Chapter 27

100 Years of Progress in Understanding the Stratosphere and Mesosphere

MARK P. BALDWIN, THOMAS BIRNER, GUY BRASSEUR, JOHN BURROWS, NEAL BUTCHART, ROLANDO GARCIA, MARVIN GELLER, LESLEY GRAY, KEVIN HAMILTON, NILI HARNIK, MICHAELA I. HEGGLIN, ULRIKE LANGEMATZ, ALAN ROBOCK, KAORU SATO, AND ADAM A. SCAIFE

ABSTRACT

The stratosphere contains ~17% of Earth’s atmospheric mass, but its existence was unknown until 1902. In the following decades our knowledge grew gradually as more observations of the stratosphere were made. In 1913 the ozone layer, which protects life from harmful ultraviolet radiation, was discovered. From ozone and water vapor observations, a first basic idea of a stratospheric general circulation was put forward. Since the 1950s our knowledge of the stratosphere and mesosphere has expanded rapidly, and the importance of this region in the climate system has become clear. With more observations, several new stratospheric phenomena have been discovered: the quasi-biennial oscillation, sudden stratospheric warmings, the Southern Hemisphere ozone hole, and surface weather impacts of stratospheric variability. None of these phenomena were anticipated by theory. Advances in theory have more often than not been prompted by unexplained phenomena seen in new stratospheric observations.

From the 1960s onward, the importance of dynamical processes and the coupled stratosphere–troposphere circulation was realized. Since approximately 2000, better representations of the stratosphere—and even the mesosphere—have been included in climate and weather forecasting models. We now know that in order to produce accurate seasonal weather forecasts, and to predict long-term changes in climate and the future evolution of the ozone layer, models with a well-resolved stratosphere with realistic dynamics and chemistry are necessary.

1. Introduction

The history of stratospheric and mesospheric discoveries over the past ~100 years is a fascinating story of perplexing observations, followed by experimentation, theory, and iterative modeling of the unexplained phenomena to identify their physical and chemical origins. Advances in our understanding have been made possible by 1) improved and more detailed observations of both dynamical and chemical quantities (including in situ, ground-based remote sensing, and remote sensing from satellites); 2) theoretical advances, especially in understanding the behavior of waves and their...
interaction with the background flow; 3) increases in computational power and methods that allow ever more realistic numerical simulations; and 4) reanalysis and data assimilation in which global observations and models are combined to produce gridded output for analysis.

By observing weather in mountainous regions, it has long been known that temperature decreases with altitude (Fig. 27-1). Ground observations alone, however, could give no indication that temperatures might start to increase at some higher level, so the existence of the stratosphere was not anticipated. The stratosphere (from the Latin “stratum,” meaning layered, stratified) was discovered independently by Teisserenc de Bort (1902) and Assmann (1902) using balloon flights to obtain direct temperature measurements. These observations showed that the decrease of temperature with height observed in the troposphere ceased near 10–12 km and was replaced by an isothermal layer up to the greatest heights (about 17 km) sampled by the balloons in use at the time (see, e.g., Hoinka 1997). Throughout the first two decades of the twentieth century the established view was that the atmosphere consists of the troposphere overlain by a nearly isothermal stratosphere. While the altitude range of in situ temperature observations slowly increased, this was limited by the capability of high-altitude platforms (even in the 1930s the highest balloon ascents reached only about 30 km). The first indication that the stratosphere was not an isothermal layer came from the work of Lindemann and Dobson (1923). Based upon their interpretation of meteor trail observations, they concluded that “between 60 and 160 km…meteor observations…all indicate densities very much greater than those calculated on the assumption of a uniform air temperature of 220 K but consistent with a considerably higher temperature.” Over the following two decades additional indirect determinations of air temperatures above the balloon ceiling were made using acoustic measurements as well as spectrographic observations of airglow and auroral emissions.

It was not until after World War II that rockets were used to probe directly the atmosphere to great heights. By 1947, temperature profiles could be inferred from in situ pressure information (returned via telemetry), along with radar observations of the altitude and speed of the rocket (Best et al. 1947). These observations confirmed the existence of the stratopause, mesosphere, and mesopause (see section 4). In subsequent decades, new meteorological rocket platforms, along with better methods of in situ observation of air temperature, refined our knowledge of the climatological temperature structure of the upper stratosphere and mesosphere. Deduction of the winds in those regions would reveal the seasonally reversing pole-to-pole circulation that has been called “Earth’s grandest monsoon” (Webb et al. 1966). In contrast to the discovery of the tropopause (see section 8), the discovery of the stratopause and mesosphere unfolded over a quarter century, a period bookended by the investigations of Lindemann and Dobson (1923) and Best et al. (1947).

By the end of the nineteenth century, Hartley (1880) had detected the presence of ozone in the upper atmosphere, and the ozone layer itself was discovered in 1913 by the French physicists Charles Fabry and Henri Buisson using measurements of the sun’s radiation (see section 11). Ground-based remote sensing of the upper atmospheric composition also began in the early twentieth century, one focus of which was the measurement of ozone (see section 11). This began with a set of spectrophotometers around Europe established in the 1920s by Dobson and colleagues. Since then the network of Dobson spectrophotometers has been expanded globally, with a particular push during the International Geophysical Year (IGY) in 1957 (which was also the year when accurate ground-based measurements of carbon dioxide were initiated by Keeling, supported by Roger Revelle; see Keeling 1960).
Ground-based and airborne remote sensing techniques to measure ozone column amounts and vertical profile evolved at great pace in the post–World War II period. Techniques include passive remote sensing such as differential optical absorption spectroscopy (DOAS), Fourier-transform infrared spectroscopy (FTS), microwave radiometry, and active remote sensing using differential absorption lidar. Absorption or emission features of ozone as a function of wavelength provide “fingerprints” that can be used to determine amounts along the line of sight, providing profiles of ozone in different parts of the atmosphere. These remote sensing devices are not confined to measuring ozone, and the Network for the Detection of Stratospheric Change (NDSC) was created at the beginning of the 1990s to provide a network of instruments that measure many other important upper-atmospheric constituents, in addition to the Dobson and Brewer ozone spectrophotometers.

Understanding the chemical reactions that control ozone, beginning with the work of Chapman (1930), and the influence of the global Brewer–Dobson circulation (BDC; see section 2) that work in concert to control ozone amounts and their distributions took many more decades. Molina and Rowland (1974) recognized the importance of the growing release of chlorofluorocarbons, which are long lived in the troposphere but release chlorine through photolysis or reaction with excited oxygen atoms when transported to the stratosphere. The resultant ClO₂ (Cl, ClO) acts as a catalyst in an extremely efficient “odd oxygen cycle” that destroys ozone. More recently, the threat to stratospheric ozone levels from the release of long-lived bromine compounds, which are also long lived in the troposphere but photolyzed in the stratosphere, was also recognized (Wofsy et al. 1975).

By 1985, theoretical and laboratory work suggested that man-made chlorine and bromine compounds could, and would, increasingly reduce ozone concentrations in the upper stratosphere at midlatitudes. However, no one anticipated the dramatic ozone destruction now referred to as the “ozone hole,” that was first observed to occur over the South Pole each springtime (Farman et al. 1985). The ozone loss was primarily within the polar stratospheric vortex over Antarctica. The phenomenon could only be explained after dedicated theoretical, laboratory, and field campaign efforts, which demonstrated unambiguously that stratospheric ozone in the polar vortex above Antarctica was being destroyed by previously unforeseen catalytic cycles in late winter and spring. The catalytic destruction cycles involve heterogeneous, strongly temperature-dependent equilibrium reactions that take place on the surface of or within small particles, that is, aerosol particles and polar stratospheric clouds (PSCs). Thus, the Antarctic ozone hole, as it has come to be known, was directly linked to the human generation and release of chlorofluorocarbon and organo-bromines into the troposphere. Today, the ozone layer is recovering. However, if those early observations of the ozone hole had not been made, or if the rapid response involving international actions to reduce harmful emissions had not been taken, then the continued depletion of the ozone layer could have led to dire consequences for human life and the biosphere (Morgenstern et al. 2008; Newman et al. 2009).

