The stratosphere: a review of the dynamics and variability

Neal Butchart
Met Office Hadley Centre (MOHC), Exeter, EX1 3PB, UK
Correspondence: Neal Butchart (neal.butchart@metoffice.gov.uk)

Abstract. Large-scale, intra-seasonal to inter-annual variability of the stratosphere is reviewed. Much of the variability is dynamical and induced by waves emanating from the troposphere. It is largely characterised by fluctuations in the strength of the polar vortex in winter and a quasi-biennial oscillation in the equatorial winds. Existing theories for the variability are generally formulated in terms of wave mean flow interactions, with refinements due, in part, to teleconnections between the tropics and extratropics. Climate and seasonal forecast models are able to reproduce much of the observed polar stratospheric variability and are increasingly successful in the tropics too. Compared to the troposphere the models display longer predictability timescales for variations within the stratosphere. Despite containing just under 17% of the atmosphere’s mass the stratosphere’s variability exerts a powerful downward influence on the troposphere that can affect surface extremes. The stratosphere is therefore a useful source of additional skill for surface predictions. However, a complete dynamical explanation for the downward coupling is yet to be established.

1 Introduction

The existence of the stratosphere, or at least a second layer to the Earth’s atmosphere, was first established at the beginning of the twentieth century when Assmann (1902) and Teisserenc de Bort (1902) independently noted from balloon observations that from altitudes of 10–12 km up to around 17 km (the maximum achieved by the balloons) the atmosphere is isothermal unlike lower down where the temperatures decrease with altitude. The term stratosphere from the French stratosphère meaning “sphere of layers” was introduced by Teisserenc de Bort who also gave us the term troposphere for the lowermost layer of the atmosphere. Nowadays it is recognised that the stratosphere is not isothermal and extends up to about 50 km with temperatures, in general, increasing with height (Figure 1). Above that temperatures decrease upward in the mesosphere before increasing upward again in the thermosphere (Figure 1). These three layers, stratosphere, mesosphere and thermosphere are often collectively referred to as the “middle atmosphere” (Andrews et al., 1987). They contain just under a fifth of the mass of the Earth’s atmosphere with the majority of that in the stratosphere (Baldwin et al., 2019). The focus of this review is the dynamics and dynamically driven variability of the stratosphere with some consideration given to how this influences the troposphere.

The stratosphere is also the home of the ozone-layer (McElroy and Fogal, 2008) which protects us from harmful ultra violet (UV) radiation (Boucher, 2010). In order to explain some early ozone observations Dobson et al. (1929) suggested

1Earlier stratosphere had been used as a geological term for part of the Earth’s crust though that use is now obsolete.

the concept of a global-scale mass circulation which, with later refinements, evolved into what we now know as the Brewer-Dobson circulation (Butchart, 2014). This is a dynamically driven circulation (see Section 2.2) and one consequence is it drives the seasonal and zonally averaged temperatures of the extratropical stratosphere away from radiative equilibrium. Instead, on seasonal timescales and large spatial scales, the climate of the stratosphere can be described as being in radiative-dynamical equilibrium (Section 2.2).

A similar conceptual global transport circulation was invoked by Brewer (1949) to explain why the stratosphere was observed to be dry. With very low amounts of water vapour the dynamical effects of latent heating due to water experiencing phase transitions can be ignored in the stratosphere. Negligible latent heating effects and strong stable stratification due to the temperatures increasing with height distinguishes stratospheric dynamics from tropospheric dynamics which generally involve moist processes.

Variability in stratospheric winds and temperatures arises mostly through dynamical processes though fluctuations in incoming solar radiation, temporary volcanic aerosol injections and changes in composition such as ozone depletion and recovery also contribute variability on a range of timescales. This review considers only the large-scale variability of dynamical origins for which the two most distinctive phenomena are sudden stratospheric warmings (SSWs; Section 2.5) which occur in high latitudes in winter, mostly in the Northern Hemisphere (NH), and the quasi-biennial oscillation (QBO; Section 3.1) observed in the
equatorial stratospheric winds. Comprehensive reviews of the QBO and SSWs can be found in the Baldwin et al. (2001, 2021, respectively) series of reviews.

The first recorded SSW was over Berlin in January 1952 when Scherhag (1952) measured a rise in stratospheric temperatures of ∼30 K in just two days. Measurements in following years showed that this was neither a one-off event, nor local, with similar SSWs occurring roughly every other year in the Boreal winter and the rapid increase in temperatures also seen in the zonal mean (Baldwin et al., 2021). Current theoretical understanding of SSWs originates from the pioneering research by Matsuno (1971) and relies on three principle dynamical mechanisms: propagation of Rossby waves from the troposphere, the waves interacting with the mean flow due to wave breaking and dissipation and an induced mean meridional overturning circulation (Haynes et al., 1991) leading to rapid adiabatic warming (i.e., the SSW). These same dynamical mechanisms drive the extratropical stratosphere away from radiative balance and explain the observed latitudinially varying structures of the zonal mean temperature and wind fields (Section 2.2). Intraseasonal, and interannual variability is, in turn, driven by variability in this dynamical forcing of the stratosphere away from radiative balance.

Tropical variability in the lower and middle stratosphere is dominated by the QBO which, like SSWs, was discovered from observations that became available in the 1950s (Ebdon and Veryard, 1961; Reed et al., 1961). The QBO is one of the most distinctive modes of natural variability seen anywhere in the Earth’s atmosphere that is not directly associated with the changing seasons. It is characterised by alternate layers of eastward and westward winds descending through the equatorial stratosphere roughly every 28 months. The original canonical model proposed to explain this oscillation (Lindzen and Holton, 1968; Holton and Lindzen, 1972) remains relevant today. Again this invokes vertical wave propagation from the troposphere and wave mean flow interactions though now the waves are equatorial Kelvin and Rossby-gravity waves plus a broad spectrum of gravity waves (Section 3.1).

Despite waves from the troposphere being instrumental in most of the dynamical variability seen throughout the stratosphere, fluctuations in the wave flux from the troposphere are not essential to obtain the variability. Indeed, the QBO arises mostly from the feedback between the waves and the mean flow rather than an oscillation in the wave sources and filtering within the troposphere (Anstey et al., 2022b). Likewise Hardiman et al. (2020) argue that, for monthly means, vacillations between the waves and mean flow can explain much of the sub-seasonal and interannual variability in the strength of the NH polar vortex. Teleconnections between the tropics and extratropics also feature in determining variability in the different regions with, for example, the strength of the NH polar vortex and likely occurrence of an SSW dependent on the phase of the QBO (Anstey and Shepherd, 2014). The reverse influence of the extratropics on the tropics was thought to be relatively weak (O’Sullivan, 1997) until two interruptions of the regular QBO cycles in recent years by waves from the extratropics (Newman et al., 2016; Osprey et al., 2016; Anstey et al., 2021) prompted a reassessment (Section 4.2).

Advances over the last two decades have established that variability in the stratosphere can exert a powerful downward influence on the troposphere both in the extratropics (Kidston et al., 2015) and the tropics (Haynes et al., 2021), and on a wide range of timescales. This has fundamentally changed the traditional view that the relatively small mass of the stratosphere restricts it to a passive or weak dynamical role in surface climate and weather. Instead it is now generally accepted that the dynamical coupling between the stratosphere and troposphere is two-way. Hence, a better understanding of stratospheric
variability together with improvements in its representation in global models (Gerber et al., 2012; Anstey et al., 2022a) will potentially bring benefits of more reliable climate projections (Gerber and Manzini, 2016) and more accurate predictions of surface weather (Butler et al., 2016; Domeisen et al., 2020a) and extremes (Domeisen and Butler, 2020).

2 Extratropical stratosphere

2.1 Zonal mean climate

Unlike the troposphere, the stratosphere does not feel the thermal inertia of the oceans and without dynamics would be close to radiative equilibrium at all latitudes. In this situation the thermal structure is with a thermal structure and underlying annual cycle determined by a balance between heating due to absorption of incoming radiation (mostly solar UV) and cooling due to infrared emissions (Shine, 1987). Maximum radiative heating occurs at the summer pole while the maximum cooling is located at the winter pole. At the equinoxes the maximum heating shifts to the equator with cooling at both poles. Consequently the winter pole is relatively cold while the summer pole is relatively warm. This leads to a meridional temperature gradient and, by gradient wind balance, an eastward2 circumpolar vortex in the winter hemisphere with westward flow in the summer hemisphere. These features are seen for the observed zonal mean stratospheric climate (Figure 2) though there are significant departures from the radiatively determined state particularly in winter [cf., Figure 2 with Shine (1987, Figures 6 and 8)]. During the polar night when there is no incoming solar radiation the observed high latitude stratosphere is significantly warmer than would be expected if temperatures were determined by the radiative cooling alone, especially in the NH (Shine, 1987). Correspondingly, the eastward “polar night jets” are weaker than for the radiatively determined state with the NH jet peaking at just over 40 ms\(^{-1}\) in the upper stratosphere (Figure 2e) compared to \(\sim 100\) ms\(^{-1}\) in the Southern Hemisphere (SH) (Figure 2d), which is also closer to the radiative equilibrium value (Shine, 1987).

Inter-hemispheric differences are also visible in the annual cycle of the climatological temperatures and zonal wind at 10 hPa (Figures 2e and 2f). The annual cycle amplitude is larger in the SH than the NH, and the transition from eastward to westward winds in the SH occurs eight and not six months after the corresponding transition in the NH. Despite these differences the underlying annual cycle throughout the extratropical stratosphere is fundamentally a radiatively driven phenomenon with an amplitude comparable to the dynamically driven intraseasonal and interannual variability (see below).

The departure of the stratosphere from radiative equilibrium is due to a hemispheric-scale mean meridional overturning circulation, sometimes referred to as the “diabatic circulation.” It consists of ascent at the tropical tropopause and throughout the tropical stratosphere, and descent in the extratropical stratosphere causing adiabatic cooling and warming, respectively. Murgatroyd and Singleton (1961) were the first to attempt a calculation of this circulation though a full theoretical understanding was not obtained until the 1970s (Section 2.2). Nonetheless, Murgatroyd and Singleton’s qualitative description of the structure of the circulation agrees well with current understanding of stratospheric dynamics.

