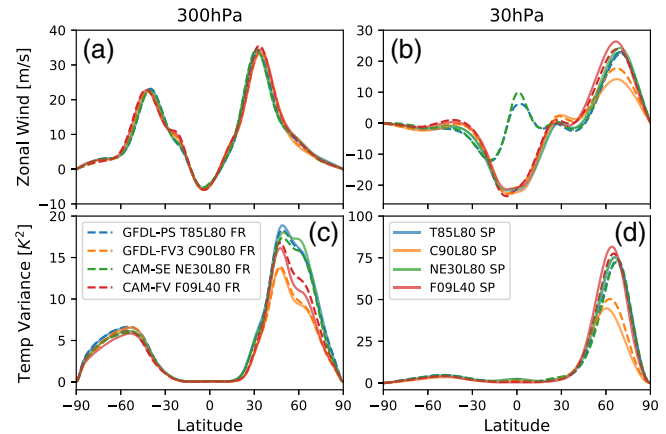
Outside the tropical stratosphere, the winds in the SP and FR integrations are pretty comparable, as is the spread between models. However, there is more spread in the strength of the stratospheric polar vortex in the SP integrations. In GFDL-FV3, specifying the tropical winds leads to a weaker vortex, while in CAM-FV, the vortex increases in strength, as seen in Figure 4b. This results in a slightly larger variance of $\bar{u}$ in the vortex region for the SP experiment than for the FR experiment.

Figure 3c shows the ensemble mean of the eddy temperature variance in the troposphere for the SP runs; the results are nearly identical in the FR integrations and so not shown. Our choice of a perpetual northern winter climatology endows a hemispherical asymmetry in the eddy temperature variance structure not observed in the HS94 test, the variance being much larger in the winter hemisphere. The variance is maximum near the surface and decreases higher up in the troposphere up to roughly 400 hPa. Above that, variance in the upper troposphere slightly increases, associated with the midlatitude jet variability and baroclinic eddy heat exchange between stratosphere and troposphere.

Figure 4c, d shows the temperature variance in the individual models for select levels. At 300 hPa, which captures the upper-troposphere maximum, variability is weaker in both of the finite-volume-based cores relative to the spectral-based cores. This suggests that finite-volume cores are more strongly damped; although dominated by synoptic scales, this variability increases with resolution

(not shown). Figure 3c was truncated at 100 hPa, as the variability in the polar vortex is very large, exceeding 150 K^2 at 10 hPa. Differences in the variance at 30 hPa, shown in Figure 4d, are characteristic of differences at other levels. There is notably less temperature variability of the vortex in the GFDL-FV3 integrations.

Figure 3d shows the vertically integrated zonal spectra for the eddy kinetic energy (EKE) for the SP runs. The concentration of EKE at wavenumber 2 is associated with stationary waves induced by the topography, and the asymmetry between the hemispheres (not present in HS94) is due to the wintertime climatology added by Polvani and Kushner (2002). We do not show this metric for our FR runs, as the vertically integrated spectrum for the ensemble is nearly identical for both the FR and SP tests.

5.2 | Coupled stratosphere–troposphere variability benchmark: SSW events

The wave-2 topography was developed by Gerber and Polvani (2009) to increase the planetary-scale wave forcing on the stratosphere, and to obtain realistic frequency of SSW events. A SSW is identified by an abrupt reversal of the stratospheric polar vortex, specifically the zonal mean zonal winds at 60°N and 10 hPa, following World Meteorological Organization (WMO) convention (Mcinturff, 1978). Figure 5a illustrates a time series of the polar vortex winds in CAM-SE, a thousand-day period during which there were three SSWs (marked by vertical dashed lines), in addition to a minor warming at the start of the period, when the vortex weakened, but did not completely reverse. The breakdown of the vortex is associated with a dramatic warming of the polar stratosphere, which occurs approximately every other year in the boreal hemisphere (e.g., Butler *et al.*, 2017). SSWs are a source of surface weather predictability on subseasonal to seasonal time-scales, as a poleward shift of the tropospheric jets follows the breakdown of the vortex (e.g., Kidston *et al.*, 2015).

Climate prediction models have struggled to properly capture the correct frequency of SSWs (Charlton-Perez *et al.*, 2013), although more recent climate models show improvement (Ayarzagüena *et al.*, 2020). In general, reasonably fine vertical resolution of the stratosphere and a model lid above the stratopause is needed (albeit not sufficient) to simulate them. However, Figure 6, which shows the SSW frequency among the four dynamical cores considered in the study, suggests that the choice of model numerics may also play an important role. Overall, most of the dynamical cores simulate between two and three SSWs per thousand days of integration. In observations, SSWs are observed every other year, but recall that we
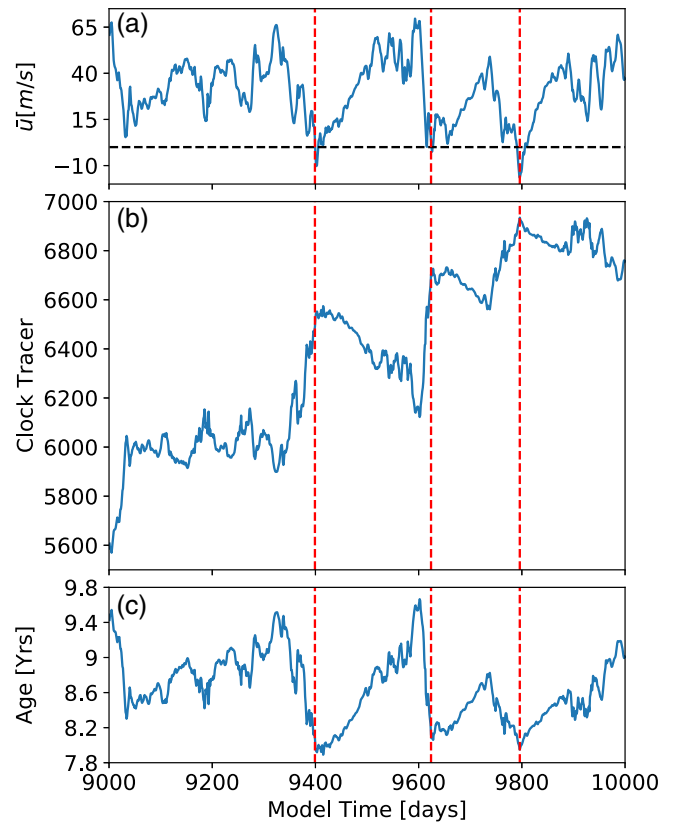

FIGURE 5 (a) Zonal mean zonal wind, (b) clock tracer concentration, and (c) age-of-air at 60°N and 10 hPa for a 1,000-day interval for the CAM-SE NE30L80 integration. Dashed red lines mark reversals of the zonal mean zonal wind from westerly to easterly, indicating major Sudden Stratospheric Warming events

simulate a perpetual winter state. While the GFDL-PS and CAM-SE models' SSW statistics appear insensitive to changes in horizontal or vertical resolution, the two finite-volume-based cores appear sensitive to horizontal resolution.

The CAM-FV core stands out at low (2°) horizontal resolution, simulating very few events (if any at all). However, when run at 1° resolution, its statistics become consistent with the other models (with the exception of the L80 run with specified tropical winds, which appeared anomalous in many respects). In contrast, the GFDL-FV3 model appears to simulate only half as many warmings as the other dynamical cores when run at higher horizontal resolution. The fact that SSWs are rare events makes it difficult to establish the statistical significance of this difference. Even with 8,000 days of integration time, the 2-sigma error bounds (not shown) still overlap.

Since 8,000 days is insufficient to assess differences in the dynamical cores' ability to represent SSWs, much longer test integrations are required to establish significant differences between the models. Limitations on available computational resources precluded us from further
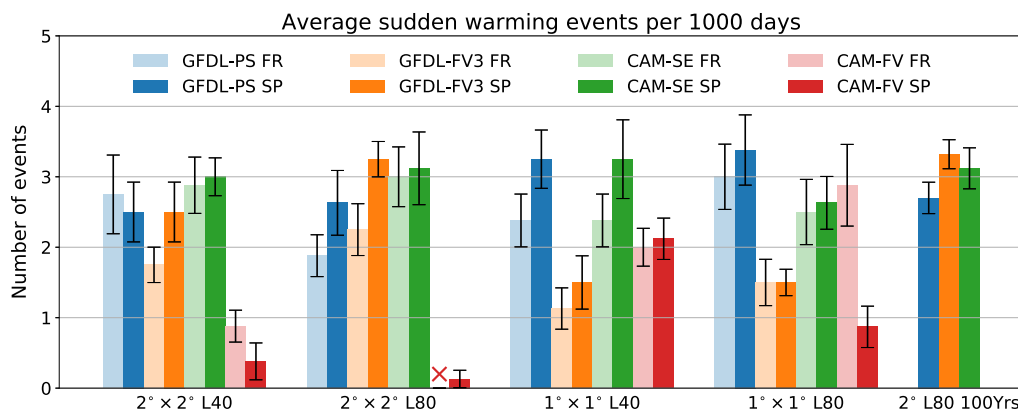
FIGURE 6 Average sudden warming frequencies per 1000 interval for the dynamical cores grouped by horizontal and vertical resolution. The histograms represent averages taken over eight independent samples and the error bars indicate the ensemble standard deviation of the mean frequency. For each resolution group, the lighter coloured bars represent the free-running (FR) integrations and the darker coloured bars represent the specified (SP) runs. The red cross in the second group indicates the absence of any sudden warming events for the F19L80 FR integration. For great statistical confidence, three of the four cores were integrated at 2° L80 resolution for 100 years: the statistics of these long integrations are shown in the rightmost group

investigating the SSW statistics of the 1° resolution integrations. However, we did probe the 2° L80 runs in more detail, running three of the models for 100 years each. This provided a factor of 4 increase in sampling, and hence narrowed the uncertainty of the mean frequency by a factor of 2, as shown at the far right of Figure 6. Based on these results, we believe that 100 years of integration time would be sufficient to clarify whether GFDL-FV3 behaves significantly different at 1° resolution.

It might come as a surprise that the SSW frequency is more sensitive to horizontal resolution than to vertical resolution; this is due to the fact that even the 40-level integrations exhibit a well-resolved stratosphere compared to a typical CMIP model (table 1 in Anav *et al.*, 2013, for instance). We compare the vertical resolution of our integrations with those associated with full AGCM integrations of the GFDL and CESM models in Figure 7. Not only is 40 layers fairly high vertical resolution, the simplicity of the HS94 forcings does not require fine resolution of the surface boundary layer, allowing us to place 17 levels between 100 and 1 hPa. The dependence on horizontal resolution is not easy to explain. The 2° CAM-FV model simulates an anomalously strong stratospheric polar vortex, which may inhibit planetary wave propagation. But for the 1° GFDL-FV3 integrations, the vortex is actually anomalously weak, as seen in Figure 4b. A weaker vortex is usually associated with greater wave forcing on average (and a correspondingly larger meridional overturning circulation).

When the tropical winds are specified, a slight (albeit consistent) increase in the average SSW frequency is observed for all cores at all resolutions (again with an exception of the CAM-FV core). This increase can

be explained by the (Holton and Tan, 1980) effect. Tropical easterlies inhibit the equatorial propagation of planetary-scale Rossby waves, trapping them in the high latitudes, leading to a more disturbed polar vortex and hence more SSWs. In contrast, a westerly jet in the Tropics shifts the critical lines for planetary waves equatorward, allowing them to percolate into the Tropics (and even the summer hemisphere). This reduces the planetary wave activity in the winter midlatitudes, resulting in fewer SSW events (Shuckburgh *et al.*, 2001; Garfinkel *et al.*, 2012).

The polar stratospheric vortex is associated with sharp meridional gradients of potential vorticity, which inhibit exchange of air between the polar and midlatitude stratosphere, allowing very old air to accumulate in the polar vortex (McIntyre and Palmer, 1983; Sobel *et al.*, 1997). The breakdown of the polar vortex during an SSW destroys this barrier to mixing, homogenizing air between the midlatitude surf zone and the pole. A signature of this can be observed in the age of air, as illustrated in Figure 5b, c. Overall, the concentration of the clock tracer averaged over the polar cap (Figure 5b) grows linearly in time, indicating that the age of air in the polar vortex has converged (Figure 5c). This growth, however, is highly episodic, accomplished almost exclusively during SSWs, when young air (with a clock tracer concentration closer to the surface value) is mixed into the vortex. The tracer concentration spikes by several hundred "days" over the course of a month, associated with a drop in the age of air of over a year (Figure 5c).