Progress in our understanding of dynamical processes has also been aided by a number of surprising observations. Sudden stratospheric warmings (SSWs; see section 5), in which the usual boreal wintertime westerly stratospheric circulation breaks down in a few days, were first observed by Scherhag (1952), and occur about every other year in the Northern Hemisphere (NH). The first theoretical model (and numerical simulation) of a sudden warming by Matsuno (1971) combined theoretical aspects of vertical wave propagation and the effect of the waves on the mean flow, producing a realistic result (see section 3). The mechanism involved vertically propagating large-scale waves “breaking” in the stratosphere and slowing the mean flow. It was assumed that the absence of observations of SSWs in the Southern Hemisphere (SH) meant that they were only possible in the NH, but this assumption was proved wrong in the austral spring of 2002 when the first SH SSW was observed. Today, a requirement of stratospheric models is the ability to produce realistic SSWs at roughly the observed frequency in the NH.

A surprising observation in 1961 was the quasi-biennial oscillation (QBO; section 7). This is the largest of Earth’s jet streams—a concentrated, intense, elongated flow (Baldwin et al. 2007b)—and accounts for approximately 4% of atmospheric mass. It spans ~20°S–20°N and ~100–500 hPa, and it consists of downward-propagating easterly and westerly wind regimes that repeat at irregular intervals averaging 28 months (Baldwin et al. 2001). At the time of its discovery in 1961 there was no theoretical explanation. The initial breakthrough in understanding came in 1968, when it was realized that vertically propagating and dissipating waves provide the force needed to drive the phenomenon (Lindzen and Holton 1968, hereafter LH68). Even now, the QBO is still producing surprises. In 2016 an unanticipated disruption of the usual cycle was observed, which likely resulted from unusual wave forcing from the NH (Osprey et al. 2016; Newman et al. 2016).

All of the above examples of advances in our understanding of the stratosphere have occurred as a result of improved observations, but observing the stratosphere...
is particularly challenging. Prior to the 1950s there were few observations of stratospheric winds and temperatures, and most of these were in the NH. Weather balloons tend to burst at low pressures, and few balloons ascend above \(\sim 6\) hPa, even today. Rocketsondes (e.g., Baldwin and Gray 2005) provided sparse measurements extending into the mesosphere from the 1940s to the mid-1990s. However, after the birth of the space age in 1957, coinciding with the IGY, meteorological parameters, and also bulk and trace gas composition, began to be probed from space. The majority of the reliable observations of temperatures, winds, and chemical constituents from instruments on satellite platforms began in the 1970s and have provided routine coverage of the global stratosphere since 1978. Separate monograph chapters are available that describe atmospheric observing systems (Stith et al. 2019) and satellite observations (Ackerman et al. 2019), and they discuss in more detail the advances over the past 100 years. Today, atmospheric reanalysis products that assimilate all available historical observations produce fairly reliable records of stratospheric conditions (at least up to \(10\) hPa), beginning in 1958 for the NH and from late 1978 in the SH (Fujisawa et al. 2017). Prior to 1978 there were insufficient observations to ascertain even the basic large-scale flow in the SH, and reanalysis products prior to 1978 are not considered to be reliable because there are simply too few observations to anchor these early reanalyses.

More recently, long-duration balloons designed to float at stratospheric levels for periods of weeks to months have been developed and employed in a number of coordinated observing campaigns (e.g., Knudsen et al. 2006). These balloons provide quasi-Lagrangian air trajectories and have been used as a platform for temperature and chemical measurements. The high-frequency oscillations (periods up to \(\sim 10\) h) of such balloons can also provide information on the local inertia–gravity wave field (Bocars et al. 2008). A further exciting development has been the deployment of long-duration stratospheric balloons for terrestrial radio communication providing “platforms of opportunity” to make stratospheric measurements (Friedrich et al. 2017).

In the past 50 years, we have lived through a pioneering and exploratory period of atmospheric passive remote sensing from space-based platforms (e.g., SPARC 2017). The advantages and disadvantages of measurements at different wavelengths, frequencies, and energy spectral regions have been probed and exploited to achieve an impressive array of observations. A key challenge is to deliver good vertical resolution profiles, through limb or occultation measurements, so that the dynamical and chemical processes that influence ozone, temperature, and circulation patterns can be better understood and represented in the climate models employed to predict how our atmosphere is likely to evolve in the future. However, we are currently in a period where many relatively long-lived satellite missions have ended or are well over their guaranteed lifetime in space. It is unclear whether there will be an adequate set of satellite observations to meet the future needs of the scientific community. The lack of an adequate continuous set of measurements providing vertical profiles of the required meteorological parameters including chemical species now seems likely.

In parallel to the major advances in observational capabilities since the 1970s, massive gains have been made in computational power—in processing speed, data storage, and transmission speeds. This has enabled substantial increases in horizontal and vertical resolution of weather and climate models so that they more accurately represent the relevant dynamical, radiative, and chemical processes. These have stimulated theoretical advances, particularly in geophysical fluid dynamics, and resulted in improved understanding and numerical simulations of stratospheric phenomena such as the BDC, SSWs, and the QBO. Increases in computational power are particularly important for numerical model studies to understand phenomena that are nonlinear.

In addition, increased computational power has enabled weather centers to increase the vertical extent of the models used for routine weather forecasting to fully encompass the stratosphere, in response to the need to assimilate satellite observations that span both the stratosphere and the troposphere, and in recognition of the influential role that the stratosphere has on the underlying weather and climate (see section 15). Similarly, climate model centers are moving to include more extensive and detailed stratospheric processes, including interactive chemistry, in recognition of the coupled nature of the stratosphere–troposphere dynamical system and the interaction of ozone chemistry and climate (see section 14). For example, previous assessments from the Coupled Model Intercomparison Project (CMIP) that provide input to the Intergovernmental Panel on Climate Change (IPCC) included very few climate models with a fully resolved stratosphere, but in the current CMIP6 exercise many more will do so (see section 16), and several will include fully interactive chemistry schemes. By doing so, the climate models are better able to represent and predict impacts from climate forcings that involve the stratosphere, including the impact of explosive volcanic eruptions that inject large amounts of sulfate aerosol precursors into the lower stratosphere (Robock 2000; see section 13) and the 11-yr cycle in solar radiation that is
known to influence stratospheric ozone and circulation (Gray et al. 2010; see section 14).

2. The Brewer–Dobson circulation

The BDC describes the mass circulation of tropospheric air into and through the stratosphere. “It is particularly prominent because of its widespread controlling influence on the stratosphere...it has important roles in determining the thermodynamic balance of the stratosphere, the lifetimes of chlorofluorocarbons (CFCs) and some greenhouse gases, the temperature of the tropical tropopause, the water vapor entry into the stratosphere, the period of the tropical quasi-biennial oscillation, and the transport and redistribution within the stratosphere of aerosols, volcanic and radioactive debris, and chemical tracers such as ozone.” (Butchart 2014, p. 178).

The concept for this circulation had been proposed by Brewer (1949) and Dobson (1956) to explain observations of water vapor and ozone, respectively, though initial speculation about this conceptual model originates in the work of Dobson et al. (1929). Brewer deduced from water vapor measurements that there must be a mean meridional circulation in the stratosphere, but he then noted that he could not explain the angular momentum balance of the air in such a circulation. An adequate explanation would require invoking the ideas of wave momentum fluxes that came much later, and these are discussed below in sections 5 and 7. Dobson (1956) was trying to explain why observed total columns of ozone were lower in the tropics than in polar regions, even though most ozone is produced in the tropical stratosphere. Both Brewer and Dobson argued that their observations implied a global mass circulation in which tropospheric air enters the stratosphere in the tropics and then moves upward and poleward before descending in the mid- and high latitudes (Fig. 27-2).

More evidence for such a circulation (Sheppard 1963) was found in the 1950s and early 1960s in the patterns of radioactive fallout from atmospheric testing of nuclear weapons. Beginning with Newell (1963) the circulation has been typically referred to in the literature as the “Brewer–Dobson circulation.”