\(^2\)Throughout this review eastward and westward will be used when referring to wind direction rather than westerly and easterly, respectively, as traditionally chosen by practical meteorologists.
Figure 2. Zonal and monthly mean 1979–2020 climatology for 1980–2016 based on MERRA2 (Gelaro et al., 2017) and hourly ERA5.1 re-analyses data (Hersbach et al., 2020) for January and July: (a) temperatures (K; colour shading), (c) and (d) zonal winds (m s\(^{-1}\); black contours) and climatological annual cycle at 10 hPa of monthly and zonal mean for (ea) temperatures January, (Kb) July and (fc) zonal winds (m s\(^{-1}\)) the annual cycle at 10 hPa. The contour interval for the winds is 10 apart from in panel (a) where it is 5. Negative m s\(^{-1}\) with negative values are indicated by dotted contours. Data kindly retrieved from the ERA5 archive and processed by Dr. Martin Andrews.
2.2 Wave-mean flow dynamics

Concurrent with Murgatroyd and Singleton’s calculations, Charney and Drazin (1961) deduced from theory that the upward propagation of Rossby waves into the stratosphere can only occur through zonal mean winds that are eastward relative to the wave phase speed, and then only if those winds are not too strong. Even when these conditions are met all but the largest scale waves are inhibited from propagating. The planetary scale waves that can propagate are mainly forced by surface topography and land-sea contrasts (Held et al., 2002; Garfinkel et al., 2020b), with some additional forcing resulting from baroclinic processes (Tung and Lindzen, 1979; Boljka and Birner, 2020). Based on the observed climatological zonal mean zonal winds (Figures 2e and 2da and 2b) this then explains why the circulation in the summer stratosphere is relatively zonal symmetric and why the deviations from zonal symmetry are larger in the NH winter than the SH winter. In general, the zonal asymmetries are dominated by the first and second zonal wave harmonics as illustrated by the example daily maps of geopotential height at 10 hPa shown in Figure 3. Also noticeable in the figure is the absence of zonal asymmetries in summer (panels (b) and (c)) and the weaker wave amplitudes in the SH compared to the NH (cf., panels (a) and (d)) which is due to the difference in the surface topography and the land-sea contrast between the hemispheres.

**Figure 3.** Geopotential height (m) and vectors of horizontal wind (m s\(^{-1}\)) at 10 hPa for (a) and (b) 14 January 2021, and (c) 16 July 2020 for the NH, and (b) 14 January 2021 and (d) 16 July 2020 for the SH. From UK Met Office daily operational analysis for 12Z.
Charney and Drazin (1961) also showed that the steady, non-dissipated Rossby waves they considered, had no effect on the mean flow. Similar theoretical relationships between waves and mountain (gravity) waves and the mean flow were derived by Eliassen and Palm (1961) and form the basis for a “non-acceleration theorem,” or “Eliassen-Palm theorem,” that states that, to second-order in wave amplitude, waves that are steady and conservative (i.e., waves that are not time-varying and not subjected to any forcing or dissipation) do not change the mean flow. Following further theoretical developments in the 1960s and ’70s (e.g., Dickinson, 1968; Matsuno, 1971; Uryu, 1973, 1974; Boyd, 1976), Andrews and McIntyre (1976, 1978) introduced a generalized form of the Eliassen-Palm and Charney-Drazin theorems expressed in terms of the “transformed Eulerian mean (TEM) equations,” which are now widely used for analyzing wave-mean flow interactions (Held, 2019, Section 8). The generalisation of Andrews and McIntyre (1976, 1978) is ubiquitous as it is relevant to both gravity and Rossby waves and is applicable in the tropics as well as the extratropics.

Andrews and McIntyre (1976, 1978) derived the TEM equations by defining a residual mean meridional circulation

\[
(\overline{v}^*, \overline{w}^*) = \left( \overline{v} - \rho_0^{-1} \frac{\partial}{\partial z} \left( \rho_0 \frac{\nu' \theta'}{\theta} \right), \overline{w} + \frac{1}{a \cos \phi} \frac{\partial}{\partial \phi} \left( \cos \phi \frac{\nu' \theta'}{\theta} \right) \right)
\]

and transforming the zonal mean thermodynamic and zonal momentum equations in latitude, log-pressure coordinates \((\phi, z)\) to

\[
\frac{\partial \theta^*}{\partial t} + \overline{v}^* \frac{\partial \theta^*}{\partial \phi} + \overline{w}^* \frac{\partial \theta^*}{\partial z} = Q - \rho_0^{-1} \frac{\partial}{\partial z} \left[ \rho_0 \left( \frac{\nu' \theta' \cos \phi}{a \theta^*} + \overline{w' \theta'} \right) \right]
\]

and

\[
\frac{\partial \pi^*}{\partial t} + \overline{v}^* \left[ \frac{1}{a \cos \phi} \frac{\partial}{\partial \phi} \left( \pi \cos \phi \right) - 2 \Omega \sin \phi \right] + \overline{w}^* \frac{\partial \pi^*}{\partial z} = \nabla \cdot \mathbf{F} + \overline{X},
\]

respectively. Here an overbar refers to a zonal average and a prime the deviation therefrom or “eddy.” Zonal, meridional and vertical velocities are denoted by \(u, v, \) and \(w, \) respectively; \(\theta\) is potential temperature; \(Q\) is the zonal mean diabatic (radiative) heating; \(\rho_0 = \rho_s \exp(-z/H)\) the basic state density with \(H\) defined as scale height for log-pressure coordinates, generally taken to be \(\sim 7\) km, and \(\rho_s\) a sea level reference density. The rotation rate and radius of the Earth are \(\Omega\) and \(a,\) respectively. On the right hand side of Equation 3 \(\mathbf{F}^\phi - \mathbf{F}^z\) is the Eliassen-Palm flux vector defined as

\[
\mathbf{F}^\phi = \rho_0 a \cos \phi \left( \frac{\nu' \theta'}{\theta^*} - u' \theta' \right)
\]

and

\[
\mathbf{F}^z = \rho_0 a \cos \phi \left[ \left( 2 \Omega \sin \phi - \frac{(\pi \cos \phi)_{\phi}}{a \cos \phi} \right) \frac{\nu' \theta'}{\theta^*} - u' \theta' \right],
\]

with

\[
\nabla \cdot \mathbf{F} = \frac{1}{a \cos \phi} \frac{\partial}{\partial \phi} (\mathbf{F}^\phi_{\cos \phi}) + \frac{\partial}{\partial z} \mathbf{F}^z.
\]

In the thermodynamic Equation 2 the second term on the right hand side can generally be neglected (Andrews and McIntyre, 1978) such that, for steady state conditions, the equation is isomorphic to that used by Murgatroyd and Singleton (1961) to
calculate the diabatic circulation. Dunkerton (1978) used this TEM equation to estimate the residual mean meridional circulation \((\overline{v^*}, \overline{w^*})\) for steady conditions and, as might be expected, his results agree remarkably well with those of Murgatroyd and Singleton (1961). From this Dunkerton deduced that, rather than the Eulerian mean meridional circulation \((\overline{v}, \overline{w})\), it is this residual mean circulation \((\overline{v^*}, \overline{w^*})\) incorporating dynamical contributions from the eddy heat and momentum fluxes that is responsible for driving the stratosphere away from radiative equilibrium. Full theoretical details of the actual dynamical mechanisms involved are presented in the classic paper on “downward control of the diabatic circulation” by Haynes et al. (1991) and are not repeated here. One aspect that is worth noting is the residual mean meridional circulation approximates Lagrangian transport and, on an hemispheric scale, forms the advective component of the Brewer-Dobson circulation, which also includes two-way mixing (Butchart, 2014).

For the TEM equations the non-acceleration relations derived by Eliassen and Palm (1961) reduce to the simple expression \(\nabla \cdot \mathbf{F} = 0\), i.e., the forcing term on the right hand side of the zonal momentum Equation 3 is zero. \(\mathbf{F}\) indicates the direction (Jucker, 2021) and magnitude of wave propagation in the meridional plane and, for small-amplitude waves, represents the flux of wave activity (Andrews, 1987). The observed climatological winter mean \(\nabla \cdot \mathbf{F}/(\rho_0 a \cos \phi)\) is negative throughout the stratosphere (Figure 4) and therefore has a westward acceleration (eastward deceleration) effect on the circumpolar vortex. This explains why the polar night jets (Figure 2c) are weaker than would be expected from purely radiative considerations (Section 2.1). The arrows in Figure 4 indicate that the Rossby waves are propagating upward from the troposphere though on reaching the stratosphere they turn equatorward. Essentially this is because of the variation in latitude of the refractive index (due mainly to its dependence on the Coriolis parameter \(2\Omega \sin \phi\)) in the wave equation (Matsuno, 1970) which determines the direction of propagation of Rossby waves in the \((\phi, z)\) plane (Butchart et al., 1982). As expected the wave fluxes are stronger in northern winter than in the southern winter due to the relatively weak surface forcing in the SH.

![Figure 4](image_url)

**Figure 4.** Climatological mean \(\mathbf{F}\) (arrows) and \(\nabla \cdot \mathbf{F}/(\rho_0 a \cos \phi)\) (contours at intervals \(\pm 0.5, \pm 1.0, \pm 2, \pm 4, \pm 8, \ldots \) ms\(^{-1}\)day\(^{-1}\), with dashed contours denoting negative values) for (a) December-February in the NH, and (b) June-August in the SH, based on ERA-Interim (Dee et al., 2011) for the period of 1979–2018. Note the equator is on the left in both panels. For a clearer visualization of the direction of wave propagation \(\mathbf{F}\) is scaled as \((p_\pi/p)^{0.6} \times (F(\phi), 100 \times F(z))\), where \(p\) is pressure, and \(p_\pi\) a reference sea level pressure taken as 1000 hPa. Figure kindly provided by Dr. Hua Lu and adapted from Figure 1b of Lu et al. (2015), with the SH climatology also included.
Rossby waves in the stratosphere are often transient and almost certainly thermally damped (Andrews et al., 1987) but a significant cause of a nonzero $\nabla \cdot \mathbf{F}$ and therefore acceleration or deceleration and, therefore, an eastward deceleration (westward acceleration) of the mean flow is wave-breaking, at least in the extratropics. Waves propagating vertically, that do not encounter a critical layer (i.e., where the phase speed approaches that of the mean flow and propagation is inhibited), attain large amplitudes due to the exponentially decreasing density with height and the waves eventually break. Waves also break on encountering a critical layer either horizontally or vertically. Both gravity and Rossby waves experience amplitude growth with height though in the extratropics the gravity waves. In the extratropical (and tropical) stratosphere most gravity waves predominantly propagate upward tend not to break until with some reaching the mesosphere where they contribute their breaking contributes significantly to driving the temperature away from radiative equilibrium (Leovy, 1964). On the other hand, as Rossby waves are mostly refracted equatorward before reaching the mesosphere (e.g., Figure 4) the breaking occurs in the stratosphere and is now commonly diagnosed in terms of potential vorticity.

2.3 Potential vorticity

Like potential temperature $\theta$, potential vorticity $PV$ is a quasi-conservative quantity in the stratosphere. With $\theta$ as the vertical coordinate

$$PV = -g(\zeta_\theta + 2\Omega \sin \phi) \frac{\partial \theta}{\partial p},$$

where $\zeta_\theta$ is the relative vorticity evaluated along isentropic surfaces (surfaces of constant $\theta$) and $g$ gravitational acceleration. In the stratosphere isentropic surfaces are quasi-horizontal. $PV$ generally increases poleward due to the Coriolis parameter $2\Omega \sin \phi$ in Equation 7 and it is these gradients that support Rossby-wave propagation. A key advance in our understanding came when McIntyre and Palmer (1983) presented coarse grained maps of the isentropic distribution of $PV$ derived from newly available satellite observations which showed that $PV$ gradients in the middle stratosphere were not uniform at all latitudes. Instead they identified a “main vortex” with steep $PV$ gradients at its edge surrounded by a “surf zone” exhibiting relatively weak large scale gradients (Figure 5). Today the main vortex is known as the “polar vortex.” McIntyre and Palmer (1983) argued that the tongues of $PV$ seen drawn out from the main vortex in their maps (cf., Figure 5b) gave the first “reasonably convincing direct view of the breaking of planetary scale Rossby waves,” and McIntyre and Palmer (1984) went on to claim these were, in fact, the “world’s largest breaking waves.”