After a warming event, the polar vortex re-establishes itself, isolating the polar air again. As a result, the age of air begins to increase again, as is noticeable between days 9,400 and 9,600, as the age increases from 8 years to
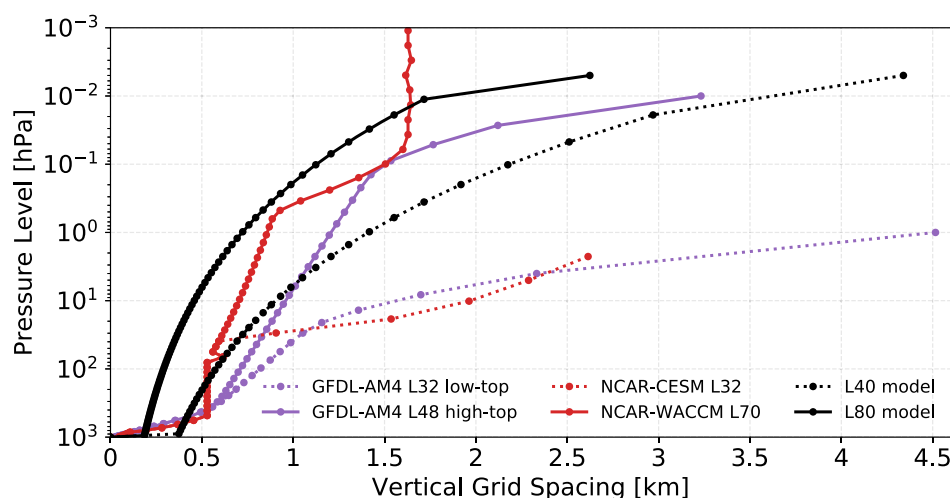
FIGURE 7 Approximate vertical grid spacing as a function of pressure (in km assuming log pressure with a fixed 7.5 km scale height), used in the dynamical cores for the 40- (dashed black) and 80- (solid black) level configurations. The vertical spacing is compared with the coordinates used by the low-top L32 (dashed violet) and high-top L48 (solid violet) AM4 AGCM associated with the GFDL-FV3 dycore, and the low-top L32 CESM model (dashed red) and the high-top L70 WACCM model (solid red) associated with the CAM dycores

9.6 years within just 200 days, associated with slow descent of very old air from above. It is these large variations in the age-of-air in the polar vortex associated with SSWs that required a long averaging period (3000 days) to obtain a reliable estimate of the mean age. For most other regions, the age-of-air can be reliably estimated with a much shorter averaging window.

5.3 | Transport benchmark: Age-of-air

Figure 8 shows the ensemble mean age-of-air in the stratosphere, taken across the same high-resolution runs as in Figure 3. Figure 8a shows results based on the FR test and the right, the SP test; in both, solid contours show the age-of-air and the colour shading, the fractional deviation in the age across the ensemble mean. The most conspicuous difference between the two tests is the fractional age deviation: the FR integrations vary by a factor of up to 25% in the tropical stratosphere and winter midlatitudes, corresponding to a spread of up to 2 years in the age of air. The deviation for the SP test are substantially lower throughout the stratosphere (just 6% in the Tropics). A closer look also reveals significant differences in mean age between the two ensembles. For instance, the ensemble mean age at 45°N and 10 hPa is over 10% higher in the FR runs than in the SP integrations.

Before investigating these differences, we first highlight the large-scale structure of the transport profiles. The overall structure of the mean age in the models is consistent with observation-based estimates of the mean age of the atmosphere, but with a substantial bias towards older values. The age in the upper stratosphere exceeds 9 years in the models, compared to 6 years in observations (Waugh and Hall, 2002, their figure 7). The offset is due to a multitude of factors. First, Waugh and Hall (2002)

define the age relative to the tropopause, as opposed to the surface boundary layer ($p \geq 700$ hPa in our idealized configuration). However, the air at the tropical tropopause in our models is approximately 0.5–0.75 years old, explaining only part of the offset. The remaining difference is due to the fact that the tropopause is artificially low in the PK02 stratosphere, as shown by the solid and dashed black curves in Figure 3a, b. As the vertical ascent of air is much slower in the stratosphere than in the troposphere, increasing the mass of the stratosphere increases the age at any given pressure level: air making it to 100 hPa has already spent upwards of 1.5 years in the dycore models' stratosphere. Finally, the overturning circulation strength in the model is weak compared to that of the real atmosphere. Gravity wave driving plays an important role in the meridional overturning of the stratosphere, particularly at upper levels, but is only crudely approximated in these integrations by a sponge layer above 0.5 hPa. However, the older air helps emphasize differences in numerical schemes, as small errors accumulate over time.

In the vertical, the stratosphere always becomes older with height, even in the midlatitudes where the mean transport of mass is downward. While the mean overturning tends to bring air upward in the Tropics and downward in the mid and high latitudes, isentropic mixing between the upward and downward branches of the cells ensures that the oldest air is found aloft (Neu and Plumb, 1999). The horizontal structure of the age distribution, which can be more clearly seen in Figure 9, reveals three distinct regions: (a) the tropical pipe, where comparatively young air is brought up into the stratosphere, (b) the midlatitude surf zone, where isentropic mixing homogenizes the age, seen most clearly at 30 hPa in Figure 9b, and (c) the polar vortex, where on average the oldest air is found. The three regions are separated by sharp gradients in age, associated with barriers to mixing. The tropical barrier, ~15–20°N is

FIGURE 8 Ensemble mean age-of-air (years) in the stratosphere (solid contours) for (a) the free-running FR integrations and (b) the specified tropical wind SP integrations. The colours show the fractional age deviation (%) across the ensemble mean. The contour interval is 1 year, except for an additional contour at 0.5 to highlight the structure in the lower tropical stratosphere
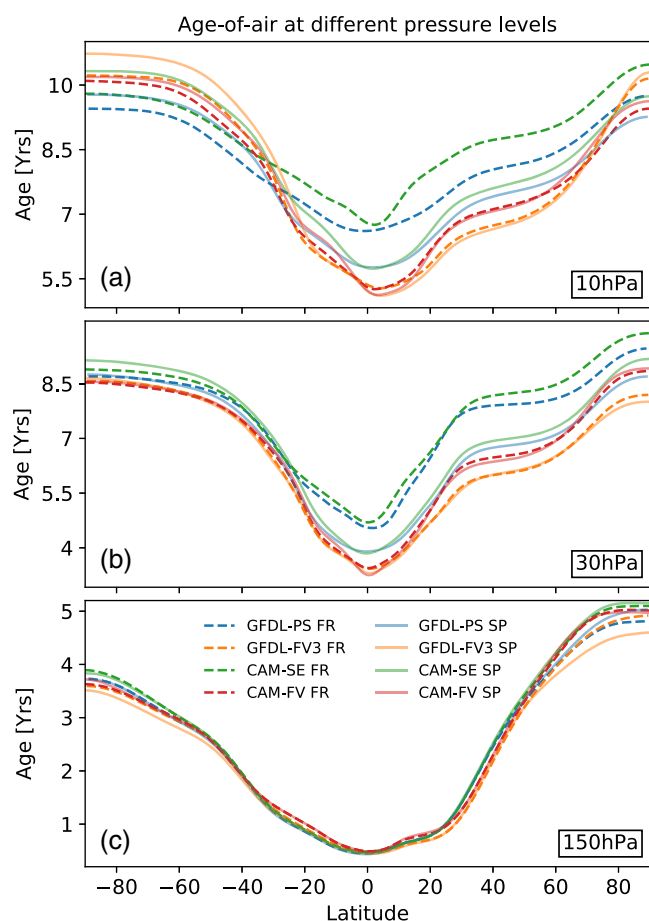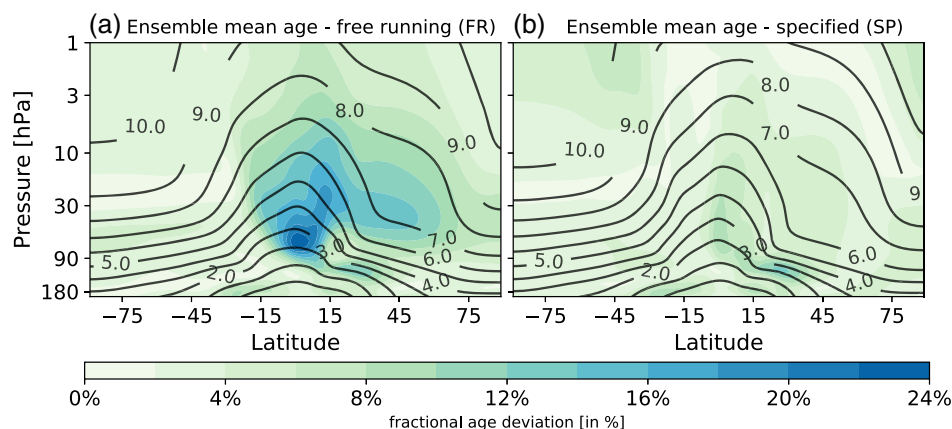

FIGURE 9 Age-of-air (years) for the four dycores at three different pressure levels: (a) 10 hPa, (b) 30 hPa, and (c) 150 hPa. FR and SP integrations are shown using dashed and solid curves respectively. The three levels are chosen to represent the upper, middle and lower stratosphere respectively

associated with the tropical limit of planetary-scale wave breaking, while the polar barrier at 60°N is associated with strong potential vorticity gradients surrounding the strong polar vortex (Plumb, 2002).

We have focused on the Northern (winter) Hemisphere, as perpetual summer in the Southern Hemisphere prevents the formation of a surf zone and a polar vortex. Easterly flow in the summer hemisphere inhibits Rossby wave propagation, limiting both isentropic mixing and the meridional overturning circulation.

Returning to the difference between FR and SP results, the key factor contributing to the large age variance in the FR integrations is a split in the vertical structure of age in the spectral-based models versus the finite-volume-based models. This is immediately evident in Figure 9, which shows the age-of-air among different models at three different pressure levels in the stratosphere, chosen to represent the lower, middle and upper stratosphere. At 150 hPa, the age is very similar among all models, independent of the model numerics (colours) or treatment of tropical winds (dashed versus solid). The strong gradient between the Tropics and higher latitudes reflects the contrast between fresh air brought up into the stratosphere in the Tropics and the comparatively old air returning down to the troposphere at higher latitudes. The gross age gradient is a function of the diabatic circulation strength alone, independent of isentropic mixing (Neu and Plumb, 1999; Linz *et al.*, 2016), and reflects the consistency of the meridional overturning across the models.

However, higher up, at 30 hPa and 10 hPa, a key difference between the dynamical cores emerges, especially in the FR integrations. In the FR integrations, the spectral-based cores (GFDL-PS and CAM-SE) exhibit an older age-of-air all the way from the subtropics of the summer hemisphere to the winter pole (dashed blue and green). The air in the midlatitudes is as much as 2.5 years older (i.e., around 25% older, at lower levels) when compared to the finite-volume-based cores (dashed red and orange). The spread between the finite-volume and spectral models grows slightly with height, but less so than the mean age, so that the relative (fractional) variance decreases with height (Figure 8).

Specifying the tropical winds has little effect on the finite-volume models, but leads to a significant decrease in the age in the spectral models. The age in the winter

midlatitudes drops by as much as 2 years. This explains both the drop in mean age and the dramatic drop in the spread between the FR and SP ensembles. The consistency between the two models with finite-volume numerics and the two models using spectral methods is striking. Perhaps this should not come as a surprise for the finite-volume models, which differ primarily in their grid (lat–lon versus cubed sphere), although they have different treatment of vertical advection. The two "spectral" models, however, are rather different, particularly in the treatment of tracer transport; GFDL-PS is a very old core based on decomposition of the flow into the spherical harmonics, while CAM-SE is a very modern core based on a cubed-sphere grid and uses spectral elements for horizontal advection of both momentum and tracer. We therefore speculate that differences between the spectral and finite-volume models stem from their representation of the resolved dynamics, where the use of explicit hyperdiffusion makes them less diffusive on intermediate scales.

The region with peak ensemble age variance in the FR experiment coincides with spread in winds (Figure 3a). As seen in Figure 2, the winds abruptly shift toward a westerly jet at about 80 hPa in the two spectral models. The age in the Tropics diverges between the models at this level, with the spectral models remaining consistently older at all levels above. However, the difference in age spreads all the way to the pole, even though the winds are similar outside the Tropics. As seen in Figure 9, the difference between the finite-volume and spectral models increases a bit with height, but to first approximation, the difference could be attributed to a uniform (in latitude) jump in age at 80 hPa, which remains roughly constant at all levels above. This abrupt change in the vertical gradient of age suggests a difference in mixing, which we investigate in Section 6.