Alongside these results obtained from observations of tracers, Murgatroyd and Singleton (1961) deduced a remarkably similar circulation (often referred to as the “diabatic circulation”) based on stratospheric heating and cooling rates. Like Brewer (1949), Murgatroyd and Singleton noted there were problems with the angular momentum budget that they were unable to address. At the same time, other researchers argued that eddy motions could provide an alternative explanation for both the ozone transport (e.g., Newell 1963) and the heat budget of the stratosphere (Sawyer 1965) without the need for a mean meridional circulation. Vincent (1968) attempted to understand the circulation with an Eulerian-mean analysis including the eddy effects but discovered that, instead of the single hemispheric cell of Brewer’s and Dobson’s proposed model, his analysis indicated two cells, with a reverse cell in the high latitudes with upward motion in the poleward region and downward motion in midlatitudes.

These apparent inconsistencies between the different descriptions of the circulation and explanations for the ozone observations, plus the problems with the angular momentum budget, were eventually resolved in the mid-1970s when Andrews and McIntyre (1976, 1978a,b,c) and Boyd (1976) independently came up with a fluid dynamical explanation. By introducing the so-called transformed Eulerian mean (TEM; see section 3) formulation of the equations, the angular momentum budget could now be balanced by including the contribution from wave momentum fluxes in a way that was physically consistent with a single hemispheric cell description of the circulation.

Importantly, these developments were also instrumental in helping to establish in the 1990s that the BDC is essentially a wave-driven phenomenon (Haynes et al. 1991). We now understand that while diabatic heating associated with seasonal changes in insolation clearly influences the pole-to-equator temperature gradients that determine the background zonal winds through which the waves propagate, the BDC is nevertheless primarily wave driven, and the diabatic (primarily radiative) heating responds to balance the adiabatic cooling and warming patterns induced by the BDC. Briefly, upward-propagating waves from the troposphere transport and deposit westward angular momentum into the stratosphere. To conserve angular momentum this forces a circulation with poleward movement of air at mid-latitudes, upward motion in the tropics, and downward motion at mid- and high latitudes, a process sometimes referred to as “gyroscopic pumping” (Holton et al. 1995).

The major theoretical advances in the 1970s were followed in the 1980s and onward by a growing number of observations stimulated by concerns over the stratospheric ozone layer together with a new capability to measure global trace-gas distributions from Earth-orbiting satellites (see section 10). Complementing this were rapid advances made possible by the development and improvement of stratosphere-resolving general circulation models (GCMs; e.g., Pawson et al. 2000; Gerber et al. 2012), chemistry–climate models (CCMs; e.g., Eyring et al. 2005; SPARC 2010) in the last two decades, and reanalyses (e.g., Iwasaki et al. 2009; Seviour et al. 2012).
Another major theoretical development in the 1990s was the introduction of the concept of “mean age of air” (Hall and Plumb 1994), which is based on the mean transit time for air to reach a particular location in the stratosphere after entry from the troposphere. This single metric combines the effects of the transport by the BDC with those of the mixing by eddies (Kida 1983). Climate model projections suggest a shorter transit time in the future (i.e., “younger” age of air) as a result of climate change (Butchart and Scaife 2001). Apart from in the subtropical lower stratosphere, where this can only result from a strengthened BDC, the younger age generally indicates both the possibility of a strengthening of the BDC and/or weaker mixing or recirculation of the stratospheric air between the tropics and mid-latitudes (Garny et al. 2014).

In recent years a synergy of these developments has led to a more quantitative, dynamically based analysis of the BDC and its driving mechanisms, and most notably its projected response to climate change. An important gap in current knowledge about the BDC is a comprehensive understanding of how wave driving of the stratosphere has changed in response to climate change. Changes in the troposphere could influence the strength of the upward propagating waves, while changes in the background state of the stratosphere will change the way in which the waves propagate (e.g., Bell et al. 2010), and how much of their momentum will get deposited in the flow (e.g., Lubis et al. 2018). Critical-layer control on Rossby wave breaking has been invoked as a possible mechanism (Shepherd and McLandress 2011). However, changes in wave-driving in response to climate change are still uncertain, especially the relative changes between wave-forcing due to planetary waves (zonal wavenumbers 1–3), which drive the deep branch of the BDC (Plumb 2002);
3. Middle atmosphere dynamics theory

The thermodynamic state and the flow in the middle atmosphere are governed by dynamics as well as the complex balance between radiation and photochemical processes that heat and cool the atmosphere. Heating is dominated by absorption of solar radiation by ozone in the stratosphere and molecular oxygen in the thermosphere, while cooling is dominated by infrared emission, mostly by carbon dioxide (CO₂) (e.g., Murgatroyd and Goody 1958). Large deviations of the temperature field from the state of radiative equilibrium are caused by adiabatic heating and cooling processes, which are driven by waves (Leovy 1964). The three principal theoretical paradigms that are applied to middle atmosphere dynamics are as follows: 1) wave propagation, 2) wave mean–flow interaction, and 3) the mean overturning circulation response to radiative forcing and wave driving.

The most important wave modes for middle atmosphere theory are atmospheric gravity waves (see section 6), whose restoring force is buoyancy due to gravity and stable stratification, and Rossby waves (see section 5), for which a combination of differential planetary rotation and stratification provides the restoring force. On spatial scales larger than a few hundred kilometers, gravity waves are modified by Earth’s rotation and are known as inertia–gravity waves. A third type of wave is the atmospheric Kelvin wave, which is analogous to coastal Kelvin waves in the ocean. It exists in the atmosphere because of the change in sign of the Coriolis parameter at the equator, which provides geostrophic balance in the latitudinal direction, but its restoring force is otherwise buoyancy (Holton and Lindzen 1968).

Wave propagation theory tells us how the waves propagate and where they are likely to get absorbed, or if they will get reflected back to their source region. Wave propagation differs among the different wave types, but their interaction with the mean flow shares some common features (e.g., Eliassen and Palm 1961), namely, that under steady, nondissipative conditions, the waves conserve wave pseudomomentum flux. The pseudomomentum indicates the strength of the drag on the flow when the waves are dissipated. Thus, waves affect the flow nonlocally, by essentially transporting momentum from their source region to where they dissipate. This dissipation exerts a drag on the mean flow, which modifies it both directly and indirectly by driving an overturning circulation in response (see section 2).

The theory of atmospheric gravity waves can be traced back to the works of eminent eighteenth-century scientists such as Euler, Lagrange, Laplace, and Newton on water waves, but the crucial role of buoyancy in atmospheric gravity waves began with the works of
Väisälä (1925) and Brunt (1927), who independently derived the frequency of oscillation of an air parcel displaced vertically in a stably stratified dry atmosphere, which now bears their name, the Brunt–Väisälä frequency,

\[ N = \sqrt{\frac{g}{T} \left( \frac{\partial T}{\partial z} + \frac{g}{c_p} \right)}, \]  

(27-1)

where \( g \) is the acceleration due to gravity, \( T \) is temperature, \( z \) is altitude, and \( c_p \) is the specific heat at constant pressure. This buoyancy restoring force acting on slanted displacements gives rise to gravity waves that propagate both vertically and horizontally (see Lindzen 1973). The dissipation of gravity waves results from several processes, namely, radiative damping (Fels 1984), gravity wave breaking (Lindzen 1981), and other nonlinear wave–wave processes (see the review by Fritts and Alexander 2003). All these processes are strongest near critical levels—where the horizontal phase speed of the wave equals the mean flow speed (e.g., Booker and Bretherton 1967; Lindzen 1981). Gravity wave drag is especially strong in the mesosphere and is responsible for reversing the meridional temperature gradient (summer pole is coldest) and forcing the strong meridional summer-to-winter pole circulation (Lindzen 1981).