$PV$ maps provide a useful diagnostic for separating the reversible effects of Rossby wave propagation, such as the distortion and displacement of the polar vortex, from the irreversible effects of the waves breaking. The latter is characterized by tongues of $PV$ drawn from the polar vortex (e.g., Figure 5) and the eventual nonlinear mixing of $PV$ into the surf zone. A corollary of this is an erosion of the vortex with a sharpening of the $PV$ gradients at its edge (Figure 5). Therefore both reversible and irreversible processes can contribute to the variability of the polar vortex.

---

3Vertically propagating gravity waves in the extratropics typically have wave lengths that are too small to be resolved by the reanalysis used for Figure 4 or, indeed, nearly all current climate models and, therefore, their effect on the mean flow is included in the term $\mathbf{X}$ in Equations 3 and not $\nabla \cdot \mathbf{F}$. 

McIntyre and Palmer (1983, 1984) conjectured that wave breaking and the concomitant vortex erosion occurs almost continuously throughout the winter. By simply measuring the size of the vortex in terms of the area enclosed within contours of constant $PV$ on the isentropic maps (Figure 6a), Butchart and Remsberg (1986) confirmed this for the winter of 1978/79, with the confirmation subsequently extended to all northern winters 1964-82 by Baldwin and Holton (1988). Figures 6a shows that, with wave breaking, the vortex weakens throughout winter compared to what would be expected if there were no waves and the stratosphere evolved radiatively (Figures 6b). If the amplitudes of the waves become sufficiently large the vortex is displaced off the pole or even split in two and usually this corresponds with the occurrence of an SSW (Section 2.5).

2.4 Polar vortex variability

In terms of the interannual standard deviation of the monthly and zonal mean zonal winds, the maximum variability of the polar night jet occurs in high latitudes in the NH winter but in the SH winter is displaced toward mid-latitudes (Shiotani et al., 1993; Kuroda and Kodera, 2001) and does not extend down into the lower stratosphere (Figures 7a and c, respectively). Instead the variability in the SH spring (Figure 7d) is more similar to that of the NH winter albeit slightly weaker, though stronger than the corresponding variability seen in the NH spring (Figure 7b). This is what is expected for variability resulting from Rossby wave forcing from the troposphere. The mid-winter SH jet is generally strong enough (Figure 2) to limit Rossby wave propagation (Section 2.2) and hence the variability, and the jet only becomes comparable in strength to the NH mid-winter jet...
Figure 6. Area enclosed by contours of potential vorticity on the 850 K isentropic surface (∼ 10 hPa) for (a) 1978/79 NH winter, (b) an idealized simulation of the NH winter in the absence of Rossby waves. Figures 4 and 6 of Butchart and Remsberg (1986), respectively. Details of the calculations and information on the data used can be found in Butchart and Remsberg (1986).

as it weakens in spring (Figure 2f). In contrast, by April in the NH the switch to the summer westward circulation has often already commenced (Figure 2f) and, again, this inhibits Rossby wave propagation.

An alternative approach for analyzing the interannual variability of the polar night jets is to use Empirical Orthogonal Functions (EOFs; Feser et al. (2000)). For reanalysis data from 1980 to 1999 the first EOF explains 87% of the variance in the zonal mean zonal wind at 50 hPa in the NH, while at the same level in the SH the first two EOFs explain 59% and 35% of the variance, respectively (Butchart et al., 2011). Variations in the jet strength are represented by the first mode of variability or EOF, while the second mode corresponds to a meridional shift of the jet. The two local maximums in the latitudinal structure of the interannual standard deviation for July (Figure 7c) are then simply a manifestation of the larger contribution from the second EOF in the SH, as compared to the NH, which results from year to year fluctuations in the latitude of the jet maximum.

Because the strength of the polar night jet modulates Rossby wave propagation from the troposphere (Section 2.2, Charney and Drazin, 1961) there is potential for the jet strength to vacillate. A weakening, followed by reversal of the eastward jet due to wave-mean flow interactions diminishes wave propagation from the troposphere. This then reduces the deceleration effect of the waves and radiative effects restore the jet's strength. More wave propagation is then allowed and wave-mean flow interactions begin to weaken the jet again, thus a vacillation cycle is created. Stratospheric vacillations of this form were first identified in idealised numerical experiments (Holton and Mass, 1976) and subsequently in a general circulation model (GCM) where they had a timescale of approximately 100 days (Christiansen, 1999). Establishing the role of vacillations in the observed variability is more challenging. Hardiman et al. (2020) found that for the 39 years of ERA-Interim (Dee et al.,
In contrast, Newman et al. (2001) concluded that the interannual variability of Arctic lower-stratospheric temperatures during spring is almost completely determined by the variability of the (meridional) eddy heat flux near the tropopause slightly earlier in the year. For large scale Rossby waves the heat flux ($\vec{v}''T''$, where $T$ is temperature) is approximately proportional to the vertical component of the EP-flux ($F^{(z)}$, Equation 5) and, hence, the vertical flux of wave activity. Over the 22 years 1979–2001 Newman et al. (2001) observed that mean temperatures poleward of 60°N at 50 hPa and averaged from 1 to 16 March were significantly correlated with the 45–75°N mean eddy heat flux at 100 hPa averaged from 15 January to 28 February. In particular, they found that a strong planetary wave eddy heat flux in the 100–400 hPa and 45–75°N region during January-February results in a warm March polar lower stratosphere, while a weak planetary wave eddy heat flux results in a cold March polar lower stratosphere. Newman et al. (2001) noted that their result is consistent with a theoretical framework based on linear wave propagation and also robust to the choice averaging periods and latitude ranges. A similar relationship exists between Antarctic lower-stratospheric temperatures in spring and the eddy heat flux emerging from the troposphere slightly earlier in
the year and for both hemispheres this relationship appears robust across a range of model simulations (Austin et al., 2003; Manzini et al., 2003; Eyring et al., 2006; Osprey et al., 2010).

During the extended winter season temperatures and zonal winds at high latitudes in the middle stratosphere can vary quite significantly from day to day in both hemispheres, as well as from year to year. For example, when the standard deviations of the daily variations in mean temperature poleward of 60°N, and zonal mean zonal wind at 60°N, are calculated at 10 hPa for October to March for each winter and then averaged for the 41 winters 1979/80 to 2019/20 the values are ∼7 K and ∼13 ms⁻¹, respectively, and comparable to the inter-annual spread in daily values shown by the light gray shading in Figures 8a and b. As noted above the strongest inter-annual (and also daily) variability in the SH occurs roughly two months later in the annual cycle than in the NH. Therefore the corresponding six month period in the SH is June to November and for the 41 winters 1980 to 2020 the averages of the standard deviations of the daily 10 hPa time series of the mean temperature poleward of 60°S, and zonal mean zonal wind at 60°S are ∼4 K and ∼9 ms⁻¹, respectively, slightly weaker than their NH counterparts. In both hemispheres daily variability is not uniform over the six months but includes periods when temperatures and winds evolve more steadily, similar to the summer months (Figure 8) when the dynamical forcing is weak or nonexistent. At other times significant and rapid rises in temperature and deceleration of eastward polar night jet (e.g., purple curves in Figure 8) are observed, especially in the NH. Corresponding rapid falls in temperature and/or a sudden acceleration of the jet are not

Figure 8. Daily temperature and winds from MERRA2 data (Gelaro et al., 2017) for 1979-2020. The black line is the multi-year average of the daily time series with 30th/70th and 10th/90th percentiles, and range shown by the dark, medium and light gray shading, respectively. The purple lines show the daily values for year 2008/09 in the NH and 2009 in the SH.
observed as the vertically upward propagating Rossby waves can only decelerate the polar night jet (Section 2.2)\(^4\). Instead, a strengthening of the polar vortex and decline in the anomalously high polar temperatures occurs on the longer radiative damping timescales as illustrated by the behaviour seen in 2008/09 (Figure 8a and b). Depending on the definition used, the events where there is a significant rapid rise in polar temperatures and a reversal from eastward to westward wind at 60° latitude, such as in January 2009 (e.g., Manney et al., 2009), are generally referred to as “sudden stratospheric warmings.”\(^5\) (Butler and Gerber, 2018).

2.5 Sudden stratospheric warmings

In the decade following Scherhag (1952)’s first observation of a mid-winter stratospheric warming (see Introduction) many more reports of anomalous warming events appeared in the literature (e.g., Keegan, 1962; Scherhag, 1960; Craig and Hering, 1959; Palmer, 1959; Teweles, 1958; Teweles and Finger, 1958) and by the early 1960s such events had acquired the name “stratospheric sudden warmings” (Reed et al., 1963), though today it is considered more appropriate to refer to these events as “sudden stratospheric warmings” (Butler et al., 2015). Initially SSWs were simply categorised as either mid-winter warmings or final warmings which are a manifestation of the year-to-year variability in the timing and vertical structure of the spring transition from an eastward to westward circulation in the middle and high latitude stratosphere (Matthias et al., 2021). In the 1960s warmings were further separated into “minor” and “major” (Butler et al., 2015). Minor warmings occur when a significant temperature increase is observed (i.e., at least 25 K in a period of a week or less) at any stratospheric level in any area of the wintertime hemisphere but the criteria for a major warming is not met. A warming is “major” if at 10 hPa or below the latitudinal mean temperature increases poleward from 60° latitude and an associated circulation reversal is observed (i.e., mean eastward winds poleward of 60° latitude are succeeded by mean westward winds in the same area). This criteria for distinguishing major from minor warmings is still used today and the occurrences of major warmings can be identified in Figures 8b and d as the times when the zonal winds fall below the zero wind line in the figures. This occurs only once in the SH (in 2002), i.e., only one major warming has been observed in the SH since 1979 and indeed since records began\(^5\). For any particular date during the NH winter the lightest gray shading in Figure 8b shows that there are westward winds in less than 10% of the years but, as SSWs can occur at any time during winter, major warmings are actually observed roughly every other year (Baldwin et al., 2021), though the exact frequency of occurrences varies from decade to decade (Domeisen, 2019) for reasons that are not yet fully understood.

Early research into understanding SSWs contributed to the development of the TEM theory of wave mean-flow interactions (Section 2.2) and it is now accepted that the weakening and reversal of the polar night jet that occurs during warmings results from anomalously large negative values of \(\nabla \cdot \mathbf{F}\) in Equation 3. This generally results from an amplification and a refraction of upward propagating Rossby waves poleward where they deposit westward wave-momentum mainly due to nonlinear wave

\(^{4}\)Even though there are no rapid falls in zonal mean temperatures, locally rapid temperature rises can be followed by rapid cooling (e.g., Scherhag, 1952) if, for instance, the polar vortex is displaced from and then back toward the pole.