Even when the the tropical winds are specified, there is a split between the age in the spectral and finite-volume models. The spectral models consistently show older age throughout the winter stratosphere, by as much as 0.75 to 1 years in the winter Extratropics. Moreover, the reduction in the ensemble variance is only seen in the winter hemisphere and summer hemisphere Tropics, as the age-of-air in the extratropical summer hemisphere remains largely unchanged. These differences between the age in the spectral and finite-volume models, and their dependence on resolution, are investigated in more detail in Section 7.

6 | SENSITIVITY TO VERTICAL RESOLUTION

Transport in the stratosphere is affected by both the Lagrangian mean overturning of mass and by the

mixing of tracers along isentropes (layers of constant entropy, or potential temperature). An increase in the meridional overturning will enhance the rate at which the stratospheric trace gases are flushed out of the region, thus reducing the age-of-air, all other things being equal. As discussed by Linz *et al.* (2016), the meridional overturning can be directly linked to the horizontal gradient of age. An increase in the isentropic mixing, on the other hand, results in an overall increase in the age in both the Tropics and Extratropics, as seen in the leaky pipe model (Neu and Plumb, 1999).

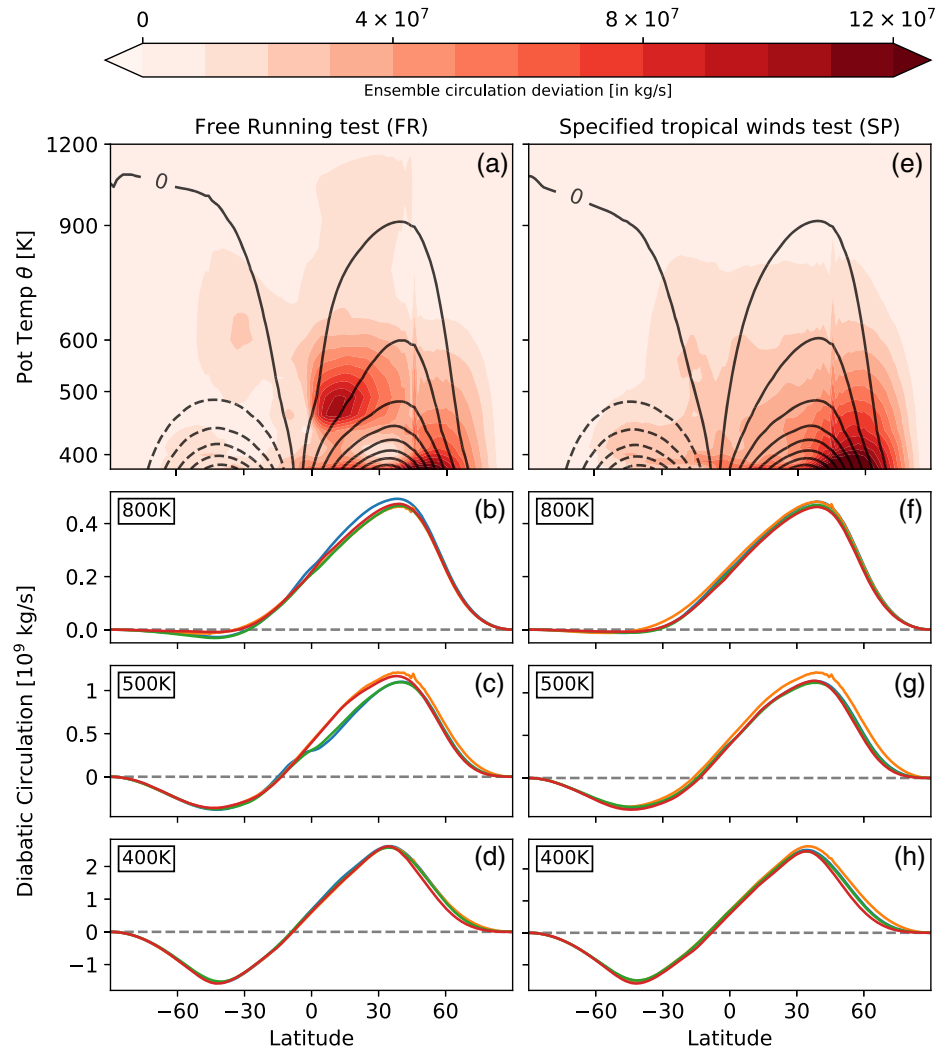## 6.1 | Differences in the diabatic circulation

To quantify the Lagrangian overturning circulation of mass in the stratosphere, we use the isentropic mass streamfunction. The Eulerian mean mass streamfunction averaged in isentropic coordinates provides a quantitative estimate of the overturning comparable to other measures, for example, the Transformed Eulerian Mean, or residual circulation (Edmon *et al.*, 1980; Townsend and Johnson, 1985; Linz *et al.*, 2019). As in Pauluis *et al.* (2009), the time-averaged Eulerian streamfunction in isentropic coordinates can be expressed as an integral of the meridional mass flux,

$$\psi(\phi, \theta_0) = A \int_{T_1}^{T_2} \int_0^{2\pi} \int_0^{p_s} v H[\theta(\lambda, \phi, p, t) - \theta_0] \mathrm{d}p \ \mathrm{d}\lambda \ \mathrm{d}t, \quad (3)$$

where ψ is the mass streamfunction averaged over a given time interval $t \in [T_1, T_2]$, $A = R\cos\phi/g(T_2 - T_1)$, v is the meridional velocity, H is the Heaviside function which by construction equals 1 for non-negative values and 0 otherwise, $\lambda \in [0, 2\pi]$ is the longitude, $\phi \in [-\pi/2, \pi/2]$ is the latitude, $\theta(\lambda, \phi, p, t)$ is the potential temperature at a given position, R is the radius of the earth and g is the acceleration due to gravity. In our analysis, the time-averaged diabatic circulation strength is obtained by averaging the streamfunction over the last 8,400 model days.

Figure 10a shows the ensemble mean diabatic circulation (in black) among the highest-resolution FR runs (with the exception of CAM-FV on the 40-level grid as before). The ensemble mean diabatic circulation strength is shown using solid black contours and the colours show the standard deviation of the circulation strength across the ensemble. The circulation qualitatively matches the observed mean meridional stratospheric overturning circulation. In particular, it shows that, in a mean Lagrangian sense, air moves upwards in the Tropics and downwards in the higher latitudes, gradually moving polewards in between. A perpetual winter results in a much stronger

FIGURE 10 Ensemble-averaged diabatic circulation in the stratosphere (solid black contours) and ensemble deviation in circulation (colour shading) in potential temperature coordinates for (a) FR integrations and (e) SP integrations. (b)–(d) show the diabatic circulation strength from the FR integrations on three isentropes: (b) 800 K, (c) 500 K, and (d) 400 K. (f)–(h) shows the same but for the SP integrations. As before, the most resolved runs were considered for the ensemble averaging with the exception of CAM-FV. As in previous figures, GFDL-PS is shown in blue, GFDL-FV3 in orange, CAM-SE in green, and CAM-FV in red. Contour interval for the diabatic circulation is 0.4×10^9 kg·s^{-1}. The small spikes in the GFDL-FV3 curves are an artifact of the mass-conserving interpolation used to stitch the tiles together near 45°. The spikes are more prominent at lower horizontal resolution (2°) and decrease as the horizontal grid is further refined

mass transport in the winter hemisphere, and higher up in the stratosphere the tropical diabatic upwelling intrudes all the way into the summer hemisphere.

Similar to the results shown for the ensemble mean zonal winds and age-of-air (Figures 3 and 8, respectively), the FR models exhibit large variations in the diabatic circulation strength in the Tropics (red shading, Figure 10). Between 450 and 600 K, the circulation strength varies by as much as 15%. In contrast to the zonal wind variance, these differences pervade all the way into the winter midlatitudes. Figures 10b–d show the variations among different cores at three different levels in the stratosphere. The models tend to agree on the circulation structure and strength at all levels, except at 500 K, where they disagree in the Tropics and winter midlatitudes. The GFDL-PS and the CAM-SE cores, which resolve tropical westerlies, tend to have a much weaker diabatic circulation in the region than the two finite-volume cores. This difference results in the high variance obtained in Figure 10a.

These variations in mass transport are eliminated once the tropical winds are specified, as seen in Figure 10e

which shows the same metrics as Figure 10a, but for the SP integrations. The models no longer disagree on the diabatic circulation strength in the Tropics. A closer look at three different isentropic levels in the stratosphere (Figures 10f–h) confirms this observation: all the cores agree on the circulation strength in the SP integrations. This suggests that differences in mass transport were due to differences in the resolved tropical winds. Tropical westerlies in the spectral cores induce a secondary mean meridional circulation in the Tropics, akin to circulation anomalies associated with the westerly phase of the QBO. Plumb and Bell (1982) show that the easterly and westerly phases of the QBO induce opposite secondary circulation in the Tropics. While the easterly phase induces a clockwise circulation in the Tropics, the westerly phase induces a secondary anti-clockwise circulation.

Specifying the tropical winds does not eliminate the inter-model differences altogether. As is seen in Figure 10a, e, while constraining the tropical winds eliminates the uncertainties in diabatic circulation strength in the Tropics, it also leads to a slight increase in circulation

spread at higher latitudes in the lower stratosphere around 400 K. This spread is seen more clearly in Figure 10h, where the models tend to have slightly different circulation at 400 K in the 45–75°N region. While the spread is reduced as we look higher in the stratosphere, Figure 10g indicates that the GFDL-FV3 core maintains a slightly stronger circulation strength than the other cores at 500 K.

6.2 | Tropical westerlies and isentropic mixing

Jets associated with the QBO also impact isentropic mixing between the Tropics and midlatitudes. Enhanced mixing between the Tropics and the Extratropics is observed during the westerly QBO phase, as the westerly winds allow midlatitude Rossby waves to penetrate deeper into the Tropics (e.g., Holton and Tan, 1980). The leaky pipe model of stratospheric transport in turn shows how enhanced mixing will increase the age-of-air throughout the stratosphere (Neu and Plumb, 1999). The enhanced mixing flux transports older midlatitude air into the Tropics, increasing the age in the Tropics. Naively, one might expect the age of midlatitude air to decrease as younger tropical air is brought in, but this effect is counteracted by an increase in the age of the air circulating to the Extratropics from the Tropics (Neu and Plumb, 1999; Linz *et al.*, 2016 give more detailed explanations). The net result is that mixing increases the age above, leaving the gross gradient untouched.

To establish that the tropical jets drive this mixing effect, we perform an additional specified wind profile experiment with the low-resolution (C48L40) GFDL-FV3 core. This allows us to compare the SP integration, with easterly winds throughout the tropical stratosphere (using the standard profile given in Appendix B), to a second integration, where the tropical winds were damped to a westerly profile. The westerly profile was taken from the 1° L80 (NE30L80) FR integration with the CAM-SE model (Figure 2, solid green). Winds from the 1° CAM-SE model were chosen because the tropical westerly jet was strongest in this integration. The lower vertical resolution FV3 grid was intentionally chosen to show that most of the transport differences can be explained solely by the differences in tropical winds.