Many of the gravity waves have horizontal and vertical scales that are too small to be resolved in climate models and must be parameterized (e.g., Lindzen 1981; McFarlane 1987; Hines 1997; Alexander and Dunkerton 1999; Warner and McIntyre 2001). The dispersion relation for atmospheric gravity waves, neglecting the effects of Earth’s rotation and compressibility for simplicity, is

\[ \hat{\omega}^2 = \frac{N^2 k_h^2}{k_h^2 + m^2 + \frac{1}{4H^2}}, \]  

(27-2)

where \( k_h \) is horizontal wavenumber satisfying \( k_h^2 = k^2 + l^2 \), with \( k, l \) being the zonal and meridional wavenumbers, respectively; \( \hat{\omega} \) is the intrinsic wave frequency; \( \omega = \hat{\omega} - ku_0 - hv_0 \), with \( \omega \) being the wave frequency and \( u_0 \) and \( v_0 \) the zonal and meridional mean flow velocities, respectively; \( m \) is the vertical wavenumber; and \( H \) is the pressure-scale height. The vertical group velocity, \( \partial \omega / \partial m \), is oppositely directed relative to the vertical phase velocity in the frame of reference relative to the mean wind. Assuming a small mean-wind Doppler shift, one can easily derive the direction of the vertical group velocity, which is the sense of wave energy propagation (see, e.g., Fritts and Alexander 2003).

Conservation of potential vorticity (PV) gives rise to atmospheric planetary waves, also known as Rossby waves after C.-G. Rossby, who introduced them in Rossby (1939). The equation for Rossby wave propagation from the troposphere into the stratosphere was first derived by Charney and Drazin (1961) and later extended to include latitudinal propagation by Dickinson (1968) and Matsuno (1970). It indicates that Rossby waves can only propagate to the stratosphere if the zonal flow is westerly and below a certain critical value, and if the wavenumber is small. This explained why stratospheric Rossby waves are planetary scale and only found in winter (Charney and Drazin 1961). Moreover, the equations also indicate the existence of two kinds of surfaces that block wave propagation—critical surfaces that lead to wave absorption (Eliassen and Palm 1961; Matsuno 1971) and turning surfaces that reflect the waves (Sato 1974; Harnik and Lindzen 2001).

The dispersion relation for atmospheric Rossby waves, using the quasigeostrophic (QG) approximation, is

\[ \hat{\omega} = -\frac{\partial q_0 k}{\partial y} \frac{k_h^2 + m^2 f^2}{N^2} - \frac{f^2}{4N^2H^2}, \]  

(27-3)

where \( f = 2\Omega \sin \phi \) is the Coriolis parameter; \( \Omega \) being the rotation frequency of Earth; \( \phi \) is latitude; \( \partial q_0 / \partial y \) is the meridional gradient of mean flow quasigeostrophic potential vorticity (QGPV), which has a planetary component \( \beta = \partial f / \partial y = (2\Omega \cos \phi) / a \) and a contribution from the meridional and vertical curvature of the zonal-mean zonal wind (see, e.g., appendix of Harnik and Lindzen 2001); \( y \) is the latitudinal distance coordinate; and \( a \) is Earth’s mean radius, and we have simplified the equations by assuming QG dynamics and a zonal-mean flow (typically assumed for Rossby waves), with constant \( N^2 \) and an exponentially decreasing pressure with scale height \( H \).

The linear theory for atmospheric gravity waves can be modified easily to include the local effects of planetary rotation leading to a generalization of the dispersion relation (27-2) for “inertia–gravity waves” (e.g., Holton and Hakim 2013, 153–154). Including the effects of Earth’s rotation near the equator leads to another class of wave modes denoted “equatorial waves” (Matsuno 1966). The gravitational influence of the sun and moon, as well as the sun’s heating effects, gives rise to the atmospheric tides, waves with frequencies related to the astronomical frequencies of the solar and lunar days [see Chapman and Lindzen (1970) and section 4]. The following fundamental relationships for atmospheric wave–mean flow interactions have their origins in Eliassen and Palm (1961). For adiabatic flow with \( f = 0 \), no wave transience, and \( u_0 - c \neq 0 \),

\[ \overline{p' \omega'} = -\rho_0 (u_0 - c) \overline{u' w'}, \]  

(27-4)
where \( p \) and \( w \) are the pressure and vertical velocity, respectively; \( \rho \) is the basic state density; \( u_0 \) is the basic state flow, taken here to be the zonal-mean wind; the primes indicate wave quantities; \( c \) is the wave phase velocity; and overbars denote averages over wave phase. Equation (27-4) indicates that if \( u_0 > c \), an upward (group) propagating wave (upward energy flux \( \rho' u' w' > 0 \) will have negative vertical flux of horizontal momentum \( \rho_0 u_0 w_0 \)); that is to say, the waves will tend to decelerate the mean flow toward the wave phase velocity in the presence of dissipative effects. It also indicates that waves cannot propagate vertically through critical levels.

Equation (27-5) implies that in the absence of wave transience and critical levels and diabatic effects, there is no interaction between waves and the mean flow. The Eliassen–Palm relationships in Eqs. (27-4) and (27-5) have been generalized by Andrews and McIntyre (1978a,b,c) and Boyd (1976) and have led to the TEM (see section 1) formulation of the dynamics, which relates the Eulerian zonal-mean fields to the approximate Lagrangian overturning meridional-vertical circulation, considering wave-induced Stokes drift effects. The TEM equations are discussed in the general circulation chapter of this monograph (Held 2019).

The Eliassen–Palm relations imply the well-known nonacceleration theorem; that is to say, for steady waves with no dissipation and no critical levels, waves propagate through the mean flow without leading to accelerations or decelerations. An important counterpart to this is the nontransport theorem, which states that under the conditions for nonacceleration, no net transport by the waves occurs for chemical species that have lifetimes much longer than the dynamic time scales. While the nonacceleration theorem appears to be a negative result, it is important in that it identifies those factors that give rise to wave mean–flow interactions.

SSWs are a spectacular example of wave–mean flow interaction in the middle atmosphere involving Rossby waves, first successfully modeled by Matsuno (1971) and described in detail in section 5 below. Why SSWs occur during specific winters and not others is not yet clear, however, it has been shown that downward reflection of waves dominates the daily variability during most of the winters which lack SSWs (Perlitz and Harnik 2003). Many of the interannual differences can be rationalized as arising from the different effects which SSWs (wave absorption) and wave reflection have on the mean flow deceleration (Perlitz and Harnik 2004), on the resulting overturning circulation (Shaw and Perlitz 2014), and correspondingly on poleward ozone transport and concentrations (Lubis et al. 2017).

Examples of wave–mean flow interaction involving gravity waves are shown in Figs. 27-3 and 27-4, using results from some NASA Goddard Institute for Space Studies (GISS) models. Note that the upper-stratospheric winds are excessive in Fig. 27-3b, and are much more excessive in the winter NH than in the summer SH. The Fig. 27-3c winds, while not agreeing perfectly with those in Fig. 27-3a, are much more realistic in both hemispheres. Comparing Figs. 27-3b and 27-3c, we see that gravity waves play a major role in making the upper-stratospheric winds more realistic in Fig. 27-3c in the winter hemisphere, while they play a lesser role in the summer hemisphere.

Another aspect of the importance of the wave–mean flow interactions is shown in Fig. 27-4. The QBO (described in more detail in section 7) is a quasiperiodic reversal in the equatorial mean zonal wind in the lower stratosphere with a mean period of about 28 months.
The first successful explanation for the QBO was given by LH68, in terms of wave–mean flow interaction. It involves a constant wave flux of easterly and westerly momentum at a bottom boundary, notionally taken to be the tropopause. When the mean wind is greater than zero ($u_0 > 0$), there is preferential absorption of westerly momentum at lower levels, and the easterly momentum flux penetrates to high levels, giving rise to easterly ($u_0 < 0$) wind at high levels. The easterly and westerly winds descend until easterly winds prevail at lower levels. Then, the situation repeats giving rise to the QBO [see Plumb (1977) for a schematic illustration of this process].

Figure 27-4a shows an example of the QBO oscillation in the mean zonal wind averaged between 4°S and 4°N from the ERA-40 reanalysis. Figures 27-4b and 27-4c show the same plot from a GISS model with a gravity wave parameterization, with gravity wave momentum flux at 100 hPa at the equator equal to 0.5 and 3.0 mPa, respectively.