\(^{5}\)In mid-September 2019 a significant weakening of the SH polar vortex occurred but as no actual wind reversal was observed poleward of 60° this event was not classified as a major SSW.
breaking (Section 2.3). A ramification of the non-zero $\nabla \cdot \mathbf{F}$ is an induced residual mean meridional circulation (Haynes et al., 1991) with adiabatic descent producing the observed warming of the polar stratosphere. Matsuno (1971) was the first to propose a model for SSWs based on planetary waves propagating from the troposphere and in this model SSWs are classified as either (zonal) wave number one or wave number two warmings depending on which is the most dominant wave (O’Neill and Taylor, 1979). Higher wave number warmings do not occur as only the largest scale Rossby waves can propagate up into stratosphere as a result of the Charney and Drazin (1961) theorem (Section 2.2). This then provides a dynamical distinction between major and minor warmings as the reversal of the flow at 10 hPa prohibits upward wave propagation beyond that level following a major warming. Consequently, after a major warming radiative cooling is the more dominant process and polar temperatures the strong radiative cooling causes the polar temperatures to fall, though at a slower rate than the preceding rise in temperatures, and SSWs are often followed by a period when the polar vortex is relatively cold and quiescent (e.g., see the winter 2008/09 curves in Figures 8a and b).

One aspect of the mechanism for SSWs that is less well understood is the cause of the amplification of the waves. The rarity of SH SSWs provides compelling evidence for the importance of the wave flux from troposphere as the SH topography and land-sea contrast is unable to produce sufficiently strong planetary wave forcing for more frequent SSWs as observed in the NH. However even in the NH anomalously large upward wave fluxes compared to the climatological values shown, for instance, in Figure 4 are essential for SSWs. Some studies have argued that the anomalous fluxes result from blocking (e.g., Quiroz, 1986; Martius et al., 2009) and/or other tropospheric precursor events (e.g., Taguchi and Hartmann, 2006; Cohen and Jones, 2011), while other studies have stressed the importance of the state of the stratosphere. One possibility originally proposed by Palmer (1981a) to explain the occurrence of the 1979 SSW is the stratosphere is “preconditioned” to favor enhanced upward wave propagation (Albers and Birner, 2014; Hitchcock and Haynes, 2016), as originally proposed by Tim Palmer (1981a) to explain the occurrence of the 1979 SSW propagation of Rossby and gravity waves (Albers and Birner, 2014; Hitchcock and Haynes, 2016). An alternative possibility, first suggested by Plumb (1981), is internal wave mean-flow feedbacks within the stratosphere leads to a resonant growth of Rossby wave amplitudes (Matthewman and Esler, 2011). Most likely both the state of stratosphere and tropospheric precursors are relevant for the generation of SSWs since it has been estimated that only about a third of SSWs are preceded by extreme anomalous upward planetary wave fluxes from the lower troposphere (Birner and Albers, 2017).

Despite the uncertainty over the precise mechanism for the generation of SSWs the geometric vortex parameters associated with the warming are known to be important (Albers and Birner, 2014). Hence the traditional mathematical classification of SSWs in terms of the dominant wave number is often supplemented by more physically based classifications of SSWs linked to the morphology of the polar vortex as diagnosed from maps of the isentropic distribution of $PV$ (Waugh, 1997). Charlton and Polvani (2007) were the first to use such an approach and classified SSWs as either “(vortex) displacement,” characterized by a clear shift of the polar vortex off the pole, or “(vortex) split,” when the polar vortex breaks into two separate vortices of comparable size. Examples of displaced and split warming events are shown in Figure 9b and c, respectively. A simple relationship between the geometric vortex parameters and a Fourier decomposition of the fields does not exist (Waugh, 1997) but wave number one warmings generally result in a vortex displacement and the presence of wave number two is essential for a split vortex. Classifying SSWs according to whether they are vortex splitting or vortex displacement events has proved useful
for tracking the nature and timing of the surface impacts of SSWs (Hall et al., 2021a). Therefore, understanding the generation of SSWs is likely to lead to a better understanding of their surface effects (Section 6.3).

![Figure 9](image)

**Figure 9.** Daily averaged PV (10^{-6} \text{ m}^2\text{s}^{-1} \text{ K kg}^{-1}) fields for three different states of the NH polar vortex on the 850 K isentropic surface: (a) the stable state of the vortex, (b) a displaced vortex, and (c) a vortex that has split into two daughter vortices. Black crosses mark the North Pole. Figure 1 from Mitchell et al. (2011).

### 3 Tropical stratosphere

Modes of variability in the tropical stratosphere are associated with distinct peaks in spectral density, or power, in the periodogram of the monthly and zonal mean zonal winds at the equator (Figure 10) (Pascoe et al., 2005). The first peak occurs at six months and represents the so-called “semi-annual oscillations (SAO)” which is essentially a mesospheric phenomenon that extends down into the upper stratosphere. It is driven mainly by the transport of zonal momentum by vertically propagating equatorial and gravity waves plus cross-equatorial meridional advection of mean momentum which locks the SAO to the seasonal cycle (Kawatani et al., 2020).

A second peak in the spectral density represents the annual cycle and is present throughout the depth of the stratosphere though the amplitude of the cycle is much smaller than that for either the SAO in the upper stratosphere or the QBO (see below) in the lower and middle stratosphere, at least for the zonal mean zonal wind (Figure 10). However, for the zonal mean temperature the annual cycle becomes the dominant mode at the tropical tropopause. It is driven by variations in adiabatic ascent (cooling) that result from the seasonal cycle in the diabatic (Brewer-Dobson) circulation (Yulaeva et al., 1994). As this seasonal cycle is a consequence of the inter-hemispheric differences in the wintertime extratropical wave driving (Section 2.2) the annual cycle in the tropical tropopause temperature is dynamically driven (e.g., Jucker and Gerber, 2017) in contrast to the radiatively determined annual cycle in the extratropical winds and temperature (Section 2.1).

The third peak in the spectral density in the periodogram shown in Figure 10 is broader than those for the SAO and annual cycle thereby indicating a range of frequencies, or periods — 22 to 40 months with an average of 28.5 months for the years.
Figure 10. Periodogram (Fourier analysis) of equatorial zonal mean zonal wind for 44 years of reanalysis data from 1958 to 2001. The Fourier harmonic axis indicates the number of cycles that a given periodic mode experienced during the 44 years. The periods of some Fourier harmonics are also indicated. Contours are drawn at Fourier amplitudes of 1, 2, 4, 8, and 16 ms\(^{-1}\). Figure 3 from Pascoe et al. (2005).

1958–2001 shown in the figure. It is also the dominant signal of variability in the equatorial zonal winds in the lower and middle stratosphere and results from the QBO.

3.1 Quasi biennial oscillation

Ebdon and Veryard (1961) and Reed et al. (1961) are independently credited with the discovery of quasi-periodic reversals of the prevailing winds at the equator from eastward to westward and back again roughly every 28 months in the lower and middle stratosphere. Corresponding oscillations were also identified in temperature and ozone and soon after its discovery this irregular oscillation became known as the “quasi-biennial oscillation” or QBO (Angell and Korshover, 1964). The original discovery was based on single station data near the equator but similar behaviour is seen for the zonal means. For the zonal mean zonal wind, the signal is centered at the equator with a latitudinal half width of about 12° (Baldwin et al., 2001). The most iconic image of the QBO is that of a Hovmöller diagram showing alternate descending layers of eastward and westward winds such as in Figure 11. In this figure repeated transitions in monthly averaged zonal mean zonal winds at the equator appear to originate in the upper stratosphere and then propagate downward, at least until 2016. These repeating irregular cycles have, in fact, been continuously observed since the 1950s before being unexpectedly interrupted in 2016 (Newman et al., 2016; Osprey et al., 2016). A second interruption occurred in 2019/20 (Anstey et al., 2021) in the zonal wind above Singapore (Anstey et al., 2021, Figure 1a), though the occurrence of this interruption is not so apparent in the zonal mean zonal wind shown in Figure 11.
In the years following its discovery various dynamical mechanisms were proposed for the QBO but it was difficult to explain key aspects, such as the QBO’s downward propagation from 10 to 40 hPa without loss of amplitude and the super rotation of the atmosphere at the equator during the eastward phase, without vertical transport of momentum by equatorial waves. Lindzen and Holton (1968) were the first to consider such processes. They argued that with a broad spectrum of gravity waves propagating up from the troposphere there would be eastward acceleration of the mean flow in regions of eastward vertical shear from the waves with eastward phase speeds, and westward acceleration in regions of westward vertical shear from the waves with westward phase speeds. Moreover since selective filtering by the mean flow meant that only waves with westward phase speeds could propagate through strong eastward winds, and vice-versa for waves with eastward phase speeds, the shear zones would descend over time. This proposed mechanism was corroborated with results from idealized numerical experiments (Lindzen and Holton, 1968).

Instead of gravity waves Holton and Lindzen (1972) considered vertical transport of momentum by thermally and mechanically damped upward propagating planetary scale Kelvin and Rossby-gravity waves and also obtained a realistic QBO in an idealized numerical model. However subsequent studies suggest that the observed large scale waves are unable to fully explain the observed QBO accelerations (Dunkerton, 1997). Most likely a combination of equatorial Kelvin and Rossby-gravity waves and small-scale gravity waves cause the alternating eastward and westward accelerations, though for the shear zones to descend the wave forcings (accelerations) have to exceed the effects of mean upward advection by the residual mean meridional or diabatic circulation (Section 2.2), especially in the lower stratosphere (Match and Fueglistaler, 2020).

The two-way interaction between the mean flow and vertical fluxes of wave-momentum is now well established as the fundamental process leading to the QBO but major challenges remain in determining the details of the waves involved. Small scale gravity waves are, in particular, difficult to observe (Hertzog, 2020) and hence the precise partitioning of the various wave-types in the forcing of the QBO is uncertain. Estimating the wave-momentum flux from the available observations is also problematic as it relies on numerous assumptions (Vincent and Alexander, 2020). In addition the relative importance of the different wave dissipation mechanisms (e.g., damping or critical layer wave-breaking that occurs when the phase speed is close to the wind speed) is poorly understood (Anstey et al., 2022b).
Strictly the QBO is not a wave, but a sequence of alternating eastward and westward wind regimes propagating downward. Therefore the cycles are not pure sinusoids and display variability from cycle to cycle (Pascoe et al., 2005). Many cycles show the descent of the region of westward vertical shear apparently stalling around 30 hPa followed by the eastward winds persisting longer in the lower stratosphere (Figure 11). Some of the variability most likely results from variations in the sources (mostly convection) of the tropical waves (Schirber, 2015), which are not well understood. Another possibility suggested by Dunkerton (1983) was laterally propagating Rossby waves from the NH winter could modulate the strength of the QBO during its eastward phase, though evidence from idealized simulations (O’Sullivan, 1997) did not support this. Hence, prior to 2016, the consensus was the laterally propagating Rossby waves from the extratropics had only a small effect (e.g., Shuckburgh et al., 2001; Hamilton et al., 2004) on the QBO. Since 2016 horizontal wave-momentum transport from the winter hemisphere has been implicated (Anstey et al., 2021) with initiating interruptions to the QBO cycling in 2016 and 2019 and this has prompted renewed interest in the role of the extratropics in modulating tropical stratospheric variability (see Section 4.2). A tendency for the phases of the QBO and annual cycle to align is thought to arise from the annual cycle in the tropical upwelling in the equatorial lower stratosphere that opposes the QBO’s downward progression (Hampson and Haynes, 2004; Rajendran et al., 2018).