Figure 11 compares the diabatic circulation (thick curves) and age-of-air (thin curves) between the GFDL-FV3 and CAM-SE runs at three different isentropic levels in the stratosphere. At 380 K in the lower stratosphere, Figure 11c, the tropical winds have virtually no impact on the age-of-air and mass transport. While there are some differences in the age-of-air between 15 and
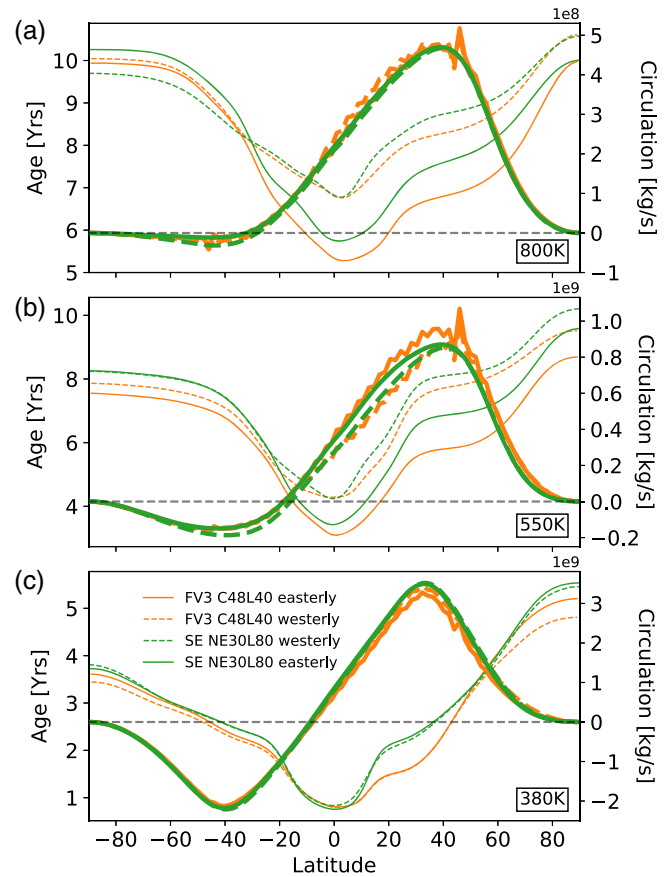

FIGURE 11 Age-of-air (bold curves; left axis) and diabatic circulation (thin curves; right axis) at three different isentropic levels in the stratosphere: (a) 800 K, (b) 550 K, and (c) 380 K, for the FR and SP integrations of CAM-SE, the FR integration of GFDL-FV3, and the experiment in which tropical westerlies were imposed on GFDL-FV3. The solid curves show the age and circulation for the CAM-SE and GFDL-FV3 core with easterly winds, i.e. NE30L80 (SP) and GFDL-FV3 (FR). The dashed curves show the same but for the runs with westerly winds, i.e. NE30L80 (FR) and GFDL-FV3 (SP). The GFDL-FV3 is shown in orange and the CAM-SE in green. As noted in Figure 10, the spikes in the GFDL-FV3 circulation profile are an artifact of the mass-conserving interpolation to a lat–lon grid

40°N at 380 K, these are unrelated to the tropical winds themselves, rather reflecting differences in synoptic-scale mixing between the two cores.

The situation changes as we proceed higher up in the stratosphere. At 550 K (~50 hPa), there is a split between integrations with tropical westerlies (dashed) and tropical easterlies (solid), Figure 11b. The diabatic circulation is reduced in both the low-resolution GFDL-FV3 and high-resolution CAM-SE integrations with tropical westerlies. The change in mass transport is accompanied by a change in the age. Thus, we see that the underlying numerics of the dynamical core strongly affect the transport, but chiefly through their impact on the resolved tropical winds structure (and hence the diabatic circulation and
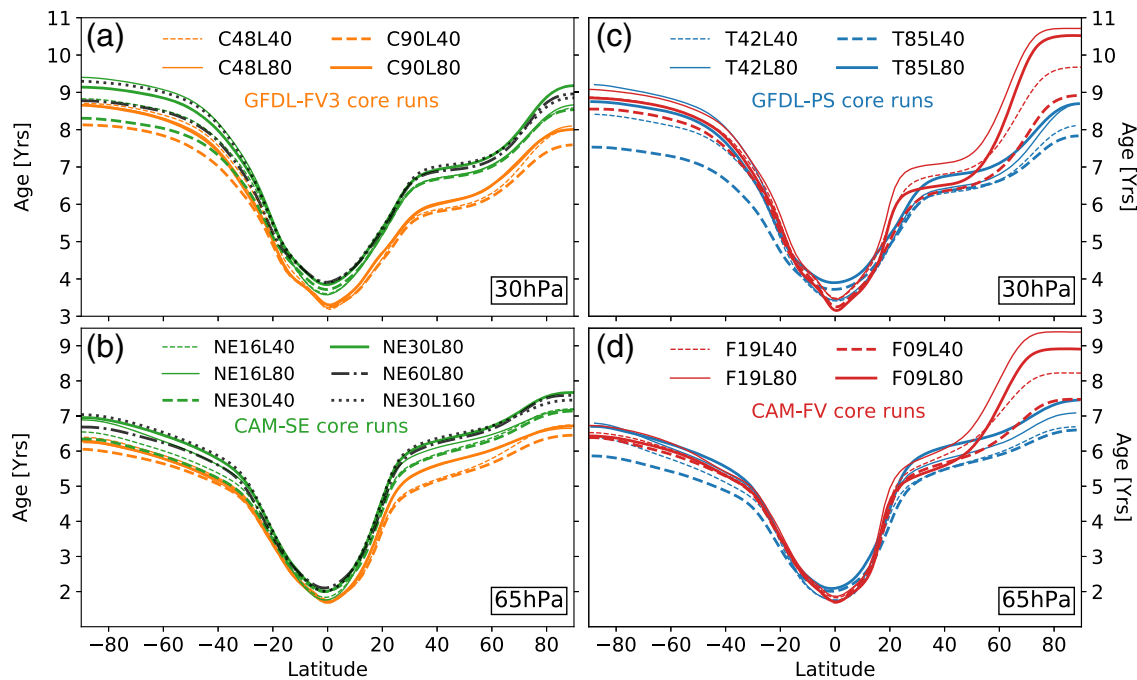
FIGURE 12 Age-of-air (years) from the specified (SP) integrations from all four dycores at all four resolutions: (a, b) show the age for GFDL-FV3 and CAM-SE at (a) 30 hPa and (b) 65 hPa; (c, d) show the age for GFDL-PS and CAM-FV at (c) 30 hPa and (d) 65 hPa. The low (high) horizontal resolution runs are shown using lighter (darker) curves and the low (high) vertical resolution runs are shown using dashed (solid) curves. GFDL-PS is shown in blue, GFDL-FV3 in orange, CAM-SE in green, and CAM-FV in red. In (a, b), higher horizontal (0.5° L80; NE60L80) and higher vertical (1° L160; NE30L160) resolution CAM-SE runs are shown using dashed-dotted and dotted black curves

isentropic mixing), not because of major differences in the tracer advection suite.

At 800 K (∼10 hPa), Figure 11c, the diabatic mass transport is similar across all the integrations, consistent with the fact that at this level, the climatological winds are similar among the four runs. However, the differences in age-of-air continue to increase. This is a signature of a change in mixing. The increase in isentropic mixing across the barrier between the Tropics and midlatitudes, facilitated by the westerly jets, makes the air older throughout the whole stratosphere above it (Neu and Plumb, 1999). At 800 K, the GFDL-FV3 and the CAM-SE westerly runs exhibit very similar age profiles, both nearly 2 years older than their easterly counterparts.

The findings suggest the critical importance of tropical momentum balance in stratosphere-resolving models. In comprehensive climate models, biases in representation of tropical stratosphere would lead to significant biases in transport of active trace species including ozone and water vapour. While tuning a gravity wave parametrization to force the QBO can help eliminate these biases, it is unlikely that the parametrization will provide an accurate representation of gravity wave activity. Rather the parametrization is being used to correct for other biases in the model numerics. There is substantial spread in the response of the QBO period to global warming, which is chiefly attributed

to uncertainty in the parametrization of gravity waves (e.g., Richter *et al.*, 2020). Based on the behaviour of the dycores in our study and other experiments (Yao and Jablonowski 2015; 2016), we suspect that part of this uncertainty also stems from the numerical formulation of the dynamical cores.

7 | NUMERICAL IMPACTS ON TRANSPORT

Specifying the tropical winds in the SP test brings all the models into the same flow regime, independent of resolution. However, as observed in Figures 8b and 9, differences persist in their representation of age. In particular, we still observe a split between the cores with spectral versus finite-volume numerics, although the difference is not nearly as large as in the FR runs. It is tempting to attribute this to fundamental differences between finite-volume and spectral numerical schemes. Apart from their spectral convergence, however, the numerical schemes in the GFDL-PS and CAM-SE differ in almost all respects. Likewise, the two FV cores have been built on different grids, with important differences in their treatment of vertical advection. To further probe these differences, we document differences in the age distribution as a function of model resolution.

7.1 | Impact of grid resolution

Figure 12 compares the age-of-air across all SP integrations at two levels chosen to characterize the structure of the lower and upper stratosphere. We first focus on the winter (Northern) hemisphere. For the CAM-FV core (red), the age throughout much of the winter midlatitudes and poles is very sensitive to both vertical and horizontal resolution. The integrations with higher vertical resolution (F19L80 and F09L80) exhibit notably older air in the polar region, indicating a more isolated polar vortex. This is consistent with the relatively low frequency of sudden warming events in the core, as seen in Figure 6. As noted earlier, we are concerned that the SP runs with this model may not accurately reflect the performance of the core. Following HS94, we consider the small-scale damping a part of the numerical method, and used the default settings for all cores. For CAM-FV, this entailed second-order divergence damping. Lauritzen *et al.* (2012) found that CAM-FV simulates inordinately high polar vortex winds at high horizontal resolution (1° and 0.5°), unless fourth-order divergence damping is used.

The age simulated by the other models is more robust to changes in resolution, particularly in the more modern cores (GFDL-FV3 and CAM-SE). In the GFDL-PS core, transport is sensitive to vertical resolution. The age throughout the winter hemisphere is nearly identical in the two low vertical resolution (T42L40 and T85L40; dashed blue lines in Figure 12c,d). At higher vertical resolution, the age increases, particularly in the T85L80 integration (dark blue), where age within the polar vortex increases by approximately 1 year. The increase in age with vertical resolution is consistent with a reduction of vertical diffusion, an expected impact of increasing the vertical resolution.

The age increases with vertical resolution in the GFDL-FV3 and CAM-SE cores as well, albeit to a lesser extent. Transport by the GFDL-FV3 model is remarkably insensitive to the horizontal resolution; the age profiles for the two 40-layer and 80-level models lie over each other. The split between L40 and L80 integrations is not simply an issue of diffusion, however, as the age in the Tropics is consistent across all the GFDL-FV3 integrations. Rather, there is an increase in the gradient between the Tropics and Extratropics with resolution, which indicates a decrease in the overturning circulation (Linz *et al.*, 2016). A weaker overturning circulation implies a reduction in wave driving, consistent with the reduction of the frequency of SSWs in Figure 6. We therefore speculate that the key difference is the sensitivity of the model's resolved circulation, not the sensitivity of the tracer transport scheme.

As with GFDL-FV3, CAM-SE's representation of the age is fairly insensitive to changes in resolution.

The age profile is nearly identical for all the runs (Figure 12a,b; green curves) with a minor exception of the highest-resolution integration, NE30L80, which develops a slightly older age (about 0.2 years) throughout the winter hemisphere. For this core alone, we were able to perform two additional integrations, doubling the horizontal resolution (NE60L80, half-degree resolution) and doubling the vertical resolution (NE30L160, where the horizontal resolution is kept at 1°). The age in both of these integrations lines up on top of the NE30L80 integration, indicating that the age distribution has converged.

The summer hemisphere in our idealized forcing is perpetually quiescent, exhibiting weak isentropic mixing and diabatic overturning (Figure 10). It therefore provides a challenging test on the model numerics, as diffusion becomes comparatively more important. Here, the model spread in age increases with height, leading to greater intermodel age differences at 30 hPa compared to 65 hPa (Figure 12a–d). The GFDL-PS model at lower vertical resolution stands out in particular (bold dashed blue curve in Figure 12c, d). However, for all models, the age increases and becomes more consistent in the high vertical resolution integrations, suggesting the importance of limiting spurious vertical diffusion by the numerical scheme.

7.2 | Impact of numerical schemes

Our results highlight the importance of employing high vertical resolution to accurately model stratospheric transport. As shown in Figure 7, even our 40-layer models contain fairly high resolution in the stratosphere relative to high-top configurations of comprehensive AGCMs. Accurate simulation of the formation and recovery of the ozone depends not only on a proper treatment of the chemistry, but also an accurate representation of the transport (Karpechko *et al.*, 2013). We also find fundamental differences between models employing finite-volume versus spectral schemes. The age distribution in both the GFDL-FV3 and CAM-SE models show evidence of convergence, but the models are converging to different age distributions, as seen in Figure 12a, b. For example, the age in the midlatitude surf zone at 30 hPa and 45°N is 6 years in GFDL-FV3, but 7 years in CAM-SE. This difference exceeds the internal spread between the integrations of each model with varying resolution.