PV is another concept that is extremely useful in middle atmosphere dynamics, and is defined as

$$PV = -g(\partial \theta / \partial p)(\zeta_0 + f),$$

where $g$ is gravity, $\theta$ is potential temperature, $p$ is pressure, $\zeta_0$ is relative vorticity evaluated along isentropic surfaces, and $f$ is the Coriolis parameter. During the 1980s, advances were made in our understanding of stratospheric dynamics by applying “PV thinking” (Hoskins et al. 1985). PV has large gradients between the troposphere and stratosphere near the tropopause. In fact, this has led to defining the “dynamical tropopause” in terms of a given PV value in the extratropics (e.g., see Hoskins et al. 1985). See section 8 for a more general discussion of the tropopause.

PV generally increases poleward, largely due to $f$ increasing, but in winter there is usually a particularly large PV gradient at the edge of the winter polar vortex, due to large variations in the horizontal shear of the zonal winds, and this can act as a transport barrier. It was pointed out by McIntyre and Palmer (1983) that planetary wave breaking gives rise to the mixing of chemical constituents and PV at midlatitudes, and while this tends to reduce latitudinal PV gradients where the mixing occurs, in the region referred to as the “surf zone” (McIntyre and Palmer 1984), it also serves to strengthen the gradients at the subtropical and polar edges of the surf zone. One consequence of this is the sharpening of the large PV gradients at the edge of the vortex. The very large PV gradients at the equatorward edge of the SH polar vortex that acts as a transport barrier has been referred to as a “containment vessel,” where the air inside that vortex is largely isolated from the lower-latitude air. This was a crucial aspect of explaining the Antarctic ozone hole (see Solomon et al. 1986). Erosion of the large PV gradients at the edge of the NH polar vortex was also suggested by McIntyre (1982) to be crucial in setting the conditions for SSWs. Finally, examination of transport processes in the vicinity of the large PV gradients at the equatorward edge of the surf zone, which can also impede transport across the subtropics, has led to the concept of the tropical “leaky pipe” (see Plumb 1996).

There are many more theoretical concepts in middle atmosphere dynamics. Because of space limitation, we have concentrated on the wave–mean flow interaction as the main paradigm in the field. Many of the theoretical approaches outlined above have been quasi-linear, in the sense that the waves interact with the mean flow, but not with each other. This is an unrealistic assumption, but such models have served the field well as a template for understanding middle atmosphere dynamics.

4. Atmospheric thermal tides

Atmospheric thermal tides are global-scale, periodic oscillations that are excited mainly by absorption of solar radiation by ozone and water vapor, and by latent
heating due to tropical deep convection. The thermal tides were first documented through their signature in surface pressure. These “barometric tides” are remarkable in that the semidiurnal oscillation has much greater amplitude than the diurnal, even though solar heating is obviously dominated by its diurnal component. The paradox was noted explicitly by Kelvin (Thomson 1882), who hypothesized that the larger amplitude of the semidiurnal tide relative to the diurnal could be explained by the existence of a “free,” or resonant, solution of Laplace’s tidal equation with period near 12 h.

The effort to substantiate Kelvin’s hypothesis led to systematic exploration of Laplace’s tidal equation as applied to Earth’s stratified atmosphere. These studies, together with the increasing ability to observe temperature and winds above the tropopause using radiosondes and—beginning in the late 1940s—rocket sondes, shaped our current understanding of the tides throughout Earth’s atmosphere. It was found that heating due to absorption of solar radiation by ozone in the stratosphere and water vapor in the troposphere were the leading sources of excitation (e.g., Siebert 1961; Butler and Small 1963). Kelvin’s resonance hypothesis was eventually discarded, and the unexpectedly small amplitude of the diurnal surface pressure tide was shown to arise from the propagation characteristics of the tidal “modes” that are solutions to the tidal equations. Specifically, the diurnal component of ozone heating in the stratosphere projects most strongly on modes that are nonpropagating, or “trapped,” whereas the semidiurnal component can propagate to the surface (Kato 1966; Lindzen 1966, 1967). The history of this work, together with the development of the mathematical theory of the tides, is summarized in Chapman and Lindzen’s (1970) monograph on the subject.

The introduction of satellite-borne observing systems in the late 1970s enormously enhanced the ability to document the global behavior of the tides from the troposphere to very high altitudes in the ionosphere. At the same time, rapidly increasing computational capabilities allowed numerical solution of the tidal equations in atmospheric global models and detailed comparisons between numerical predictions and observations. The amplitude of nondissipating waves in a stratified atmosphere grows with altitude, \( z \), as \( \exp(z/2H) \), where \( H \) is the scale height, such that the temperature and wind perturbations associated with the tides become very large in the mesosphere. At still higher altitudes, in the thermosphere, growth of these waves ceases as they are damped by molecular diffusion. Amplitude growth can also be limited by dissipation due to wave “breaking” if the tides become dynamically or convectively unstable (Lindzen 1981). These processes can make a substantial contribution to the momentum and thermodynamics budgets of the thermosphere (e.g., Becker 2017).

Much recent work on atmospheric thermal tides has focused on their behavior in the range of altitude from the tropopause (10–15 km) to the lower thermosphere (~150 km), as discussed in the recent review article by Oberheide et al. (2015). England (2012) has reviewed the tides at even higher altitudes, in the ionosphere. Sassi et al. (2013) used a global model extending to 500 km to study the migrating and nonmigrating tides in the thermosphere and showed that the diurnal tide undergoes a striking change in structure in the lower thermosphere, where the upward-propagating (1, 1) mode disappears due to dissipation by molecular diffusion. The (1, 1) designation refers to the westward-propagating, wavenumber 1, first mode of the inertia–gravity wave manifold (see Chapman and Lindzen 1970), which is the main component of the upward-propagating diurnal tide. Above ~120 km, the (1, 1) mode is replaced by a latitudinally broad, nonpropagating external mode, which is forced by in situ extreme UV solar heating.

Nonmigrating or, more properly, non-sun-synchronous tides have been documented recently in observations of the mesosphere and lower thermosphere. These are oscillations whose periods are harmonics of the solar day but do not propagate westward following the sun. They arise from diurnal but spatially fixed forcing, associated principally with the diurnal cycle of deep convective heating in the troposphere (Lindzen 1978; Hamilton 1981; Forbes et al. 1997). Along these lines, Gurubaran et al. (2005) and Pedatella and Liu (2012) have documented an apparent modulation of tidal amplitudes by El Niño–Southern Oscillation (ENSO), which is a principal source of interannual variability in tropical convection. Several nonmigrating tides have been observed (Talaat and Lieberman 1999, 2010; Forbes and Wu 2006; Li et al. 2015), including an eastward-propagating wavenumber-3 diurnal oscillation (DE3), which features prominently in satellite observations of the mesosphere and lower thermosphere. The structure of DE3 is shown in Fig. 27-5, which is constructed from observations made by the Sounding of the Atmosphere Using Broadband Emission Radiometry (SABER) infrared radiometer (Russell et al. 1999) using squared coherence analysis, as detailed by Garcia et al. (2005). The role of DE3 in coupling the lower thermosphere to the ionosphere has been demonstrated by Immel et al. (2006), who documented a link between longitudinal variability in ionospheric density in the F region (250–400 km) and the amplitude of nonmigrating diurnal tides. The link operates mainly via tidal modulation of
the electric field in the E region (100–150 km), which in turn couples to the F region (Hagan et al. 2007; Xiong and Lühr 2013). This discovery established a link between “space weather” and the tropospheric weather (tropical convection) that excites DE3. England et al. (2010) have also explored the impact of tides on the ionosphere and further illustrated the coupling between tides in the lower thermosphere and electron density in the ionosphere.