Due to thermal wind balance a corresponding QBO signal is also observed in the temperatures at the equator, with warm and cool anomalies where there is eastward and westward vertical wind shear, respectively (Baldwin et al., 2001). Maintaining the balance requires a secondary residual meridional circulation with ascent at the equator in the regions of westward vertical shear and descent in regions of eastward vertical shear (Figure 12). Plumb and Bell (1982) noted that vertical advection associated with this secondary circulation could account for some of the observed phase asymmetries by, for instance, retarding the downward propagation of the westward shear zone (cf., the stalling noted above) and speeding up the descent of the eastward shear zone. The secondary circulation also acts to concentrate the strongest eastward accelerations close to equator (Dunkerton, 1991) which again is consistent with the observed meridionally narrower eastward phase compared to westward phase (Coy et al., 2016; Pahlavan et al., 2021a).

Vertical advection by the secondary circulation along with the QBO temperature anomalies are the primary reason for a similar period oscillation in ozone, which was first seen in subtropical data by Funk and Garnham (1962). At the equator, transport (advection) is the dominant process in the lower stratosphere and the ozone and wind QBO anomalies are generally in-phase. Above about 15 hPa the effects of the temperature dependent ozone-chemistry becomes more relevant and the ozone anomalies change sign (Chipperfield et al., 1993). Feedbacks from these ozone anomalies, in turn, modulate the dynamical variability. For example, Butchart et al. (2003) found that in model simulations ozone related diabatic feedbacks cause a ~10% lengthening of the QBO period and a stronger temperature signal in the lower stratosphere (Butchart et al., 2003) and zonal wind (Tian et al., 2006) in the lower and middle stratosphere. However, there is uncertainty due to the precise role of chemical processes in the ozone QBO and there is uncertainty in the strength of the feedbacks and this remains an area of active research (Zhang et al., 2021).
Figure 12. Schematic showing the secondary residual meridional circulation (gray arrows) associated with the QBO equatorial temperature anomalies and the asymmetry in the circulation between the summer winter hemispheres (Peña Ortiz et al., 2008). The locations of the warm and cool anomalies are shown in red and blue respectively with the font size giving a qualitative guide to strength of the anomalies. Solid (dashed) contours indicate eastward (westward) zonal mean zonal winds at intervals of 10 ms$^{-1}$ starting at $\pm 5$ ms$^{-1}$ for (a) westward shear zone and (b) eastward shear zone. Adapted from Figure 1 of Plumb and Bell (1982).

4 Tropical-extratropical coupling

One process providing a coupling between the tropics and extratropics in the stratosphere is the residual meridional circulation (Equation 1). The extratropical planetary waves drive the diabatic circulation (Section 2.2) which is the principal reason for an annual cycle in temperature at the tropical tropopause. Likewise the vertical component of the QBO’s secondary meridional circulation causes adiabatic warming and cooling in the subtropical stratosphere and thereby forces a similar period oscillation in the temperatures in the region extending out to $\sim 45^\circ$ latitude (Randel et al., 1999). In the subtropics the secondary
Figure 13. Holton-Tan effect for January in the NH and November in the SH. Latitude-height cross sections of the monthly and zonal mean zonal wind averaged over all years when the winds at the equator are $\geq 2.5 \text{ ms}^{-1}$ minus the average over all years when the winds at the equator are $\leq -2.5 \text{ ms}^{-1}$ at 50 and 30 hPa in January and November, respectively. The contour interval is 2 ms$^{-1}$ for values between -20 and +20, otherwise the interval is 10 ms$^{-1}$. Negative values are shown by dotted contours and shading indicates where the differences are statistically significant at the 95% level according to Student's $t$-test. Based on MERRA2 the same ERA5.1 reanalysis data (Gelaro et al., 2017) (Hersbach et al., 2020) for the 1980–2016 1979–2020 as used in Figure 2.

Other mechanisms through which the tropical variability, and in particular the QBO, has a remote impact on the extratropical stratosphere are less well understood (Anstey and Shepherd, 2014). These impacts also extend down into the troposphere (Section 6.3) and upward into the mesosphere and are commonly referred to as “QBO teleconnections” (Anstey et al., 2022b).

4.1 QBO teleconnections

Angell and Korshover (1964) provided the first evidence of the QBO possibly influencing the high latitude stratosphere when they reported a QBO signal in polar temperatures (and ozone) in both hemispheres. With more years of observations (1962-1977) Holton and Tan (1980) found that the strength of the NH winter polar vortex correlated well with the phase of the equatorial QBO: a warmer, weaker vortex coincided with westward equatorial winds at 50 hPa whereas when the winds were eastward the vortex was colder and stronger. A corollary of this so-called “Holton-Tan (HT) effect” first reported by Labitzke (1982) is that the occurrence of an SSW is more likely when the QBO winds are westward at 50 hPa than when the winds are eastward (Baldwin et al., 2021). Holton and Tan (1980) also noted a QBO signal in the SH zonal wind, this time in spring. An illustration of the HT-effect in both hemispheres is presented in Figure 13. Comparing this figure with the inter-annual standard deviation of the monthly mean zonal winds (Figure 7) highlights the importance of QBO teleconnections in modulating the variability of the polar vortex.
A full understanding of the underlying mechanisms for the QBO polar vortex teleconnection remains elusive, despite numerous investigations into the HT-effect following its discovery (Anstey and Shepherd, 2014). Holton and Tan (1980) noted that the zero zonal wind line at the edge of the tropical QBO was typically further poleward during the westward phase than in the eastward phase and argued that this contributed to confining the extratropical planetary waves more toward high latitudes, thereby leading to the warmer, weaker vortex. This is often referred to as the “Holton-Tan mechanism.” In contrast, Garfinkel et al. (2012) suggested that enhanced polar refraction of planetary waves results from the QBO’s secondary meridional circulation modifying the Rossby wave refractive index in the subtropics. As the extratropical planetary wave activity is further modulated by the strength of the vortex (Section 2.4) this has, so far, hampered attempts to isolate which of these two proposed mechanisms for the HT-effect is more relevant (Anstey et al., 2022b). More recently, Yamazaki et al. (2020) have suggested an alternative tropospheric pathway for the Boreal winter in which a convection anomaly resulting from the QBO effect on the tropical troposphere (see Section 6.2) generates a Rossby wave train into the mid-latitude troposphere which, in turn, influences the upward propagation of planetary waves into the polar vortex.

Model studies suggest that, in general, the HT-effect is sensitive to the state of the polar vortex (Anstey and Shepherd, 2014) though in some models QBO-teleconnections are also effected by circulation biases (Karpechko et al., 2021). A dependency on the vortex strength is one possible reason for the observed intra-seasonal variations and inter-hemisphere differences in extratropical QBO response: in the NH the HT-effect is most robust in December and January (Zhang et al., 2019) while in the SH the winter polar vortex is much stronger (Section 2.1) and the response is more distinctive around the time of the final warming (Baldwin and Dunkerton, 1998). Interactions found between the QBO stratospheric teleconnections and the solar cycle (Naito and Hirota, 1997) and El Niño Southern Oscillation (ENSO) (Garfinkel and Hartmann, 2007) makes the causes of these differences uncertain as both the QBO and polar vortex variability are, separately, impacted by these longer timescale modes of variability. The QBO-altitudes which exert the strongest influence on the extratropics is another uncertain aspect (Anstey et al., 2022b).

4.2 Interruptions to the QBO

In February 2016 a shallow layer of westward winds began to form at the equator within the decaying eastward phase of the QBO (Newman et al., 2016; Osprey et al., 2016). At the time this was a complete surprise and over the next few months the usual oscillation in the equatorial winds appeared to cease (Figure 11), raising concerns among leading researchers of the QBO entering a “death spiral” (Dunkerton, 2016) and disappearing altogether. We now know that this did not happen and by late 2016 the regular cycling of the equatorial winds had recommenced (Figure 11), albeit with the phase progression shifted from what would be expected if the interruption had not occurred (Anstey et al., 2021).

Compared to previously observed QBO cycles the anomalous behavior in 2016 was unprecedented (Newman et al., 2016; Osprey et al., 2016). It could not be accounted for by the established canonical model of the QBO, which relies only on the vertical transport of momentum (Section 3.1). Instead, Osprey et al. (2016) and Coy et al. (2017) attributed the anomalous westward acceleration near 40 hPa to enhanced horizontal EP-flux, \( F(\phi) \) (Equations 3), entering the tropics from the NH. Arguably this was the first observational evidence of the extratropical stratosphere directly influencing the variability of the
tropical stratosphere, as opposed to indirectly though the meridional circulation. Further evidence for the direct influence was obtained when the QBO was surprisingly interrupted again in December 2019, though this time the horizontally propagating Rossby waves came from the SH (Anstey et al., 2021).

A dynamical explanation for the interruptions is still in its infancy. While several studies have examined the role of different wave types and found that tropical mixed Rossby-gravity waves have an important role in the anomalous westward accelerations (Lin et al., 2019; Kang et al., 2020; Kang and Chun, 2021), others (Hitchcock et al., 2018) have stressed the importance of feedback mechanisms. Nonetheless, all conclude that during the disrupted cycles there was a significant extratropical influence on the variability in the tropical stratosphere. One consequence of this direct influence by the extratropical waves is that the paradigm in which the tropical zonal winds retain memory from year to year [the low-latitude “fly-wheel” effect (Scott and Haynes, 1998)], and thereby provide a mechanism for inter-annual variability, is possibly less appropriate than previously thought and may need to be reconsidered once a full understanding of the QBO interruptions is obtained.

5 Models and predictability

5.1 Stratosphere resolving models

First attempts to investigate stratospheric dynamics in a general circulation model (GCM) were by Manabe and Hunt (1968), though the model they used was only hemispheric with a simulation of less than one year. In general, early GCMs typically neglected the stratosphere based simply on its small fraction (~17%) of the atmospheric mass and need to optimize use of available computing resources. When the stratosphere did feature in GCMs (e.g., Kasahara and Sasamori, 1974; Fels et al., 1980) long simulations of the variability were not feasible due, again, to computational constraints. Computers used for modelling the atmosphere evolved into supercomputers around 1980 and a number of models subsequently preformed multi-year simulations of the stratosphere (Rind et al., 1988; Boville, 1995; Hamilton, 1995; Manzini and Bengtsson, 1996; Butchart and Austin, 1998). In these simulation the stratosphere was typically closer to radiative equilibrium than observed (cf., Section 2.1) with significant cold biases in the high latitudes, especially during the SH winter and spring (the so-called “cold-pole problem”), and the QBO was virtually absent from the tropical stratosphere (Pawson et al., 2000).