There are two key structural differences between the results obtained with the four models. First, the age within the Tropics is more homogeneous in the spectral cores, with little asymmetry between the hemispheres. In the two finite-volume cores, air is youngest right at the Equator, with a sharper gradient in the summer hemisphere. This indicates differences in isentropic mixing

within the Tropics. Second, the spectral cores tend to age more strongly in height; the gulf between CAM-SE and GFDL-FV3 widens at 30 hPa relative to 65 hPa. This difference in the vertical gradient indicates a higher rate of isentropic mixing between the Tropics and Extratropics in CAM-SE. Both of these differences therefore point to the representation of mixing and diffusion along isentropic surfaces. The fact that the models exhibit a similar mean overturning circulation indicates that they agree on the mixing of potential vorticity. Hence this may be exclusively an issue of how trace gas transport is represented by the numerical scheme.

Differences within the Tropics could be related to the conservation of Axial Angular Momentum (AAM). Past studies have focused on notable differences in angular momentum conservation between spectral- and finite-volume-based cores. In particular, Lauritzen *et al.* (2014) finds that the CAM-SE core, like the GFDL-PS core, conserves AAM very well. Moreover, the models capability to conserve AAM is not significantly affected by their vertical advection schemes. Such is not the case for the two finite-volume-based cores. Studies focused on superrotating extra-terrestrial atmospheres have found the spectral cores to represent the super-rotating jet more accurately than the finite-volume cores, indicating that the finite-volume cores might have a problem with angular momentum conservation. It is unclear how angular momentum conservation would affect tracer transport, but it may explain the differences in the FR integrations. Conservation of AAM would support the development of westerly jets (super-rotation) in the Tropics.

Numerical dissipation may be another important source of differences between the cores. Finite-volume cores employ flux-limiters to preserve stability and monotonicity in higher-order dispersive schemes, which can render them more diffusive, although these effects would be reduced with higher resolution. The spectral cores tend to be less dissipative, particularly for small-scale waves as the effects of diffusion are the strongest on these scales. Yao and Jablonowski, 2015 (their figure 12) found that the power spectrum from the CAM-FV cores was much weaker than that of the spectral cores. While they do not conclude that the weaker wave activity in CAM-FV is necessarily wrong, these differences in the representation of waves can potentially be a crucial factor in the tropical stratosphere.

Dissipation on small scales impacts the effective resolution of a model. The "effective order of accuracy" allows one to compare the numerical truncation errors as a function of resolution. The schemes employed by the GFDL-PS and CAM-SE cores are spectrally accurate, that is, exhibiting exponentially increasing accuracy with resolution, while the finite-volume schemes in GFDL-FV3 and CAM-FV are second- or third-order accurate (Lauritzen *et al.*, 2011; Lin *et al.*, 2016). We emphasize that this payoff in accuracy is achieved only in the limit of infinite resolution. For the resolutions considered in this study, it is unclear which model is most accurate.

Differences in the representation of small-scale gravity waves could possibly help explain differences in the FR integrations in particular, as their momentum fluxes play a key role in providing westerly acceleration in the tropical stratosphere (Baldwin and Dunkerton, 1999). Too much noise on small scales could lead to artificially strong mixing along isentropes, explaining the homogenization of age within the Tropics in the spectral models compared to the finite-volume cores. Small-scale noise could also increase mixing between the tropical pipe and midlatitude surf zone, and hence the increase in age with height observed in spectral models. These speculative claims are the subject of current research.

8 | CONCLUSIONS

We have proposed two tests to assess the climatological representation of the stratosphere–troposphere system in primitive equation models on the sphere, the dynamical cores of atmospheric general circulation models (AGCMs). They are a natural extension of the Held and Suarez (1994) intercomparison test, updated to probe the representation of stratosphere–troposphere coupling and tracer transport. An active stratosphere was established by using the equilibrium temperature profile developed by Polvani and Kushner (2002) and the addition of surface topography (Gerber and Polvani, 2009) to promote planetary-scale stationary waves. Tracer transport was assessed by including a clock tracer to allow computation of the age-of-air (e.g., Hall and Plumb, 1994), building on work by Orbe *et al.* (2012). Age-of-air is a measure of transport time-scales in the atmosphere, allowing one to diagnose the combined influence of the mean Lagrangian transport and isentropic mixing, as reviewed in Section 2.

To establish a benchmark for comparison, four different dynamical cores developed by the Geophysical Fluid Dynamics Laboratory (GFDL) and National Center for Atmospheric Research (NCAR) were assessed. This allowed us to compare pseudospectral, spectral element, and finite-volume numerical schemes employing latitude–longitude and cubed-sphere grids. The robustness of the circulation and transport to changes in model resolution was probed by integrating the models at four different resolutions, refining the grids in the horizontal and vertical directions, individually and in concert. The dynamical cores explored include GFDL-FV3, the finite-volume cubed-sphere core of the National Weather

Service forecasting model and CAM-SE, the newly introduced spectral element dynamical core for the Community Earth System Model.

The first test, referred to as FR for "free running", is closest in spirit to the Held and Suarez (1994) test, differing only in the revised temperature equilibrium profile and orography. However, the four models fail to develop a consistent representation of the large-scale circulation, particularly at high vertical resolution. We observed a split between the two models employing spectral methods (GFDL-PS and CAM-SE), which developed westerly jets at high vertical resolution, compared to the two cores with finite-volume numerics (GFDL-FV3 and CAM-FV), which tended to keep the easterly jets observed in all models at lower vertical resolution. It is well known that the momentum balance of the tropical stratosphere is subtle; this is the region of the atmosphere where the Quasi-Biennial Oscillation (QBO) is observed, a slow oscillation between westerly and easterly jets over a period of approximately 28 months.

While the zonal winds and meridional overturning outside the tropical stratosphere were largely unaffected by these differences, tracer transport throughout the entire stratosphere was profoundly affected. Westerly jets induce enhanced mixing along isentropic surfaces, as discussed in Section 6, leading to significant aging of the air throughout the winter stratosphere, by a factor of 25% or more. The large differences induced by the circulation made it nearly impossible to assess the importance of model numerics in tracer transport.

This motivated our second test, which we refer to as SP for "specified tropical winds", where the zonal winds in the tropical stratosphere are constrained by a weak Rayleigh drag to a prescribed easterly profile. This prescription of the tropical winds is justified in part by the fact that our idealized forcing (i.e., the absence of convection and small-scale topography) does not provide a realistic representation of gravity wave sources in the troposphere, and the horizontal and vertical resolution of the models is insufficient to accurately model the propagation and breaking (and so momentum deposition) of gravity waves in the free atmosphere (Giorgetta *et al.*, 2002; Garcia and Richter, 2019). Prescribing the winds is somewhat akin to specifying, or tuning the gravity wave parametrization to capture the QBO. The general circulation outside the Tropics is largely unaffected by this additional momentum forcing, although the SP integrations exhibit increased frequencies of SSW events in all the models. This is consistent with the observations of Holton and Tan (1980), who found that tropical easterlies inhibit the propagation of planetary waves into the Tropics, thereby steering them into the higher latitudes and inducing a weakening of the polar vortex.

With specified tropical winds, the model representation of the large-scale circulation is fairly consistent across the four cores. This reassuring result provides a modern update to the comparison between a pseudo-spectral and finite-difference core by Held and Suarez (1994) 25 years ago. Despite differences in underlying numerics, domain discretization, and resolution, the models exhibit overall agreement on the structure of the large-scale circulation, eddy temperature variance, and the eddy kinetic energy, as documented in Section 5. However, key differences still remain, suggesting that model numerics still matter for the representation of the stratosphere–troposphere system. In the troposphere, the position of the extratropical jets varies by approximately a degree of latitude, Figure 4a; while this difference is small relative to the grid spacing, it is on the order of projected changes in the jet stream (Barnes and Polvani, 2013) and consistent with biases in CMIP5 models (Wenzel *et al.*, 2016). In the stratosphere, there are notable differences in the strength of the polar vortex, the extent of the meridional overturning circulation into the polar region, and the frequency of SSWs (although the statistical significance of this final result could not be established, even with 8,000 days of integration).

We chose the age-of-air as a measure of transport because it captures the integrated effect of the mean Lagrangian circulation of mass and the mixing of trace gases along isentropic surfaces. As established by Hall and Plumb (1994) and Waugh and Hall (2002), the age provides information on the transport of actual trace gases by the atmospheric circulation. Figure 8 establishes a benchmark value for the age-of-air under the Held and Suarez (1994) and Polvani and Kushner (2002) forcing with an idealised topography as prescribed in Gerber and Polvani (2009), which can be used to test transport in dynamical cores. While differences remain, the four models provide an overall consistent estimate of the age, provided the tropical winds are constrained. We acknowledge that age distribution resulting from our prescribed forcing is considerably older than that estimated for the real atmosphere (e.g., Waugh and Hall, 2002), primarily due to the low tropopause. Nonetheless, it captures the qualitative structure of observed age, sufficient to test the ability of dynamical cores to capture the relevant transport processes.

As illustrated in the analytic "leaky pipe" model of Neu and Plumb (1999), and applied to full three-dimensional models by Linz *et al.* (2016), horizontal gradients in the age can be related to the mean overturning circulation. As explored in Section 6, we find some spread in the representation of the mean overturning circulation, particularly in the high latitudes. However, the spread in transport in the FR runs is dominated by differences in isentropic mixing. While the vertical gradient of age depends in part on the mean overturning, it is set primarily by isentropic mixing.

By comparing GFDL-FV3 and CAM-SE models with easterly and westerly jets, we found that changes in isentropic mixing driven by differences in the zonal wind structure dominate the changes in age in the FR integrations.

Our choice of a perpetual January climatology led to a considerably weaker circulation in the Southern Hemisphere, enhancing the relative role of diffusion in driving transport in the region. Even for the SP integrations, the age differences in the summer hemisphere are significant. However, for all cores, most of those differences are accounted by the low vertical resolution runs. As the vertical resolution is increased, the age increases and becomes more consistent, indicating the importance of limiting spurious vertical diffusion by the numerical scheme. Given that our "low-resolution" 40-level set-up already provides comparable resolution to high-top AGCMs (Figure 7), we speculate that resolution plays a role in transport biases in state-of-the-art models.

Even when the tropical winds are constrained, structural differences between the cores with spectral accuracy and finite-volume numerics persist. In particular, the age-of-air in the GFDL-FV3 and CAM-SE cores appears very stable to changes in the vertical and horizontal resolution, but the two models converge to different age-of-air profiles, as highlighted in Figure 12. The overturning circulation is slightly stronger in GFDL-FV3 (Figure 10), which will reduce the horizontal gradients in age and lead to an overall younger profile in this model. CAM-SE also exhibits a steeper gradient of age in the vertical, suggesting a higher rate of isentropic mixing. In addition, the finite-volume models indicate stronger gradients of age within the Tropics, with more asymmetry between the summer and winter hemispheres. This suggests that transport remains sensitive to model numerics, and future work will seek to assess the processes responsible for these changes.

Although the finite-volume cores provide a more consistent circulation across resolution in the FR test, we do not believe this necessarily indicates a better representation of the dynamics. We speculate that the spectral cores are potentially providing a more accurate representation of momentum transport by small-scale gravity waves, which tend to be more strongly damped by numerical diffusion in finite-volume numerical schemes at equivalent resolution, and are more accurately conserving angular momentum. In comprehensive models, these differences are usually eliminated by tuning a gravity wave parametrization. However, with such large differences in the resolved wave momentum transport, it is likely that these parametrizations are being used in part to compensate for numerical errors, as opposed to providing just an accurate representation of the actual gravity wave momentum transport.

In terms of the age distributions, finite-volume numerical methods are by construction designed to provide a conservative representation of tracer transport. The spread between the age structure in the spectral and finite-volume cores is likely due to differences in isentropic mixing across the barriers to transport in the subtropical stratosphere, and around the polar vortex. Future work will seek to more carefully attribute these differences to differences in resolved stirring (the filamentation of tracer gradients by breaking waves) and numerical diffusion on the grid scale.

ORCID

Aman Gupta https://orcid.org/0000-0002-2215-7135
Edwin P. Gerber https://orcid.org/0000-0002-6010-6638

REFERENCES

Anav, A., Friedlingstein, P., Kidston, M., Bopp, L., Ciais, P., Cox, P., Jones, C., Jung, M., Myneni, R. and Zhu, Z. (2013) Evaluating the land and ocean components of the global carbon cycle in the CMIP5 earth system models. *Journal of Climate*, 26(18), 6801–6843.

Andrews, D.G. and McIntyre, M.E. (1978) Generalized Eliassen–Palm and Charney–Drazin theorems for waves on axismmetric mean flows in compressible atmospheres. *Journal of the Atmospheric Sciences*, 35(2), 175–185.