Tides in the middle atmosphere display marked seasonal and interannual variability. Radar observations (Vincent et al. 1998) show that the amplitude of the diurnal tide has a prominent semiannual variation in the mesosphere and lower thermosphere (MLT). This variability is also seen in satellite observations (e.g., Hays et al. 1994; Burrage et al. 1995) Semiannual and quasi-biennial modulations in tidal amplitudes are clearly displayed in data obtained from the SABER instrument, as shown in Fig. 27-6. McLandress (2002a) used a linear mechanistic model to attribute the semiannual variation to changes in the horizontal shear of the background zonal-mean wind $u_0$, and argued that this influences the tide mainly through its contribution to the barotropic vorticity term, \[ f - \frac{(\partial u_0/\partial y)}{f} \], near the equator, where $f$ is small and comparable to $\frac{(\partial u_0/\partial y)}{f}$. Burrage et al. (1995), Lieberman (1997), and Vincent et al. (1998) have reported interannual variability in the amplitude of the diurnal tide, which is apparently related to the stratospheric QBO. This behavior has been reproduced in a numerical model by McLandress (2002b), who concluded that the mechanism responsible for the semiannual modulation of the diurnal tide also causes the quasi-biennial modulation. Smith et al. (2017) have shown recently that temperature data from the SABER and Microwave Limb Sounder (MLS) satellite instruments can be used to estimate the zonal-mean zonal winds in the tropics (Fig. 27-8). Comparison of Figs. 27-6 and 27-7 shows the relationship between the diurnal tide in the MLT and the tropical winds. It may be possible to use such data, derived from a common source, to elucidate further the relationship between tidal amplitudes and tropical mean zonal wind variations.

Despite these advances, it is apparent that accurate simulation of the tides in comprehensive numerical models remains a challenge. For example, Davis et al. (2013) used meteor radar data at Ascension Island to investigate the seasonal variability of the diurnal and semidiurnal tides and noted that two leading “high-top” models produce results that are not in general agreement with observations. A possible reason for these discrepancies is that simulation of the tropical wind oscillations, the QBO and the semiannual oscillation (SAO), is still unsatisfactory in many comprehensive numerical models.
If, as argued by McLandress (2002a,b), tidal amplitudes are modulated through the effect of the tropical background wind on the barotropic vorticity gradient, then accurate simulation of the QBO and SAO would be a prerequisite for simulating seasonal and interannual variability of the tides.

In contrast to the situation in the middle atmosphere, simulation of the tides in the troposphere and, in particular, the surface barometric oscillation, has produced results that are often in largely good agreement with observations. Numerical calculations have been successful in reproducing the amplitude and phase of the semidiurnal barometric tide and ascribing the behavior to the combined effects of sources of excitation and propagation (Siebert 1961; Butler and Small 1963; Lindzen 1966; Kato 1966). These early results have been refined in recent work. For example, the phase of the semidiurnal pressure maximum in the tropics, which is found to occur at 0930–1000 local time (LT; Haurwitz 1956; Schindelegger and Ray 2014) differs from the 0910 LT that is calculated assuming that the principal sources of excitation are heating due to tropospheric water vapor and stratospheric ozone. This led Lindzen (1978) to propose that stratospheric convective heating could account for the difference between theory and observations. A recent study by Sakazaki and Hamilton (2014) differs from the 0910 LT that is calculated assuming that the principal sources of excitation are heating due to tropospheric water vapor and stratospheric ozone. This led Lindzen (1978) to propose that stratospheric convective heating could account for the difference between theory and observations. A recent study by Sakazaki and Hamilton (2014) examined the dependence of the phase of the semidiurnal tide and found that accurate simulation of latent heat release and mechanical dissipation are necessary to obtain good agreement with observations. In particular, their model produces a realistic diurnal cycle of rainfall; suppressing the diurnal cycle of rainfall in the model changes the phase of the semidiurnal tide from a realistic 0940 to 0915 LT, consistent with what is found in linearized models that exclude convective heat release. Suppressing mechanical dissipation advances the phase further, to ~0910 LT. On the other hand, Sakazaki et al. (2017) showed that the daily cycle of tropical rainfall is itself influenced by the component of the semidiurnal tide excited by ozone heating in the stratosphere. The results of Sakazaki et al. (2017) highlight the fact that tides are global phenomena that couple the lower and middle atmosphere, the complete understanding of which requires consideration of excitation and propagation mechanisms throughout a wide range of altitudes.

In summary, despite substantial theoretical, observational, and numerical modeling advances in the last 40 years, there remain important deficiencies in our understanding of the atmospheric thermal tides. In particular, simulation of the tides in the middle atmosphere remains challenging and the relationship between tidal variability and the variability of the tropical wind needs further investigation. Realistic simulation of tides in the troposphere has, in general, been much more satisfactory. However, simulation of the amplitude and phase of the barometric tides is not uniformly successful, and may depend on the accurate representation of other processes, principally convection and mechanical dissipation, but also the details of stratospheric ozone heating, in comprehensive models.

5. Sudden stratospheric warmings

The wintertime stratospheric polar vortex is formed primarily through radiative cooling and is characterized by a band of strong westerly winds at mid- to high latitudes. The polar vortex can be disrupted by large wave perturbations, primarily planetary-scale zonal wave-number 1–2 quasi-stationary waves. Sufficient wave forcing of the mean flow by these waves (see section 3), that is, the transfer of westward momentum to the background flow, can result in an SSW, with the breakdown of the polar vortex and replacement of westerly winds by easterlies. As described in section 2, air is then forced to move poleward to conserve angular momentum, with descent at mid- and high latitudes that forms the poleward extent of the BDC. The adiabatic heating associated with this descent (thus maintaining thermal wind balance) leads to the rapid increases in polar cap temperatures on time scales of just a few days that give SSWs their name. Once the vortex is destroyed, strong radiative cooling will help to rebuild the vortex provided there is time before the end of winter, but this radiatively controlled process can take several weeks. None of this was known in 1952, when Scherlag (1952) observed “explosive warmings in the stratosphere” in radiosonde observations over Berlin.
up to 40 km (see section 1). It was not until the late 1950s that it became clear that SSWs were taking place on a hemispheric scale and that they appeared to occur randomly in around half of NH winters.

During the second half of the twentieth century, a lot of the basic knowledge of SSWs, including their typical characteristics and interannual variability, was based on observational analyses carried out by Karin Labitzke and her Stratospheric Research Group in Berlin. This included a summary of typical SSW characteristics based on radiosonde and satellite data (Labitzke 1981), the definition of different types of SSWs and the documentary descriptions of all SSWs for the years 1951/52 to 1980/81 (Labitzke 1982).

SSWs can be classified as major, minor, and final (the latter are followed by the transition to summertime easterly conditions). Various classification methods have been proposed, based on zonal winds, temperatures, PV, the northern annular mode (NAM), or other more highly derived quantities (Charlton and Polvani 2007; Butler et al. 2015), but no unambiguous standard definition has so far been agreed on. All definitions are somewhat arbitrary because the parameters vary continuously, and therefore a threshold value is needed in the definition. Conventionally, SSWs are classed as a “major SSW” if the direction of both the equator-to-pole temperature gradient reverses and the zonally averaged winds at 60°N, 10 hPa reverse to become easterly. A “minor SSW” is said to have occurred if the polar temperature gradient reverses but the zonal winds do not.

There has been only one major warming observed in the SH, in 2002. Rossby waves are generated by mountain ranges and land–sea contrasts, so Rossby wave amplitudes are smaller in the SH. There is less wave forcing of the SH winter stratosphere and the polar vortex remains less disturbed (thus providing the colder, more isolated vortex conditions that favor ozone destruction and lead to the ozone hole). Nevertheless, we now know, from the 2002 event, that even though major SH warmings are far less likely than in the NH, they are nevertheless possible.

Figure 27-8 shows the time series of temperatures at 10 hPa (65°–90°N) and zonal wind at 60°N for the winter of 2018/19, in which a major warming occurred, compared to the observed variability during 1979–2019. Note that the circulation is quiescent during summer, but highly variable during winter. The temperature rise during January occurs in less than one week, and is caused by large-scale descent over the polar cap. The corresponding reduction in zonal wind reverses the flow to easterly.