Biases in the simulated stratospheres are critical as starting with Boville (1984) it has become clear that the simulated tropospheric circulation is more sensitive to the stratosphere than would be expected from the stratosphere’s relatively small mass (see also Section 6.3). Likewise, when Farman et al. (1985) discovered a hole in the ozone layer over Antarctica, the first versions of the models developed to include chemistry (chemistry-climates models or CCMs) for assessing ozone changes were limited by the stratospheric temperature biases (Austin et al., 2003). Improvements in the simulated stratosphere were also found desirable for numerical weather-prediction models in order to better assimilate satellite radiance measurements that are often significantly weighted in the stratosphere (English et al., 2000; Polavarapu et al., 2005).

---

In the 1960s the computing community was already using the term “supercomputer” but it only came into common usage in atmospheric science with the introduction of vector machines capable of around 100 Mflops.
The cold pole problem resulted from a diabatic circulation (Section 2.1) that was too weak due to deficiencies in the wave drag ascribed to the absence of gravity waves (McLandress, 1998). Increasing resolution was found to partially alleviate the problem (Hamilton et al., 1999), though more significant was the implementation of parametrizations of non-orographic gravity waves (NOGWs) to supply the missing forcing from sub-grid scale waves (Manzini et al., 1997; Scaife et al., 2002). Despite their simplicity, and not being well constrained by observations (Plougonven et al., 2020), such schemes enabled models to simulate well the polar night jet in both hemispheres though there is still work to be done on improving the variability of the jets (Butchart et al., 2011; Ayarzagüena et al., 2020; Hall et al., 2021b). For example, while the latest state-of-the-art climate models simulate occurrences of SSWs reasonably well (Baldwin et al., 2021), the models are persistently biased towards an under representation of split compared to displaced vortex events (Hall et al., 2021b).

When Scaife et al. (2000) first tested NOGW parametrizations in their GCM they discovered that a QBO-like oscillation of descending eastward and westward winds appeared in the tropical stratosphere. Prior to this the QBO was generally absent from GCMs or very weak (Cariolle et al., 1993). An oscillation of realistic amplitude had been obtained by Takahashi (1996) without parametrized NOGW-drag but only by using fine vertical resolution (∼ 500 m) and substantially reduced sub-grid scale diffusion. Spontaneous QBO-like phenomena were also obtained in other simplified GCMs without a NOGW parametrization by increasing vertical resolution (Horinouchi and Yoden, 1998; Hamilton et al., 1999). Nonetheless, today, climate and seasonal prediction models that are able to simulate a QBO nearly always do so by implementing parametrizations of NOGWs (Butchart et al., 2018; Bushell et al., 2022; Richter et al., 2022; Stockdale et al., 2022). The number of such models has grown substantially since the pioneering work of Scaife et al. (2002) but, so far, the quality of the simulated QBOs has, on average, not improved much (Richter et al., 2020).

One characteristic of the QBO that is generally well simulated is the mean period (Figure 14a) which is typically accomplished by tuning the (weakly constrained) parametrized NOGW-drag (Bushell et al., 2022; Garfinkel et al., 2022). More difficult to achieve is the correct vertical structure with models persistently under representing the strength of the oscillation in the lower stratosphere (Richter et al., 2020; Bushell et al., 2022), despite being able to simulate, on average, realistic amplitudes in the middle stratosphere (cf., Figures 14b and d). There are also problems in reproducing the correct phase asymmetry (Schenzinger et al., 2017). On average the eastward phase is relatively too strong in models (Figure 14c), while in initialized simulations the models struggle to maintain the strength of the westward phase (Stockdale et al., 2022). In general the models under represent cycle-to-cycle variability (Bushell et al., 2022).

Problems in simulating the QBO are largely attributed to uncertainties in parametrizing NOGWs (Bushell et al., 2022). Improving the parametrizations is an area of active research and recent developments have included more physically based representations of the source of NOGWs (Richter et al., 2010; Lott and Guez, 2013; Bushell et al., 2015). Resolved equatorial waves also contribute to the QBO forcing, particularly Kelvin waves during the eastward phase (Pahlavan et al., 2021b). Again there is uncertainty (spread) among models in the resolved waves (Kelvin and mixed Rossby-gravity wave) linked to the convective sources as well as resolution and mean wind biases, with a particular affinity between the resolved wave forcing and model vertical resolution (Holt et al., 2022). Reducing the uncertainty in the wave forcing (parametrized and resolved) is expected to lead to improvements in the simulations of the QBO (Anstey et al., 2022b). In turn, this will impact on the
Figure 14. (a) QBO periods, (b) 10-hPa QBO amplitude, (c) 10-hPa ratio of westerly (eastward) to easterly (westward) QBO amplitude, and (d) 50-hPa QBO amplitude for models that participated in Phases 5 (leftmost five bars) and 6 (middle 15 bars) of the Coupled Model Intercomparison Project (CMIP) (Eyring et al., 2016; Taylor et al., 2012). CMIP5 and CMIP6 ensemble means are compared on the right with the corresponding ERA-Interim (Dee et al., 2011) results. The triangles in (a) indicate the range of periods. Periods and amplitude obtained using different diagnostics are shown by the + symbol in (a) and (b) to confirm robustness of the conclusions. Figure 2 from Richter et al. (2020).

representation of QBO extratropical teleconnections (Section 4.1) in the models which are known to be sensitive to biases in the QBO (Anstey et al., 2022c) as well as the mean stratospheric circulation (Karpechko et al., 2021).
5.2 Predictability of the stratosphere

Predicting stratospheric variations can be done either empirically using statistical methods or, as considered here, dynamically using numerical models. Miyakoda et al. (1970) pioneered attempts to develop numerical predictions of the stratosphere but their 14-day GCM simulation for March 1965 was not successful at capturing the sudden warming that occurred. The February 1979 SSW was the first to be successfully simulated, and also the first observed from space (Palmer, 1981a, b) using newly introduced operational temperature soundings of the stratosphere (Miller et al., 1980). The successful simulation of the 1979 SSW by Butchart et al. (1982) was initialized five days prior to the peak of the warming but relied on prescribed lower boundary conditions at the tropopause. In subsequent numerical forecasts Simmons and Strüfing (1983) and Mechoso et al. (1985) found this SSW was predictable up to ten and five days ahead, respectively.

With more SSWs and concomitant case studies (Mukougawa and Hirooka, 2004; Allen et al., 2006; Tripathi et al., 2016; Rao et al., 2018) the current consensus is that SSWs can, in general, be deterministically predicted one to two weeks in advance (Domeisen et al., 2020b), with higher skill more likely from models with enhanced representation of the stratosphere (Allen et al., 2006; Marshall and Scaife, 2010; Song et al., 2020). For a single model predictive skill varies between SSWs (Taguchi, 2016; Karpechko, 2018; Rao et al., 2019) due to differences in the phenomenology and generation mechanisms for vortex-splitting and vortex-displacing events (Section 2.5). These differences contribute to determining if the predictability arises mostly from the stratosphere or from the troposphere (Sun et al., 2012). Predictability arising from the troposphere is sensitive to model biases affecting planetary waves and depends on their propagation timescales (Noguchi et al., 2016), whereas predictability arising from the stratosphere is more likely to depend on the background stratospheric state prior to an SSW (de la Cámara et al., 2017). In general, vortex-splitting appears more difficult to forecast than a vortex displacement (Taguchi, 2018a; Song et al., 2020).

Compared to the troposphere the extratropical stratosphere exhibits extended predictability (Waugh et al., 1998; Lahoz, 1999), though for extreme vortex events such as an SSW or, alternatively a polar vortex intensification, the predictability is still limited by initial condition uncertainty and model errors (Tripathi et al., 2015). Among forecast systems there is, nonetheless, an affinity between higher skill in the stratosphere and higher skill in the troposphere (Domeisen et al., 2020b), and most systems consistently predict polar vortex intensification and final warmings better than SSWs (Ichimaru et al., 2016; Domeisen et al., 2020b). For the NH winter/spring stratosphere deterministic forecast skill is generally limited to a range of one to two weeks (Domeisen et al., 2020b). In the SH perturbations to the polar vortex in winter persist into spring due to the reduced intraseasonal variability (Figure 8) resulting from the weaker planetary wave forcing (Byrne and Shepherd, 2018). Hence, the SH stratosphere is potentially predictable on longer timescales, as is evident in the significant skill found by Seviour et al. (2014) in the prediction of the strength of the Antarctic stratospheric polar vortex in spring at one month average lead times (Figure 15).

During the Boreal winter the extratropical stratosphere has also been found to be predictable in a probabilistic sense on seasonal timescales well beyond the deterministic range of forecast skill (Scaife et al., 2016; Taguchi, 2018b; Portal et al., 2022). Scaife et al. (2016) noted that in seasonal forecasts capable of reproducing the overall climatological frequencies of SSWs
and strong polar vortex (SPV) events well, the proportion of ensemble members predicting the occurrence of an SSW or SPV event varied from year to year (Figure 16). For winters in which an SSW (SPV) was “observed” (based on proxy observations derived from the forecast ensemble — see Scaife et al. (2016) for details) the proportion of ensemble members forecasting an SSW (SPV) increased, on average, by 12% indicating potential probabilistic forecast skill on a seasonal timescale with significance beyond the 95% level (Scaife et al., 2016). Most likely contributors to this probabilistic skill are the ENSO and the QBO though, on average, the latest state-of-the-art seasonal forecast systems overestimate the anomalous wave forcing of the polar vortex due to ENSO while underestimating the modulation of the strength of the vortex by the QBO (Portal et al., 2022). At least for a hindcast period when the observed interannual variability of the polar vortex was significantly affected by the QBO (Portal et al., 2022).

Prediction skill of the QBO itself is very high in the lower to middle stratosphere on sub-seasonal to seasonal timescales (Lim et al., 2019b; Coy et al., 2022). Out to one year high skill is still obtained for predictions of the phase progression (Coy et al., 2020) though initialized forecasts are generally less successful at maintaining adequate QBO amplitudes (Stockdale et al., 2022). This is largely attributed to weaknesses in all models in predicting the development of a sufficiently strong westward phase which, in turn, can be linked to inadequacies in overall representation of the QBO in the models (Coy et al., 2022; Stockdale et al., 2022). Beyond 12 months decadal forecast systems have been found to exhibit predictive skill up to four years for the QBO phase (Pohlmann et al., 2013; Scaife et al., 2014) with potential for extending this further by improving the representation and prediction of ozone variability in the models (Pohlmann et al., 2019). Enhanced ENSO prediction skill also has potential for extending lead times for skillful QBO predictions as strong ENSO events have been found to lock the QBO phase across ensemble members in uninitialized simulations (Christiansen et al., 2016; Serva et al., 2020).