Andrews, D.G. and McIntyre, M.F. (1976) Planetary waves in horizontal and vertical shear: asymptotic theory for equatorial waves in weak shear. *Journal of the Atmospheric Sciences*, 33(11), 2049–2053.

Ayarzagüena, B., Charlton-Perez, A.J., Butler, A.H., Hitchcock, P., Simpson, I.R., Polvani, L.M., Butchart, N., Gerber, E.P., Gray, L., Hassler, B., Lin, P., Lott, F., Manzini, E., Mizuta, R., Orbe, C., Osprey, S., Saint-Martin, D., Sigmond, M., Taguchi, M., Volodin, E.M. and Watanabe, S. (2020) Uncertainty in the response of sudden stratospheric warmings and stratosphere–troposphere coupling to quadrupled CO_2 concentrations in CMIP6 models. *Journal of Geophysical Research: Atmospheres*, 125(6). https://doi.org/10.1029/2019JD032345

Baldwin, M.P. and Dunkerton, T.J. (1999) Propagation of the Arctic Oscillation from the stratosphere to the troposphere. *Journal of Geophysical Research: Atmospheres*, 104(D24), 30937–30946.

Barnes, E.A. and Polvani, L. (2013) Response of the midlatitude jets, and of their variability, to increased greenhouse gases in the CMIP5 models. *Journal of Climate*, 26(18), 7117–7135.

Boyd, J.P. (1976) The noninteraction of waves with the zonally averaged flow on a spherical earth and the interrelationships on eddy

fluxes of energy, heat and momentum. *Journal of the Atmospheric Sciences*, 33(12), 2285–2291.

Brewer, A.W. (1949) Evidence for a world circulation provided by the measurements of helium and water vapour distribution in the stratosphere. *Quarterly Journal of the Royal Meteorological Society*, 75, 351–363.

Butchart, N. (2014) The Brewer–Dobson circulation. *Reviews of Geophysics*, 52(2), 157–184.

Butler, A.H., Charlton-Perez, A., Domeisen, D.I.V., Garfinkel, C., Gerber, E.P., Hitchcock, P., Karpechko, A.Y., Maycock, A.C., Sigmond, M., Simpson, I. and Son, S.-W. (2018). Sub-seasonal predictability and the stratosphere, Chapter 11 in Sub-seasonal to Seasonal Prediction: The Gap Between Weather and Climate Forecasting. Robertson, A., Vitart, F. (eds), Elsevier, Amsterdam, Netherlands.

Butler, A.H., Sjoberg, J.P., Seidel, D.J. and Rosenlof, K.H. (2017) A sudden stratospheric warming compendium. *Earth System Science Data*, 9(1), 63–76.

Butler, A.H., Thompson, D.W.J. and Heikes, R. (2010) The steady-state atmospheric circulation response to climate change-like thermal forcings in a simple general circulation model. *Journal of Climate*, 23(13), 3474–3496.

Carpenter, R.L., Droegemeier, K.K., Woodward, P.R. and Hane, C.E. (1990) Application of the Piecewise Parabolic Method (PPM) to meteorological modeling. *Monthly Weather Review*, 118(3), 586–612.

Chabrillat, S., Vigouroux, C., Christophe, Y., Engel, A., Errera, Q., Minganti, D., Monge-Sanz, B.M., Segers, A. and Mahieu, E. (2018) Comparison of mean age of air in five reanalyses using the BASCOE transport model. *Atmospheric Chemistry and Physics*, 18.

Charlton-Perez, A.J., Baldwin, M.P., Birner, T., Black, R.X., Butler, A.H., Calvo, N., Davis, N.A., Gerber, E.P., Gillett, N., Hardiman, S., Kim, J., Krüger, K., Lee, Y.-Y., Manzini, E., McDaniel, B.A., Polvani, L., Reichler, T., Shaw, T.A., Sigmond, M., Son, S.-W., Toohey, M., Wilcox, L., Yoden, S., Christiansen, B., Lott, F., Shindell, D., Yukimoto, S. and Watanabe, S. (2013) On the lack of stratospheric dynamical variability in low-top versions of the CMIP5 models. *Journal of Geophysical Research: Atmospheres*, 118(6), 2494–2505.

Chen, G., Orbe, C. and Waugh, D. (2017) The role of monsoon-like zonally asymmetric heating in interhemispheric transport. *Journal of Geophysical Research: Atmospheres*, 122(6), 3282–3298.

Chrysanthou, A., Maycock, A.C., Chipperfield, M.P., Dhomse, S., Garny, H., Kinnison, D., Akiyoshi, H., Deushi, M., Garcia, R.R., Jöckel, P., Kirner, O., Pitari, G., Plummer, D.A., Revell, L., Rozanov, E., Stenke, A., Tanaka, T.Y., Visioni, D. and Yamashita, Y. (2019) The effect of atmospheric nudging on the stratospheric residual circulation in chemistry–climate models. *Atmospheric Chemistry and Physics*, 19(17), 11559–11586.

Colella, P. and Woodward, P.R. (1984) The Piecewise Parabolic Method (PPM) for gas-dynamical simulations. *Journal of Computational Physics*, 54(1), 174–201.

Collins, W.J., Lamarque, J.-F., Schulz, M., Boucher, O., Eyring, V., Hegglin, M.I., Maycock, A., Myhre, G., Prather, M., Shindell, D. and Smith, S.J. (2017) AerChemMIP: quantifying the effects of chemistry and aerosols in CMIP6. *Geoscientific Model Development*, 10(2), 585–607.

Danabasoglu, G., Lamarque, J.-F., Bacmeister, J., Bailey, D.A., DuVivier, A.K., Edwards, J., Emmons, L.K., Fasullo, J., Garcia, R., Gettelman, A., Hannay, C., Holland, M.M., Large, W.G., Lauritzen, P.H., Lawrence, D.M., Lenaerts, J.T.M., Lindsay, K., Lipscomb, W.H., Mills, M.J., Neale, R., Oleson, K.W., Otto-Bliesner, B., Phillips, A.S., Sacks, W., Tilmes, S., van Kampenhout, L., Vertenstein, M., Bertini, A., Dennis, J., Deser, C., Fischer, C., Fox-Kemper, B., Kay, J.E., Kinnison, D., Kushner, P.J., Larson, V.E., Long, M.C., Mickelson, S., Moore, J.K., Nienhouse, E., Polvani, L., Rasch, P.J. and Strand, W.G. (2020) The Community Earth System Model version 2 (CESM2). *Journal of Advances in Modeling Earth Systems*, 12(2). https://doi.org/10.1029/2019MS001916.

Dennis, J.M., Edwards, J., Evans, K.J., Guba, O., Lauritzen, P.H., Mirin, A.A., St-Cyr, A., Taylor, M.A. and Worley, P.H. (2012) CAM-SE: a scalable spectral element dynamical core for the Community Atmosphere Model. *International Journal of High Performance Computing Applications*, 26(1), 74–89.

Dobson, G. (1956) Origin and distribution of the polyatomic molecules in the atmosphere. *Proceedings of the Royal Society of London. Series A. Mathematical and Physical Sciences*, 236(1205), 187–193.

Edmon, H.J., Hoskins, B.J. and McIntyre, M.E. (1980) Eliassen–Palm cross-sections for the troposphere. *Journal of the Atmospheric Sciences*, 37(12), 2600–2616.

Engel, A., Möbius, T., Bönisch, H., Schmidt, U., Heinz, R., Levin, I., Atlas, E., Aoki, S., Nakazawa, T., Sugawara, S., Moore, F., Hurst, D., Elkins, J., Schauffler, S., Andrews, A. and Boering, K. (2009) Age of stratospheric air unchanged within uncertainties over the past 30 years. *Nature Geoscience*, 2(1), 28–31.

Eyring, V., Bony, S., Meehl, G.A., Senior, C.A., Stevens, B., Stouffer, R.J. and Taylor, K.E. (2016) Overview of the Coupled Model Intercomparison Project Phase 6 (CMIP6) experimental design and organization. *Geoscientific Model Development*, 9(5), 1937–1958.

Eyring, V., Harris, N.R.P., Rex, M., Shepherd, T.G., Fahey, D.W., Amanatidis, G.T., Austin, J., Chipperfield, M.P., Dameris, M., Forster, P.M.D.F., Gettelman, A., Graf, H.F., Nagashima, T., Newman, P.A., Pawson, S., Prather, M.J., Pyle, J.A., Salawitch, R.J., Santer, B.D. and Waugh, D.W. (2005) A strategy for process-oriented validation of coupled chemistry–climate models. *Bulletin of the American Meteorological Society*, 86(8), 1117–1134.

Froidevaux, L., Kinnison, D.E., Wang, R., Anderson, J. and Fuller, R.A. (2019) Evaluation of CESM1 (WACCM) free-running and specified dynamics atmospheric composition simulations using global multispecies satellite data records. *Atmospheric Chemistry and Physics*, 19(7), 4783–4821.

Garcia, R.R. and Richter, J.H. (2019) On the momentum budget of the Quasi-Biennial Oscillation in the Whole Atmosphere Community Climate Model. *Journal of the Atmospheric Sciences*, 76(1), 69–87.

Garfinkel, C.I., Shaw, T.A., Hartmann, D.L. and Waugh, D.W. (2012) Does the Holton–Tan mechanism explain how the Quasi-Biennial Oscillation modulates the Arctic polar vortex?. *Journal of the Atmospheric Sciences*, 69(5), 1713–1733.

Gerber, E.P., Butler, A., Calvo, N., Charlton-Perez, A., Giorgetta, M., Manzini, E., Perlwitz, J., Polvani, L.M., Sassi, F., Scaife, A.A., Shaw, T.A., Son, S.-W. and Watanabe, S. (2012) Assessing and understanding the impact of stratospheric dynamics and variability on the earth system. *Bulletin of the American Meteorological Society*, 93(6), 845–859.

Gerber, E.P. and Polvani, L.M. (2009) Stratosphere–troposphere coupling in a relatively simple agcm: the importance of stratospheric variability. *Journal of Climate*, 22(8), 1920–1933.

Gettelman, A., Mills, M.J., Kinnison, D.E., Garcia, R.R., Smith, A.K., Marsh, D.R., Tilmes, S., Vitt, F., Bardeen, C.G., McInerny, J., Liu, H.-L., Solomon, S.C., Polvani, L.M., Emmons, L.K., Lamarque, J.-F., Richter, J.H., Glanville, A.S., Bacmeister, J.T., Phillips, A.S., Neale, R.B., Simpson, I.R., DuVivier, A.K., Hodzic, A. and Randel, W.J. (2019) The Whole Atmosphere Community Climate Model version 6 (WACCM6). *Journal of Geophysical Research: Atmospheres*, 124(23), 12380–12403.

Giorgetta, M.A., Manzini, E. and Roeckner, E. (2002) Forcing of the Quasi-Biennial Oscillation from a broad spectrum of atmospheric waves. *Geophysical Research Letters*, 29(8), 86–1–86–4.

Guba, O., Taylor, M. and St-Cyr, A. (2014) Optimization-based limiters for the spectral element method. *Journal of Computational Physics*, 267, 176–195.

Hall, N.M.J., Hoskins, B.J., Valdes, P.J. and Senior, C.A. (1994) Storm tracks in a high-resolution GCM with doubled carbon dioxide. *Quarterly Journal of the Royal Meteorological Society*, 120(519), 1209–1230.

Hall, T.M. and Plumb, R.A. (1994) Age as a diagnostic of stratospheric transport. *Journal of Geophysical Research*, 99(D1), 1059.

Hall, T.M., Waugh, D.W., Boering, K.A. and Plumb, R.A. (1999) Evaluation of transport in stratospheric models. *Journal of Geophysical Research: Atmospheres*, 104(D15), 18815–18839.

Hang, J., Sandberg, M. and Li, Y. (2009) Age of air and air exchange efficiency in idealized city models. *Building and Environment*, 44(8), 1714–1723.

Haynes, P.H., McIntyre, M.E., Shepherd, T.G., Marks, C.J. and Shine, K.P. (1991) On the "Downward Control" of extratropical diabatic circulations by eddy-induced mean zonal forces. *Journal of the Atmospheric Sciences*, 48(4), 651–678.

Held, I.M. and Suarez, M.J. (1994) A proposal for the intercomparison of the dynamical cores of atmospheric general circulation models. *Bulletin of the American Meteorological Society*, 75(10), 1825–1830.

Holton, J.R., Haynes, P.H., McIntyre, M.E., Douglass, A.R., Rood, R.B. and Pfister, L. (1995) Stratosphere–troposphere exchange. *Reviews of Geophysics*, 33(4), 403–439.