Perhaps the most vivid way of viewing sudden warmings is through the lens of PV (see section 3). By examining maps of PV on isentropic surfaces, it is possible to observe the breaking of planetary-scale Rossby waves, arguably one of the most important dynamical processes affecting the stratosphere. The reason for thinking in terms of such maps is that they are fundamentally the simplest and most useful way to visualize large-scale dynamical processes (McIntyre and Palmer 1984). From this perspective stratospheric warmings can be seen to arise as a consequence of planetary-scale wave breaking, which causes erosion of the polar vortex and, ultimately, its destruction. Typically, a potential temperature level such as 850 K (~10 hPa or ~30 km) is used. During early winter the vortex strength increases (due to radiative cooling), but as winter progresses, wave breaking in the surf zone sharpens the edge of the vortex (as described in section 3), and if the wave breaking is persistent enough, the vortex becomes displaced from the pole and decreases in size. All this can be viewed on horizontal maps of PV, or indeed simply by measuring the size of the polar vortex in terms of PV (e.g., Butchart and Remsberg 1986).

Figure 27-9 is a sequence of six isentropic maps of PV at 850 K, every 7 days, during the SSW in 2018/19. The PV structures are not at all zonally symmetric and illustrate breaking waves, the reduction in size of the polar vortex, and the vortex breakup. The blue streaks illustrate how PV is stripped away from the central vortex and mixed. When averaged over many events (Fig. 27-10), it becomes clear that the polar vortex decreases in size over time during SSWs, but because PV has conservative properties, the process appears to be much more gradual than the evolution of a quantity such as temperature. When wave breaking strips PV off the edge of the vortex, it is mixed, but when viewed in terms of equivalent latitude (the latitude at which a zonally symmetric PV contour would lie if it enclosed the same area as the actual PV contour), the process appears to be quite smooth in time (Butchart and Remsberg 1986), and it becomes clear that the PV anomalies can last for 2–3 months from their onset date.

The mechanistic connection between planetary-scale wave forcing of tropospheric origin and SSWs was first established by the numerical model experiments of Matsuno (1971; see also sections 1 and 3), which showed that an SSW can be triggered by enhanced vertical planetary wave fluxes near the tropopause. Much research has been devoted since Matsuno’s seminal work to finding corresponding tropospheric precursor signals, which could be used not only to improve our understanding of the dynamics of SSWs, but also our predictive capabilities associated with these events. This includes, for example, links to tropospheric blocking.
events (e.g., Quiroz 1986; Martius et al. 2009). Tropical tropospheric variability associated with ENSO and the Madden–Julian oscillation (MJO) have also been linked to SSWs (e.g., Butler and Polvani 2011). Impacts from so-called “external” forcing are also likely to influence the nature and frequency of SSWs, including explosive volcanic eruptions (section 13), solar cycle variability (section 14), and changes associated with future climate change (section 16), some of which could also provide additional predictive capability.

Numerical experiments with idealized and comprehensive models have shown that while tropospheric wave forcing is an essential prerequisite for an SSW to occur, they do not necessarily require anomalously large wave forcing as a precursor (e.g., Scott and Polvani 2004; Sjoberg and Birner 2014; de la Cámara et al. 2017).

**FIG. 27.8.** (top) The 10-hPa 65°–90°N observed zonal-mean temperatures and (bottom) zonal-mean wind at 60°N for 2018–19. An SSW event is seen as the upward spike (red) in temperature and the reduction to less than zero in zonal wind (easterlies). The yellow line signifies the average conditions in the stratosphere for that time of year, while the gray shadings show 70th and 90th percentiles. Solid black lines show the max/min for 1979–2019. The thin green lines are forecasts. [Source: NOAA/NWS/Climate Prediction Center, https://www.cpc.ncep.noaa.gov/products/stratosphere/SSW/.]
Indeed, Birner and Albers (2017) have recently shown that only $\frac{1}{3}$ of observed SSWs appear to be associated with anomalous wave forcing from the troposphere. SSWs that are generated by a positive wave–mean flow feedback internal to the stratosphere likely fall in the category of self-tuned resonance (Plumb 1981; Matthewman and Esler 2011; Albers and Birner 2014). The onset of an SSW seems therefore to require appropriate configurations of both the stratosphere and troposphere (Hitchcock and Haynes 2016).

Stratosphere–mesosphere variability also modifies atmospheric chemistry, including the distribution of atmospheric trace gases such as ozone (Pedatella et al. 2018). SSWs cause the stratopause to descend in altitude, and chemical species that typically reside in the upper mesosphere are transported downward into the lower mesosphere and upper stratosphere. This results in anomalously large concentrations of species such as nitrogen oxides ($\text{NO}_x$) and carbon monoxide (CO). These changes alter the chemistry of the polar winter stratosphere, for example, raising levels of NO, which acts to destroy ozone.

Looking upward at impacts in the mesosphere ($\sim$50–80 km) and lower thermosphere ($\sim$80–120 km), SSWs begin a chain of events that lead to the modulation of upward-propagating waves and hence to wind and temperature anomalies in both hemispheres (Fig. 27-11; Pedatella et al. 2018) and also affect the atmospheric tides in both the hemispheres (Karlsson et al. 2007). Studies have shown that the impacts can extend even higher, throughout the thermosphere; for example, satellite drag observations show that the temperature and density of the thermosphere are affected by SSWs (Yamazaki et al. 2018).
and into the ionosphere, affecting near-Earth space weather so that irregularities affect communication and navigation signals (Chau et al. 2012; Pedatella et al. 2018).

6. Gravity waves

In the late nineteenth century and the first half of the twentieth century, meteorologists began to understand that some of the subsynoptic-scale variability in tropospheric flow could be interpreted as gravity waves. Notably, the observations and theory of waves forced by flow over topography comprised a substantial area of research (Kuettner 1939; Scorer 1949). However, the mainstream of meteorology regarded gravity waves as a kind of noise that simply complicated the forecasting of synoptic-scale and mesoscale weather. The notion that gravity waves might be important far above the troposphere was first advanced by Martyn (1950), who speculated that observed traveling ionospheric disturbances may be gravity waves, similar in their basic dynamics to those that account for some tropospheric cloud formations and microbarograph fluctuations. Hines (1960) showed that meteor wind observations at ~90 km were consistent with the random superposition of upward-propagating gravity waves. The decrease in atmospheric mean density with height has dramatic implications for gravity waves. Hines (1960) noted that disturbances associated with weather in the lower atmosphere could be expected to generate gravity waves; horizontal wind amplitudes are likely to increase (between the ground and 90 km) by a factor of ~700 due to density changes, and thus the observed upper-atmospheric winds could be produced by wave generation in the lower atmosphere, whose associated oscillatory motions there need be only a few centimeters per second, which are common in the troposphere.

In the 1960s and 1970s the importance of gravity waves in maintaining the zonal-mean circulation of the middle atmosphere became increasingly evident, first in their role in generating the QBO (LH68; see sections 3 and 7), and then in driving the deep part of the BDC in the upper stratosphere and mesosphere (see section 2). The latter results in mesopause temperatures being colder in the polar summer mesopause region than in the winter polar mesopause region (section 3). Observations showed that the upper stratospheric and mesospheric zonal circulation is dominated by a strong westerly (easterly) jet in the winter (summer) hemisphere. The jets become weaker with height in the mesosphere and, as described in section 2, this implies the presence of
meridional flow from the summer to the winter hemisphere in the mesosphere (Murgatroyd and Singleton 1961), and consequently some zonal-mean zonal momentum forcing is required to balance the associated Coriolis torque.

While this extra forcing had already been incorporated into early models of the middle atmospheric general circulation as a Rayleigh friction on the zonal flow (e.g., Leovy 1964), it was Houghton (1978) who suggested that the momentum flux divergence from vertically propagating gravity waves might provide the necessary balance to the Coriolis torque of the mean meridional flow. Lindzen (1981) and Matsuno (1982) then showed that the sign of the required drag could be explained plausibly if one assumes that a broad spectrum of upward-propagating gravity waves with both eastward and westward phase speeds is excited in the troposphere. Selective absorption of the waves with different phase speeds (i.e., filtering of the waves by the background flow) would then lead to a preponderance of westward (eastward) propagating waves in the winter (summer) mesosphere, thus providing the zonal-mean flow driving needed to explain the observed zonal-mean circulation. The significance of this gravity wave driving of the mean flow is now generally acknowledged.