Exploiting the skillful multi-year QBO predictions is currently frustrated by deficiencies in the amplitudes and vertical structure of the predicted QBOs (Pohlmann et al., 2019). Multi-model results show that similar deficiencies are already present on the seasonal-to-annual timescale and almost certainly contribute to a lack of skill in predicting the QBO’s extratropical winter teleconnections beyond the first month (Stockdale et al., 2022). The two recent interruptions to the QBO (Section 4.2) add further uncertainty to the robustness of the QBO as a source of predictability as the predictable signal was lost during each

**Figure 15.** September-to-November mean anomalies in the zonal mean zonal wind at 10 hPa and 60°S in ERA-Interim (solid curve) and hindcasts (black dots) initialized near 1 August. The dashed curve is the ensemble mean with the interannual correlation with the reanalysis of 0.73. Adapted from Figure 2 of Seviour et al. (2014).
Figure 16. Ensemble forecast evolution of zonal wind in the stratosphere. Zonal mean winds are shown (10 hPa, 60°N) in two different winters, with 24 ensemble members per winter. The horizontal lines show the threshold for SSW (0 ms$^{-1}$) events and SPV (48 ms$^{-1}$) events. The two winters shown exhibit quite different predicted probability of a SSW or SPV event, indicating the potential for forecasting the risk of these events well beyond the deterministic range of a few weeks. Adapted from Figure 1 of Scaife et al. (2016).

interruption before reemerging a few months later but with a significant shift in phase from what would have been expected (Anstey et al., 2021). Neither interruption was predicted by operational seasonal forecasts (Osprey et al., 2016; Anstey et al., 2021) though for the 2016 event Watanabe et al. (2018) found skill at lead times of around one month. Predictability on this timescale is consistent with the events been initiated from the extratropics (Section 4.2) but much reduced from the timescales associated with the QBO’s uninterrupted quasi-regular cycling.

6 Influence on the troposphere

6.1 Middle and high latitudes

When Scherhag (1952) first discovered SSWs he also presented tentative evidence that their effects reach the surface. Similarly, over a decade after co-discovering the QBO Ebdon (1975) found evidence suggesting that it features in determining the character of the NH tropospheric circulation. The earliest model results linking tropospheric circulation changes to variations
in the stratospheric polar vortex are those published by Boville (1984), with concurrent theoretical advances (Hoskins et al., 1985; Haynes and Shepherd, 1989; Haynes et al., 1991) further supporting the possibility of a downward influence. Nonetheless, up until the end of the twentieth century, the prevalent view was that the dynamical coupling between the troposphere and stratosphere was mostly one-way with the stratosphere responding to forcing from below.

The paradigm shift to a two-way coupling arose from analysis of the Northern and Southern “Annular Modes” (NAM and SAM, respectively) (Thompson and Wallace, 2000). These are the leading patterns of the large-scale variability of the extratropical circulation and dominate variability on intraseasonal to interdecadal timescales. A manifestation of the annular modes in the troposphere is a fluctuation in the latitudinal position of the jet stream, while in the stratosphere the characteristic feature is a fluctuation in the strength of the polar vortex (Kidston et al., 2015). Annular modes in the stratosphere and troposphere appear to be coupled with occurrences of a weaker polar vortex correlated with equatorward shifts of the jet stream and vice versa for occurrences of a stronger polar vortex (Kidston et al., 2015). The discovery that NAM anomalies propagate downward from the stratosphere all the way to the Earth’s surface is most often credited to Baldwin and Dunkerton (1999), though other studies had previously noted stratospheric circulation anomalies extending down into the upper troposphere (Boville, 1984; Kodera et al., 1990; Kuroda and Kodera, 1998). Downward propagation of SAM anomalies from the stratosphere to the surface was first reported by Thompson and Solomon (2002) in association with SH climate change.

Persistence times for annular mode anomalies are longest in the stratosphere but in months when the polar vortex is most active (January-February in the NH and November-December in the SH — see Section 2.4) and strongly coupled to the troposphere the timescales in the troposphere lengthen (Kidston et al., 2015; Gerber and Martineau, 2018). In these months anomalous values of the annular modes in the stratosphere are often followed by anomalies of the same sign in the troposphere which sometimes persist for up to two months in the NH (e.g., Figure 17), and even longer in the SH (Thompson et al., 2005). A slight delay before the occurrence of the concomitant tropospheric anomalies (Figure 17) suggests the possibility that the anomalies are propagating downward (Baldwin and Dunkerton, 2001) and, indeed, model studies have demonstrated that there is a genuine physical downward influence with the stratosphere affecting behaviour in the troposphere and at the surface (Kidston et al., 2015, and ref. therein) as reviewed by Kidston et al. (2015).

On the other hand, annular modes derived from EOFs (Thompson and Wallace, 2000; Baldwin and Dunkerton, 2001; Thompson et al., 2005) are not necessarily physical modes of variability (Sheshadri et al., 2018) and formulating a dynamical theory for the downward propagation of anomalies is still a work in progress. Given the comparatively light weight of the stratosphere its downward influence almost certainly results from a sensitivity of tropospheric dynamics to changes in the stratosphere with eddy feedbacks (Song and Robinson, 2004; Smy and Scott, 2009; Hitchcock and Simpson, 2014, 2016; Runde et al., 2016; Rupp and Birner, 2021) fundamental for enhancing the tropospheric response and contributing to its persistence. Non-local An alternative perspective for the stratosphere’s downward effect on the Atlantic jet, is the strong response results from a regime shift between three preferred jet positions, rather than a simple latitudinal movement in jet position (Maycock et al., 2020; Goss et al., 2021). Proposed mechanisms for communicating the dynamical effects from the
Figure 17. “Dripping paint diagram.” Composites of a non-dimensional NAM index (Baldwin and Dunkerton, 2001) showing decent of NAM anomalies from the stratosphere into the troposphere for (A) 18 weak vortex events (e.g., SSWs) and (B) 30 strong vortex events. The approximate location of the extratropical tropopause is indicated by the horizontal line. In (A) the weak events corresponding to negative values of the index (yellow and red shading) are determined by the dates on which the 10-hPa annular mode values cross –3.0. In (B) the strong events corresponding to positive values of the index (blue shading) are determined by the dates on which the 10-hPa index crosses +1.5. The shading is at intervals of 0.25 and the white contours at intervals of 0.5. Adapted from Figure 2 of Baldwin and Dunkerton (2001).

Stratosphere to the troposphere include non-local PV inversion, (Charlton et al., 2005), the downward effect of an induced secondary meridional circulation, (Song and Robinson, 2004), and the downward propagation of information by planetary waves against the prevailing upward wave flux (Perlwitz and Harnik, 2004) have each been considered as potential mechanisms for communicating the dynamical effects from the stratosphere to the troposphere. However, a single robust dynamical mechanism has yet to be identified that explains the effects resulting from upward propagating waves either being reflected back downward or changing the refractive properties of the lower stratosphere through wave-mean flow interactions (Perlwitz and Harnik, 2004). However, none of these proposed dynamical mechanism can account for the all aspects of stratosphere’s downward influence.

One complicating factor is the anomalous stratospheric events are quite diverse and do not always extend downward into the troposphere or, indeed, to the surface (Karpechko et al., 2017). In the NH the diversity is largely associated with the different types of SSWs (Section 2.5) with a surface impact only observed for about two thirds of the events (Domeisen, 2019; Hall et al., 2021a). Characteristics which determine whether an SSW has a downward influence remain elusive and it is still not possible to foresee at the onset of an SSW if it is likely to be followed by a discernible surface impact (Baldwin et al., 2021). When there is a downward influence the surface impacts of split events are seen, on average, about one week sooner than the
impacts of displacement events, at least for reanalysis data (Seviour et al., 2016; Hall et al., 2021a). On the other hand model results from Maycock and Hitchcock (2015) indicated no significant dependency on SSW type and show that there is probably insufficient evidence to establish a difference between impacts by split and displacement events.

Follow-through from studies of descending annular mode anomalies is the apparent affinity between extreme surface weather and the strength of the polar vortex (Kidston et al., 2015; Domeisen and Butler, 2020), and also the QBO phase due to its teleconnections to the polar vortex (Section 4.1). Extreme weather events involving a significant contribution from the stratosphere have been cataloged by Domeisen and Butler (2020) and they found them to be wide ranging with the stratospheric connection more firmly established for some than others. The connection in the NH is most prominent and robust in the Atlantic basin (Charlton-Perez et al., 2018) with a weakening of the polar vortex, or SSW, often associated with cold weather over Europe and the eastern USA (King et al., 2019), and cold air outbreaks over the northeastern Atlantic (Afargan-Gerstman et al., 2020). In contrast an unusually strong Arctic polar vortex, as often occurs during the QBO eastward phase (Section 4.1), is associated with stronger surface winds and more intense cyclones with increased the risk of storms and extreme rainfall over north west Europe (Kidston et al., 2015).

A weakening of the SH polar vortex and negative SAM index in the lower stratosphere generally leads to hot dry extremes over Australia (Lim et al., 2019a, 2021), cooler wetter conditions over southern South America (Lim et al., 2018) and higher Antarctic surface temperatures, except for the Palmer Peninsula which is more likely to experience cold dry conditions (Thompson et al., 2005; Lim et al., 2018). A positive SAM index often brings significant cooling to Antarctica and much of Australia, warming to the Palmer Peninsula, and warm dry conditions to southern South America, New Zealand and Tasmania, while Australia and South Africa are more likely to experience anomalously wet conditions (Gillett et al., 2006; Garreaud, 2018).

### 6.2 Low latitudes

The influence of the stratosphere on the tropical troposphere was independently reviewed recently by Haynes et al. (2021) and, hence, only a selection of the more significant dynamical developments are covered here. First indications of a possible coupling between the stratosphere and tropical troposphere came when Gray (1984) noted a correlation between the frequency of Atlantic hurricanes and the phase of the QBO, though somewhat earlier Angell et al. (1969) had tentatively suggested that the QBO had a role in determining the frequency of Pacific typhoons. Later studies (Ho et al., 2009; Camargo and Sobel, 2010), however, questioned the robustness of such relationships and a convincing dynamical explanation is still wanting (Haynes et al., 2021; Hitchman et al., 2021).