Holton, J.R. and Tan, H.-C. (1980) The influence of the equatorial Quasi-Biennial Oscillation on the global circulation at 50 mb. *Journal of the Atmospheric Sciences*, 37(10), 2200–2208.

Holzer, M. and Hall, T.M. (2000) Transit-time and tracer-age distributions in geophysical flows. *Journal of the Atmospheric Sciences*, 57, 20.

Huynh, H.T. (1997). Schemes and constraints for advection, pp. 498–503 in Fifteenth International Conference on Numerical Methods in Fluid Dynamics. Kutler, P., Flores, J., Chattot, J.-J. (eds), Springer, Berlin.

Karpechko, A.Y., Maraun, D. and Eyring, V. (2013) Improving Antarctic total ozone projections by a process-oriented multiple diagnostic ensemble regression. *Journal of the Atmospheric Sciences*, 70(12), 3959–3976.

Kent, J., Ullrich, P.A. and Jablonowski, C. (2014) Dynamical core model intercomparison project: tracer transport test cases. *Quarterly Journal of the Royal Meteorological Society*, 140, 1279–1293.

Kida, H. (1983) General circulation of air parcels and transport characteristics derived from a hemispheric GCM. Part 2. Very long-term motions of air parcels in the troposphere and stratosphere. *Journal of the Meteorological Society of Japan*, 61(4), 510–523. 10.2151/jmsj1965.61.4_510.

Kidston, J., Scaife, A.A., Hardiman, S.C., Mitchell, D.M., Butchart, N., Baldwin, M.P. and Gray, L.J. (2015) Stratospheric influence on tropospheric jet streams, storm tracks and surface weather. *Nature Geoscience*, 8(6), 433–440.

Lauritzen, P.H., Jablonowski, C., Taylor, M.A. and Nair, R.D. (Eds.) (2011) *Numerical Techniques for Global Atmospheric Models.* Springer, Berlin.

Lauritzen, P.H., Mirin, A.A., Truesdale, J., Raeder, K., Anderson, J.L., Bacmeister, J. and Neale, R.B. (2012) Implementation of new diffusion/filtering operators in the CAM-FV dynamical core. *International Journal of High Performance Computing Applications*, 26(1), 63–73.

Lauritzen, P.H., Bacmeister, J.T., Dubos, T., Lebonnois, S. and Taylor, M.A. (2014) Held–Suarez simulations with the Community Atmosphere Model Spectral Element (CAM-SE) dynamical core: a global axial angular momentum analysis using Eulerian and floating Lagrangian vertical coordinates. *Journal of Advances in Modeling Earth Systems*, 6(1), 129–140.

Lauritzen, P.H., Nair, R.D., Herrington, A.R., Callaghan, P., Goldhaber, S., Dennis, J.M., Bacmeister, J.T., Eaton, B.E., Zarzycki, C.M., Taylor, M.A., Ullrich, P.A., Dubos, T., Gettelman, A., Neale, R.B., Dobbins, B., Reed, K.A., Hannay, C., Medeiros, B., Benedict, J.J. and Tribbia, J.J. (2018) NCAR release of CAM-SE in CESM2.0: a reformulation of the spectral element dynamical core in dry-mass vertical coordinates with comprehensive treatment of condensates and energy. *Journal of Advances in Modeling Earth Systems*, 10(7), 1537–1570.

Lin, J., Benson, R., Chen, H., Chen, X., Harris, L., Liang, Z., Morin, M., Putman, W., Rees, S. and Zhou, L. (2016). A brief overview of the FV3 dynamical core. NOAA Geophysical Fluid Dynamics Laboratory, Princeton, NJ.

Lin, S.-J. (2004) A "Vertically Lagrangian" finite-volume dynamical core for global models. *Monthly Weather Review*, 132(10), 2293–2307.

Lin, S.-J. and Rood, R.B. (1996) Multidimensional flux-form semi-Lagrangian transport schemes. *Monthly Weather Review*, 124(9), 2046–2070.

Linz, M., Abalos, M., Glanville, A.S., Kinnison, D.E., Ming, A. and Neu, J.L. (2019) The global diabatic circulation of the stratosphere as a metric for the Brewer–Dobson circulation. *Atmospheric Chemistry and Physics*, 19(7), 5069–5090.

Linz, M., Plumb, R.A., Gerber, E.P., Haenel, F.J., Stiller, G., Kinnison, D.E., Ming, A. and Neu, J.L. (2017) The strength of the meridional overturning circulation of the stratosphere. *Nature Geoscience*, 10(9), 663–667.

Linz, M., Plumb, R.A., Gerber, E.P. and Sheshadri, A. (2016) The relationship between age of air and the diabatic circulation of the stratosphere. *Journal of the Atmospheric Sciences*, 73(11), 4507–4518.

Lorenz, D.J. and DeWeaver, E.T. (2007) Tropopause height and zonal wind response to global warming in the IPCC scenario integrations. *Journal of Geophysical Research: Atmospheres*, 112(D10).

Matthes, K., Marsh, D.R., Garcia, R.R., Kinnison, D.E., Sassi, F. and Walters, S. (2010) Role of the QBO in modulating the influence of the 11-year solar cycle on the atmosphere using constant forcings. *Journal of Geophysical Research: Atmospheres*, 115(D18). https://doi.org/10.1029/2009JD013020.

Mcinturff, R.M.(ed.). (1978) *Stratospheric warmings: synoptic, dynamic and general-circulation aspects*. Technical report NASA-RP-1017, Washington DC.

McIntyre, M.E. and Palmer, T.N. (1983) Breaking planetary waves in the stratosphere. *Nature*, 305(5935), 593–600.

McIntyre, M.E. and Palmer, T.N. (1984) The 'surf zone' in the stratosphere. *Journal of Atmospheric and Terrestrial Physics*, 46(9), 825–849.

Murgatroyd, R.J. and Singleton, F. (1961) Possible meridional circulations in the stratosphere and mesosphere. *Quarterly Journal of the Royal Meteorological Society*, 87(372), 125–135.

Neu, J.L. and Plumb, R.A. (1999) Age of air in a "leaky pipe" model of stratospheric transport. *Journal of Geophysical Research: Atmospheres*, 104(D16), 19243–19255.

Orbe, C., Holzer, M. and Polvani, L.M. (2012) Flux distributions as robust diagnostics of stratosphere–troposphere exchange. *Journal of Geophysical Research: Atmospheres*, 117(D1).

Orbe, C., Holzer, M., Polvani, L.M. and Waugh, D. (2013) Air-mass origin as a diagnostic of tropospheric transport. *Journal of Geophysical Research: Atmospheres*, 118(3), 1459–1470.

Orbe, C., Oman, L.D., Strahan, S.E., Waugh, D.W., Pawson, S., Takacs, L.L. and Molod, A.M. (2017a) Large-scale atmospheric transport in GEOS replay simulations. *Journal of Advances in Modeling Earth Systems*, 9(7), 2545–2560.

Orbe, C., Waugh, D.W., Yang, H., Lamarque, J.-F., Tilmes, S. and Kinnison, D.E. (2017b) Tropospheric transport differences between models using the same large-scale meteorological fields. *Geophysical Research Letters*, 44(2), 1068–1078.

Pauluis, O., Czaja, A. and Korty, R. (2009) The global atmospheric circulation in moist isentropic coordinates. *Journal of Climate*, 23(11), 3077–3093.

Plumb, R.A. (2002) Stratospheric transport. *Journal of the Meteorological Society of Japan. Series II*, 80(4B), 793–809.

Plumb, R.A. and Bell, R.C. (1982) A model of the Quasi-Biennial Oscillation on an equatorial beta-plane. *Quarterly Journal of the Royal Meteorological Society*, 108, 335–352.

Polvani, L.M. and Kushner, P.J. (2002) Tropospheric response to stratospheric perturbations in a relatively simple general circulation model. *Geophysical Research Letters*, 29(7), 18–1–18-4.

Putman, W.M. and Lin, S.-J. (2007) Finite-volume transport on various cubed-sphere grids. *Journal of Computational Physics*, 227(1), 55–78.

Richter, J.H., Butchart, N., Kawatani, Y., Bushell, A.C., Holt, L., Serva, F., Anstey, J., Simpson, I.R., Osprey, S., Hamilton, K., Braesicke, P., Cagnazzo, C., Chen, C.-C., Garcia, R.R., Gray, L.J., Kerzenmacher, T., Lott, F., McLandress, C., Naoe, H., Scinocca, J., Stockdale, T.N., Versick, S., Watanabe, S., Yoshida, K. and Yukimoto, S. (2020) Response of the Quasi-Biennial Oscillation to a warming climate in global climate models. *Quarterly Journal of the Royal Meteorological Society*. https://doi.org/10.1002/qj.3749.

Rood, R.B. (1987) Numerical advection algorithms and their role in atmospheric transport and chemistry models. *Reviews of Geophysics*, 25(1), 71–100.

Shuckburgh, E., Norton, W., Iwi, A. and Haynes, P. (2001) Influence of the Quasi-Biennial Oscillation on isentropic transport and mixing in the tropics and subtropics. *Journal of Geophysical Research: Atmospheres*, 106(D13), 14327–14337.

Simmons, A.J. and Burridge, D.M. (1981) An energy and angular-momentum conserving vertical finite-difference scheme and hybrid vertical coordinates. *Monthly Weather Review*, 109(4), 758–766.

Sobel, A.H., Plumb, R.A. and Waugh, D.W. (1997) Methods of calculating transport across the polar vortex edge. *Journal of the Atmospheric Sciences*, 54(18), 2241–2260.

Son, S.-W. and Lee, S. (2005) The response of westerly jets to thermal driving in a primitive equation model. *Journal of the Atmospheric Sciences*, 62(10), 3741–3757.

Spiteri, R.J. and Ruuth, S.J. (2002) A new class of optimal high-order strong-stability-preserving time discretization methods. *SIAM Journal on Numerical Analysis*, 40(2), 469–491.

Townsend, R.D. and Johnson, D.R. (1985) A diagnostic study of the isentropic zonally averaged mass circulation during the first GARP global experiment. *Journal of the Atmospheric Sciences*, 42(15), 1565–1579.

Vallis, G.K., Colyer, G., Geen, R., Gerber, E., Jucker, M., Maher, P., Paterson, A., Pietschnig, M., Penn, J. and Thomson, S.I. (2018) Isca, v1.0: a framework for the global modelling of the atmospheres of earth and other planets at varying levels of complexity. *Geoscientific Model Development*, 11(3), 843–859.

Waugh, D. and Hall, T. (2002) Age of stratospheric air: theory, observations, and models. *Reviews of Geophysics*, 40(4), 1–1–1-26.

Waugh, D.W., Toigo, A.D. and Guzewich, S.D. (2019) Age of Martian air: time scales for Martian atmospheric transport. *Icarus*, 317, 148–157.

Wenzel, S., Eyring, V., Gerber, E.P. and Karpechko, A.Y. (2016) Constraining future summer austral jet stream positions in the CMIP5 ensemble by process-oriented multiple diagnostic regression. *Journal of Climate*, 29(2), 673–687.

Williamson, D.L. (2007) The evolution of dynamical cores for global atmospheric models. *Journal of the Meteorological Society of Japan. Series II*, 85B, 241–269.

Williamson, D.L., Drake, J.B., Hack, J.J., Jakob, R. and Swarztrauber, P.N. (1992) A standard test set for numerical approximations to the shallow-water equations in spherical geometry. *Journal of Computational Physics*, 102(1), 211–224.

Wu, X., Yang, H., Waugh, D.W., Orbe, C., Tilmes, S. and Lamarque, J.-F. (2018) Spatial and temporal variability of interhemispheric transport times. *Atmospheric Chemistry and Physics*, 18(10), 7439–7452.

Yao, W. and Jablonowski, C. (2015) Idealized Quasi-Biennial Oscillations in an ensemble of dry GCM dynamical cores. *Journal of the Atmospheric Sciences*, 72(6), 2201–2226.

Yao, W. and Jablonowski, C. (2016) The impact of GCM dynamical cores on idealized sudden stratospheric warmings and their QBO interactions. *Journal of the Atmospheric Sciences*, 73(9), 3397–3421.