Tropical dynamics differ from extratropical dynamics due to the smallness of the Coriolis parameter which implies that, unlike in the extratropics (Section 6.1), the vertical scales inferred from balance dynamics are generally too shallow to potentially explain a coupling between the stratosphere and troposphere. Instead, proposed mechanisms for a tropical pathway between the stratosphere and troposphere are commonly based on the paradigm that variability at the tropopause and in the lower stratosphere modulates behavior throughout the depth of the tropical troposphere (Haynes et al., 2021). While the variability in the tropical stratosphere is dominated by the QBO (Section 3), 3 km, or so, above the tropopause the descending
alternating eastward and westward wind regimes begin to weaken significantly (Figure 11) and do not penetrate into the troposphere. Driven by the induced secondary circulation, the concomitant temperature QBO (Section 3.1) extends slightly lower with an identifiable signal observed up to \(\sim 15^\circ\) either side the Equator near the tropopause, albeit small \(\sim 0.5\text{K}\) compared to the dominant annual cycle (Section 3). The resultant temperature and circulation anomalies at the tropopause and in the lower stratosphere exert a downward influence on tropical convection which in turn plays a major role in coupling the QBO to the tropical troposphere (Haynes et al., 2021). Spatial variations in convection partially explain why the QBO signal in the troposphere is no longer predominately zonally symmetric as it was in the stratosphere but, so far, none of the physical mechanisms proposed can account for all aspects of the large geographical variations in the QBO’s downward influence on the tropical troposphere (Haynes et al., 2021).

The QBO’s influence on tropical convection is the main reason for its involvement (Yoo and Son, 2016; Nishimoto and Yoden, 2017; Klotzbach et al., 2019; Martin et al., 2021) with the Madden-Julian oscillation (MJO), which dominates intraseasonal variability in the tropical troposphere (Zhang, 2005). During the NH winter up to 40\% of the inter-annual variations in the MJO amplitude are explained by a coupling to the QBO (Son et al., 2017), with the MJO stronger and more persistent, by up to 10 days, when the QBO is westward in the lower stratosphere (Yoo and Son, 2016; Martin et al., 2021). In contrast an eastward QBO phase in the lower stratosphere favours a weaker than normal MJO during NH winter (Yoo and Son, 2016).

Knowledge of the QBO-MJO relationship is only just emerging (Klotzbach et al., 2019) and there is still plenty to do in developing a working mechanism (Haynes et al., 2021; Martin et al., 2021). The relationship is, nonetheless, particularly relevant as it implies a potential pathway coupling the stratosphere to the extratropical troposphere (cf., Section 6.1) via the MJO’s teleconnections to high latitudes (Anstey et al., 2022b). The corresponding ramifications for the extratropics of this modulation of the MJO’s global teleconnections by the QBO are reviewed in Martin et al. (2021) and, therefore, this aspect of the QBO-MJO relationship is not considered here.

The tropical troposphere and MJO are also affected by SSWs (Kodera and Yamada, 2004; Kodera, 2006; Bal et al., 2017; Noguchi et al., 2020; Wang et al., 2020a; Yoshida and Mizuta, 2021). One proposed mechanism is the increased wave driving during an SSW strengthens the mean meridional (Brewer-Dobson) circulation and cools the tropical lower stratosphere and tropopause. The anomalies created then impact tropical convection (Kodera et al., 2011) in a similar way to the anomalies resulting from the QBO’s secondary circulation. Tropopause level PV intrusions from the subtropics during an SSW also affect the upper tropical troposphere (Albers et al., 2016) and represent an alternative subtropical pathway from the stratosphere to tropical troposphere (Haynes et al., 2021).

Ramifications of the stratosphere’s influence on the tropical troposphere include an enhancement of convective activity and precipitation over the tropical west Pacific and a suppression over the equatorial central and east Pacific when the QBO is westward in the lower stratosphere, at least for the annual mean (Haynes et al., 2021). During SSWs equatorial convection is enhanced south of the Equator and suppressed to the north, especially for split vortex events (Bal et al., 2017), while tropical cyclones are generally more intense (Yoshida and Mizuta, 2021). Monsoon intensity and rainfall appear to be sensitive to the QBO phase with, for example, the East Asian winter monsoon weaker, on average, when the QBO is westward compared to when it is eastward at 70 hPa (Ma et al., 2021). Also over East Asia MJO-related precipitation anomalies are between 40\%
and 70% larger during the QBO westward phase than during the eastward phase, depending on the phase of the MJO (Kim et al., 2020).

6.3 Enhancing tropospheric predictions

The stratosphere’s powerful downward influences and longer prediction timescales (Section 5.2) are reasons why the stratosphere matters for tropospheric predictions (Scaife et al., 2021). Models with better resolved stratospheres generally produce superior simulations of the surface climate (Hardiman et al., 2012; Kawatani et al., 2019) and more skillful forecasts on sub-seasonal to seasonal timescales (e.g., Figure 18) (Roff et al., 2011; Charron et al., 2012; Butler et al., 2016; Marshall et al., 2017; Domeisen, Roff et al., 2011; Charron et al., 2012; Butler et al., 2016; Marshall et al., 2017; Domeisen et al., 2020a), with more accurate stratospheric initial conditions (Jia et al., 2017) and variability (Tripathi et al., 2015) further enhancing the forecast skill. In the extratropics predictive skill is enhanced (Sigmond et al., 2013) following SSWs and, in April, skill can be gained from a knowledge of the vertical structure of the final warming (Hardiman et al., 2011; Butler et al., 2019). In the tropics MJO predictability is influenced by the QBO (Marshall et al., 2017; Lim et al., 2019b; Wang et al., 2019).

Figure 18. Average improvement in predictive skill for geopotential height anomalies in the SH poleward of 60°S in ensembles of 30 day forecasts from mid-November for the years 1979–2008 when the vertical resolution in the forecast model is increased above 100 hPa (Roff et al., 2011). In the troposphere improvements of ~ 5% are seen 3–4 weeks into the forecasts. Note the non-linear scale for the shading. Figure 2a from Roff et al. (2011).

Another reason why the stratosphere matters for extratropical predictions is the pathways it provides to potential sources of extended range skill lying outside the stratosphere such as ENSO (Cagnazzo and Manzini, 2009; Ineson and Scaife, 2009; Butler et al., 2014), Eurasian snow cover (Hardiman et al., 2008; Garfinkel et al., 2020a), and Arctic sea ice (Kim et al.,
2014; Zhang et al., 2018; Kretschmer et al., 2020). Multi-year skill arising from the solar cycle (Gray et al., 2013; Dunstone et al., 2016) and volcanic eruptions (Bittner et al., 2016) also involves the stratosphere though the possible mechanisms are not considered here. A concurrent independent perspective of these aspects of the stratosphere’s role in long range prediction can, instead, be found in Scaife et al. (2021).

7 Outstanding issues

Although the observed dynamical variability of the stratosphere is reasonably well understood in terms wave mean flow theory the observations themselves continue to surprise researchers just as the original discoveries of SSWs and the QBO had surprised meteorologists of the time. In 2002 a SSW was recorded for the first time in the SH (Baldwin et al., 2021) and in 2016, and again in 2019, the QBO’s near repeatable cycles of eastward and westward winds that had been observed continuously since its discovery were interrupted (Newman et al., 2016; Osprey et al., 2016; Anstey et al., 2021). These interruptions prompted a reassessment of the extent to which the extratropical stratosphere could influence the tropics (Section 4.2) though the efficacy of the underlying canonical model remained strong. On the other hand, the interruptions, and the occurrence of the SH SSW, suggest the observational record may not be long enough (e.g., just 27 QBO cycles prior to 2016) to properly quantify the full range of dynamical variability in the stratosphere. Consequently it is not possible to determine, with confidence, if such rare or “extreme” events are part of the natural dynamical variability or a sign of human activity affecting the stratosphere (Wang et al., 2020b; Anstey et al., 2021; Jucker et al., 2021).

Internal multi-year variability of the polar vortex seen in climate models (Butchart et al., 2000) is another outstanding issue that can only be fully addressed with more years of observations. Simple models have demonstrated that there are potential internal dynamical mechanisms that can drive low frequency variability of the polar vortex, even in the absence of other external forcings (Hatfield and Scott, 2019). In addition, circumstantial evidence for low frequency variability is seen in the decadal variations in the frequency of occurrences of SSWs deduced from surface observations (Domeisen, 2019). Whether or not this circumstantial evidence is an indication of internal stratospheric variability (Butchart et al., 2000) or is due to variations in stratosphere-troposphere coupling on inter-decadal time-scales (Manzini et al., 2012; Woo et al., 2015; Omrani et al., 2016) remains an open question.

As yet our understanding of stratospheric variability and its downward influence is not deep enough to know with confidence which aspects and events will have a strong impact at the surface (Baldwin et al., 2021; Haynes et al., 2021). Likewise the coupling between the tropical and extratropical stratosphere is not fully understood (Anstey and Shepherd, 2014). Current models struggle in representing these teleconnections (Elsbury et al., 2021; Anstey et al., 2022c) and also teleconnections to the troposphere. A more in-depth understanding of the mechanisms will therefore be useful for identifying which aspects of the models require further development to enable them to more accurately predict those surface events and extremes that have an affinity with the stratospheric variability.

Another outstanding issue is how the stratospheric variability will respond to climate change and what feedbacks there will be on the tropospheric climate. In the extratropics there is currently large uncertainty over the projected response due to
the near balance between opposing effects of increased CO₂ radiative cooling and increased adiabatic heating from a faster Brewer-Dobson circulation (Ayarzagüena et al., 2020). On the other hand, in the tropics a projected weakening of the QBO appears robust (Kawatani and Hamilton, 2013; Butchart et al., 2020; Richter et al., 2022) though, again, there is large uncertainty over the response of the QBO period (Richter et al., 2022). The uncertainties in the QBO response are attributed to inadequacies in the representation (parameterization) of small scale gravity waves within models due to the lack of observational constraints (Richter et al., 2022). It has also been suggested that the rudimentary representation of small scale gravity waves in models contributes to the uncertainties in the extratropical response (Kidston et al., 2015). Reducing these uncertainties is important since, through the stratosphere’s downward influence they can contribute significantly to uncertainties in projections for regional climate (Kidston et al., 2015; Simpson et al., 2018) and extremes (Domeisen and Butler, 2020).

Finally, for over fifty years it has been recognized that the stratosphere is extremely vulnerable to the effects of human activity (Tuck, 2021), with the well-being of the ozone-layer of particular concern. The Montreal Protocol was agreed to safeguard the ozone-layer but a range of human activities such as those causing wild fires (Stocker et al., 2021), speculative geoengineering proposals for mitigating climate change (Tilmes et al., 2018; Smith et al., 2022), space tourism and high altitude leisure flights by dangerously polluting rocket powered hobby aeroplanes (Larson et al., 2017; Ryan et al., 2022), are all likely to affect the variability in the stratosphere and hence surface weather and climate. Potentially this could increase the risk of disastrous extremes and hence there is the need for continued research and monitoring of the stratosphere and new international agreements to regulate human activities that affect stratospheric variability but have little or no societal benefits.

*Competing interests.* None

*Acknowledgements.* This work was supported by the Met Office Hadley Centre Programme funded by BEIS and Defra and by the UK-China Research & Innovation Partnership Fund through the Met Office Climate Science for Service Partnership (CSSP) China as part of the Newton Fund. Figures 4 and 11 were kindly provided by Drs Hua Lu and James Anstey, respectively. James Anstey, Martin Andrews and Scott Osprey generously provided assistance in obtaining the reanalysis data used in Figures 2, 5, 7, and 13.
References


43