Zhao, M., Golaz, J.-C., Held, I.M., Guo, H., Balaji, V., Benson, R., Chen, J.-H., Chen, X., Donner, L.J., Dunne, J.P., Dunne, K., Durachta, J., Fan, S.-M., Freidenreich, S.M., Garner, S.T., Ginoux, P., Harris, L.M., Horowitz, L.W., Krasting, J.P., Langenhorst, A.R., Liang, Z., Lin, P., Lin, S.-J., Malyshev, S.L., Mason, E., Milly, P.C.D., Ming, Y., Naik, V., Paulot, F., Paynter, D., Phillipps, P., Radhakrishnan, A., Ramaswamy, V., Robinson, T., Schwarzkopf, D., Seman, C.J., Shevliakova, E., Shen, Z., Shin, H., Silvers, L.G., Wilson, J.R., Winton, M., Wittenberg, A.T., Wyman, B. and Xiang, B. (2018a) The GFDL global atmosphere and land model AM4.0/LM4.0: 1. Simulation characteristics with prescribed SSTs. *Journal of Advances in Modeling Earth Systems*, 10(3), 691–734.

Zhao, M., Golaz, J.-C., Held, I.M., Guo, H., Balaji, V., Benson, R., Chen, J.-H., Chen, X., Donner, L.J., Dunne, J.P., Dunne, K., Durachta, J., Fan, S.-M., Freidenreich, S.M., Garner, S.T., Ginoux, P., Harris, L.M., Horowitz, L.W., Krasting, J.P., Langenhorst, A.R., Liang, Z., Lin, P., Lin, S.-J., Malyshev, S.L., Mason, E., Milly, P.C.D., Ming, Y., Naik, V., Paulot, F., Paynter, D., Phillipps, P., Radhakrishnan, A., Ramaswamy, V., Robinson, T., Schwarzkopf, D., Seman, C.J., Shevliakova, E., Shen, Z., Shin, H., Silvers, L.G., Wilson, J.R., Winton, M., Wittenberg, A.T., Wyman, B. and Xiang, B. (2018b) The GFDL global atmosphere and land model AM4.0/LM4.0: 2. model description, sensitivity studies, and tuning strategies. *Journal of Advances in Modeling Earth Systems*, 10(3), 735–769.

APPENDICES

A: MODEL DESCRIPTION AND ADVECTION SCHEMES

The dycores employ very different numerical schemes for both tracer and momentum advection. Additionally, scale-separation between horizontal versus vertical motions motivates the use of different schemes for horizontal and vertical advection for both tracers and dynamical quantities, that is, heat and momentum. Table A1 provides a summary of the advection schemes employed by the dycores to transport the clock tracer.

Tracer Advection Schemes: The more recently developed dynamical cores (GFDL-FV3 and CAM-SE) implement vertical advection of both tracer and momentum using a vertically Lagrangian finite-volume scheme by Lin (2004). This is accompanied by regular remapping of the advected fields from Lagrangian levels to fixed pressure levels after every fixed number of time steps using a mass-, momentum-, and energy-conserving remapping scheme detailed in Lin (2004). The mapping algorithms slightly differ among the dycores as highlighted in the table. In the older cores, both GFDL-PS and CAM-FV accomplish vertical tracer advection using piecewise parabolic finite-volume methods (PPM; Colella and Woodward 1984), with CAM-FV using a slightly enhanced version (Carpenter *et al.*, 1990).

For tracer advection in the horizontal, the FV3 core uses a positive-definite, grid-based, finite-volume scheme proposed in Lin and Rood (1996). The advection is further constrained by numerical constraints proposed by Huynh (1997) for improved wave resolution and monotonicity. Different from GFDL-FV3, the CAM-SE uses a strong stability preserving (SSP) fourth-order Runge–Kutta (RK4) method to time-march the transport equation. Among the older cores, the CAM-FV uses piecewise parabolic method (PPM) for horizontal tracer advection (identical to vertical transport) and the GFDL-PS core uses the default spectral scheme to advect tracers in the spectral space. The core contains the options for grid tracers, but experimentation reveals conservation issues with this scheme.

Momentum Advection Schemes: Most of the dycores use very similar schemes to advect tracers and momentum, with only minor differences incorporated for operational optimizations. For instance, the GFDL-FV3 and CAM-SE cores use identical schemes to advect both tracers and momentum in the vertical. Similarly, the CAM-FV core also uses an enhanced PPM algorithm for both momentum and tracer advection in the vertical. The GFDL-PS core is distinct from the other three cores in this regard, as it uses a second-order finite-difference scheme for vertical momentum advection, in contrast to using PPM for vertical tracer advection. Although the core provides an option to use PPM for vertical momentum advection as well, we have used the default settings provided with each model for the purpose of this study. Further work could be done to optimize the behaviour of a given model.

In the horizontal, the FV3 core uses very similar schemes to advect both mass and tracer, that is, PPM with an additional monotonicity constraint by Huynh (1997), which is less diffusive than the original monotonicity constraint proposed in Colella and Woodward (1984). The CAM-FV also uses PPM for horizontal advection of momentum. The CAM-SE core uses spectral elements to discretize the momentum equation in the vertical. A different, three-stage strong stability-preserving Runge–Kutta method (Spiteri and Ruuth, 2002) is used for horizontal advection of tracers. The algorithm is formally second-order accurate (Lauritzen *et al.*, 2018). The GFDL-PS core uses the default spherical-harmonics-based spectral scheme to advect momentum in the horizontal.

Shape and mass conservation properties of the advection schemes: Temporal tracer advection in CAM-SE is both mass-conserving and shape-preserving, courtesy of the shape-preserving filters (Guba *et al.*, 2014) used by the tracer advection algorithm. In contrast, the GFDL-FV3 and CAM-FV core do not preserve tracer shapes *exactly*; however the associated errors are quite small. These cores also use local mass fixers to borrow mass locally from neighbouring cells in order to restore mass conservation and to ensure monotonicity of tracer

TABLE A1 Horizontal and vertical tracer advection schemes employed by the four dynamical cores considered in the study

Numerics	Horizontal tracer advection	Vertical tracer advection
GFDL–**P**seudo**s**pectral	Default Spectral Scheme	Piecewise Parabolic Method (Colella and Woodward, 1984)
GFDL–Cubed-Sphere **F**inite-**V**olume (GFDL-FV3)	Positive Definite Scheme as in Lin and Rood (1996) with Huynh constraint	Lin (2004) Vertically Lagrangian scheme. Remapping by monotonic and conservative cubic splines.
CAM–**S**pectral **E**lement	Spectral Elements	Lin (2004) Vertically Lagrangian scheme. PPM vertical remapping with mirroring at boundaries.
CAM–**F**inite-**V**olume	*Enhanced* Piecewise Parabolic Method (PPM) (Colella and Woodward, 1984; Carpenter *et al.*, 1990)	*Enhanced* Piecewise Parabolic Method (PPM) (Colella and Woodward, 1984; Carpenter *et al.*, 1990)

profiles. It should be noted that the vertical remapping scheme used by GFDL-FV3 is both monotonicity and mass conserving. Therefore, any small errors in mass conservation in FV3 are solely due to horizontal advection. The relatively older GFDL-PS core does not preserve monotonicity and mass in both the horizontal and vertical. Moreover, it does not employ any mass fixers to ensure global conservation of tracer. Global mass fixers can be problematic when investigating the stratosphere, as the corrections are often dominated by the troposphere, particularly for bottom-heavy tracers like water vapour.

Model resolution: We finally briefly explain the model resolution designations shown in Table 1. It is not trivial to compare grids across different numerical schemes, but we have sought grids to provide roughly comparable resolutions for the different dynamical cores. For the GFDL-PS core, we work with **T42** and **T85** truncation (**T**riangular truncation up to **42** and **85** spherical harmonics, respectively) where the model data are output to a 128×64 ($2.8° \times 2.8°$) and 256×128 ($1.4° \times 1.4°$) Gaussian grid, respectively. For the GFDL-FV3 core, we work with **C48** and **C90** horizontal native grids (six tiles with **48 × 48** and **90 × 90** points on the local grid) that are ultimately mapped onto a 192×96 ($1.875° \times 1.875°$) and 360×180 ($1° \times 1°$) latitude-longitude grid respectively using a second-order mass- and energy-conserving scheme. Assuming six points to at least marginally resolve a given wave, the notation was designed so that the numbers (i.e., T**42** or C**48**) are roughly comparable.

For the CAM-FV dycore, we use the F**19** and F**09** grids with 144×96 ($2.5° \times 1.9°$) and 288×192 ($1.25° \times 0.9°$) points on the lat–lon grid. Here, a smaller number indicates finer resolution, unlike the GFDL convention. Finally, for the CAM-SE dycore, we use the **NE16 NP4** and **NE30 NP4** grids: **16 × 16** and **30 × 30** native grid with **4 × 4** spectral element within each cell. However, as the boundaries of each spectral element are shared,

the effective resolution is 3×3 per element, so that NE16 is directly comparable to C48 resolution of the GFDL-FV3 model, and likewise for NE30 NP4 and C90. These grids are interpolated by default on to a considerably finer 256×129 and 512×257 lat–lon grid, respectively, using bilinear interpolation. For this model in particular, then, the grid resolution is not indicative of the effective resolution.

In the vertical, we use 40 levels for the coarse runs and refine it to 80 levels for high-resolution runs. Figure 7 compares our choice of model vertical levels with the levels used in more comprehensive AGCMs and coupled chemistry models from both GFDL and NCAR. To select the vertical levels, we use the angular momentum-conserving σ–pressure hybrid coordinate (Simmons and Burridge, 1981), which is a combination of pure σ (terrain-following) coordinates following the orography upto 500 hPa and pure pressure above 200 hPa. The intermediate region between 200 and 500 hPa ensures smooth linear transition from the terrain-following coordinates below to pure pressure coordinates above. The mean pressure levels in the stratosphere are decided using the equations provided in paragraph 21 of PK02.

B: SPECIFYING TROPICAL WINDS

Specifying a given wind profile in the tropical stratosphere is similar in spirit to specifying QBO winds in comprehensive climate models (e.g., Matthes *et al.*, 2010). The key differences is that we specify a fixed easterly throughout our integrations, instead of an oscillating wind pattern. The tropical zonal winds are nudged to a prescribed wind profile, u_{eq}, with a relaxation time-scale τ that depends only on the latitude ϕ and pressure p, as:

$$\partial_t u(\lambda, \phi, p, t) = -\frac{u(\lambda, \phi, p, t) - u_{eq}(p)}{\tau(\phi, p)}, \quad (B1)$$

where λ is the longitude. The analytical wind profile u_{eq} is

$$
u_{eq}(p) = \begin{cases} u_{sponge}, & 0 \leq p \leq 0.3 \text{ hPa}, \\ u_{1hPa} + \left(\dfrac{u_{sponge} - u_{1hPa}}{\log(0.3)} \right) \log(p), & 0.3 \text{ hPa} \leq p \leq 1 \text{ hPa}, \\ u_{200hPa} + \left(\dfrac{u_{200hPa} - u_{1hPa}}{\log(200)} \right) \{\log(p) - \log(200)\} + 10 \sin \left(\dfrac{\pi \log(p)}{\log(200)} \right), & 1 \text{ hPa} \leq p \leq 200 \text{ hPa}, \end{cases} \tag{B2}
$$

where $u_{sponge} = 0 \text{ m} \cdot \text{s}^{-1}$, $u_{1hPa} = -65 \text{ m} \cdot \text{s}^{-1}$ and $u_{200hPa} = -10 \text{ m} \cdot \text{s}^{-1}$. The analytic profile u_{eq} is plotted in solid black in Figure 2. Since the model sponge gradually damps the winds between 0.5 hPa and the model top, we prescribe a u_{eq} profile which steadily decreases to zero between 0.5 hPa and 0.3 hPa.

The damping extends from $-15°$ to $15°$ latitude, and from 200 hPa to the model top. It is strongest at the Equator, with a time-scale $\tau_{max} = 40$ days, which is equal to the thermal relaxation time-scale for the diabatic forcing. Above 3 hPa, the damping is increased to a time-scale of 10 days; this was found necessary, by trial and error, in order to constrain the upper stratospheric winds close to the imposed wind profile. Otherwise, on longer time-scales, the wind patterns were still dominated by mesospheric variability which varied considerably between the cores. Away from the Equator, the damping coefficient $1/\tau$ decreases as a Gaussian with a standard deviation of $\sigma = 5°$. The analytical expression for the damping coefficient is

$$
\tau = A(\phi, p)\tau_{max}, \tag{B3}
$$

where

$$
A(\phi, p) = 0.5 \, (1 + \tanh\{-0.1(p - 200)\} \\ + 3[1 + \tanh\{-2(p - 3)\}]) \exp\left(\frac{-\phi^2}{2\sigma^2} \right), \tag{B4}
$$

where ϕ is the latitude, and p is the pressure level.