A similar issue had also arisen in numerical weather prediction models of the lower atmosphere, where it was shown that systematic errors in forecasts could be reduced by incorporating the effects of subgrid-scale gravity wave drag using an appropriate parameterization scheme that describes the effect of orographic waves, that is, those generated primarily by flow over topography with phase speed of zero (Palmer et al. 1986; McFarlane 1987).

Gravity waves are now widely recognized as important phenomena that redistribute momentum and energy in the atmosphere through their generation, propagation, and dissipation (Fritts and Alexander 2003). Since the scale of these waves is generally too small for GCMs to resolve, much current effort is directed at understanding the details of gravity wave generation, propagation, and dissipation so that the wave effects can be parameterized and incorporated into GCMs as a fundamentally necessary component.

Flow over topography represents only one source of excitation for atmospheric gravity waves, and it is reasonably simple to characterize compared with more complicated sources such as convective systems, including tropical cyclones and squall lines (e.g., Sato 1993; Alexander et al. 1995; Chun and Baik 1998), and gravity wave radiation from balanced flows such as from jet-front systems (e.g., Plougonven and Zhang 2014). Transport of momentum by these “nonorographic” gravity waves is especially important in the upper stratosphere to mesosphere as described above, throughout the summer middle atmosphere where orographic gravity waves cannot propagate through the weak
winds in the lower stratosphere and also provide significant contributions to the tropical QBO and SAO (see section 7).

Observational knowledge of gravity waves has been advanced by the development of high-resolution instruments such as mesosphere–stratosphere–troposphere (MST) radars and by the analysis of high-resolution radiosonde data in the 1980s and later. Gravity waves were examined in terms of horizontal and vertical wind fluctuation spectra and vertical momentum flux spectra (e.g., Nastrom and Gage 1985; Tsuda et al. 1989; Allen and Vincent 1995; Sato et al. 2017). These spectra, particularly power spectra, often approximate an idealized universal spectrum (VanZandt 1982) analogous to the universal spectra observed in the ocean (Garrett and Munk 1972). These observations suggest that the gravity wave field in the middle atmosphere may be shaped by some nonlinear saturation process. The growth of wave amplitudes with height as the mean density decreases indeed suggests that nonlinear saturation must occur at sufficiently high altitude (e.g., Lindzen 1981; Fritts 1984). Several theoretical approaches have been advanced to explain the shape of the universal spectrum based on a simple concept that internal gravity waves will produce a local gravitational instability when they reach sufficient amplitude and then may break (in a manner analogous to the familiar breaking of surface gravity waves), and that this process may limit further growth of amplitude with height (e.g., Smith et al. 1987). This basic idea of how an individual vertically propagating plane gravity wave might break suddenly at some altitude has been the basis of many parameterizations of gravity wave effects in models. However, a more thorough consideration of the nonlinear dynamics of gravity wave propagation and dissipation suggests that the idealized saturated gravity wave concept may not adequately describe the actual behavior of the middle atmospheric gravity wave field (e.g., Dossier and Sutherland 2011; Fritts et al. 2015).

Observations of the gravity wave field from radars, balloons, and rockets have the limitation of vertically sampling the atmosphere at a single geographic location. Nowadays such observations can be supplemented with high-resolution satellite observations (e.g., Alexander 1997, 1998). Satellite platforms allow global or near-global coverage, but the limitations of space and time resolution of satellite measurements permits observations of only part of the gravity wave spectrum (Alexander 1998). Within these limitations various satellite observations have been used to characterize global distributions and seasonal and interannual variations of the gravity wave field in terms of wave energy, momentum fluxes, and phase structure (e.g., Geller et al. 2013; Alexander 2015; Gong et al. 2015; Wright et al. 2017; Ern et al. 2018). Quantities such as energy and momentum fluxes are typically derived by combining the direct observation of temperature from satellites with the results of linear gravity wave theory. Figure 27-12 (from Ern et al. 2018) shows a recent example of such an observational estimate of the global distribution of total gravity wave momentum fluxes based on SABER satellite observations of temperature. In addition, recently developed superpressure balloon technology allows estimates of gravity wave characteristics in the intrinsic frequency space (Hertzog et al. 2008; Podglajen et al. 2016).

**Fig. 27-12.** Global distributions of gravity wave absolute momentum fluxes at 30 km altitude in (left) January and (right) July from SABER. [Adapted from Fig. 13 of Ern et al. (2018), copyright Ern et al. 2018, https://creativecommons.org/licenses/by/4.0/.]
The earliest global circulation models that included domains extending to the mesopause were able to demonstrate the effects of gravity waves on the global middle atmospheric circulation (Miyahara et al. 1986; Hayashi et al. 1989; Sato et al. 1999). These studies confirmed that the simple ideas of gravity wave filtering and the wave effects on the zonal-mean circulation of the middle atmosphere advanced by Lindzen (1981) and Matsuno (1982) were indeed operative in comprehensive global models. However, these models could only explicitly represent the long horizontal wavelength end of the gravity wave spectrum (wavelengths greater than a few hundred kilometers). As computer power has improved, ultra-fine-resolution regional and global models have been used for the explicit simulation of the gravity wave field (Watanabe et al. 2008; Plougonven et al. 2013; Holt et al. 2017; Shibuya and Sato 2019). The results from these models can help provide constraints on gravity wave parameterizations used in more moderate-resolution climate models (Kim et al. 2003; Alexander et al. 2010; Geller et al. 2013).

Standard approaches to parameterizing gravity wave effects treat the problem as a single vertical column, which essentially ignores the horizontal propagation of the subgrid-scale gravity waves across grid points. However, given the high altitude of the mesosphere, three-dimensional gravity wave propagation may be important (Smith 1980; Dunkerton 1984; Marks and Eckermann 1995; Sato et al. 2009; Ehard et al. 2017). Another process that may be important for the middle atmospheric zonal-mean circulation is the secondary generation of gravity waves (Bacmeister and Schoeberl 1989; Satomura and Sato 1999; Vadas et al. 2003; Bossert et al. 2017; Yasui et al. 2018; Becker and Vadas 2018) and Rossby waves (Ern et al. 2013; Sato and Nomoto 2015) caused by the breaking of the “primary” waves propagating from the lower atmosphere.

In summary, it is now clear that moderate-resolution climate and weather prediction models that include the middle atmosphere require a parameterization of the momentum transports by the subgrid-scale gravity wave field in order to produce realistic simulations, and these must be constrained by the observations (Alexander 2010). While significant progress has been made, much further improvement is required (Geller et al. 2013).

7. The quasi-biennial oscillation

By 1920 there had been some relevant, but rather scattered, observations of the winds in the region above the equatorial tropopause, beginning with the inference of strong prevailing easterlies to explain the spread of volcanic aerosol following the August 1883 eruption of Mt. Krakatau [e.g., Wexler 1951; see also the review by Baldwin et al. (2001)]. In 1908 visually tracked “pilot” balloon observations by Arthur Berson in East Africa indicated the presence of westerlies, at least in the lowest few kilometers of the stratosphere, and subsequent widely scattered pilot balloon observations throughout the tropics sometimes showed the presence of westerlies and sometimes easterlies. The fragmentary observational record was interpreted as indicating that the winds were dominated by strong and fairly steady zonal jets. In particular the winds above the low-latitude troposphere were thought to be dominated by prevailing “Krakatoa easterlies” at most heights and latitudes but that a narrow, possibly meandering, “thread” of westerlies (“Berson westerlies”) was also present (e.g., Labitzke and van Loon 1999).

In the early 1950s daily balloon rawinsonde (a radiosonde whose position is tracked by radio techniques as it ascends to give wind speed and direction) observations began at several tropical Pacific islands. Examination of these more systematic measurements over several years revealed the existence of dramatic interannual variations in the equatorial stratospheric circulation (Sadler 1959; McCreary 1959; Graystone 1959; Ebdon 1960), and finally Reed et al. (1961) and Ebdon and Veryard (1961) demonstrated that these variations took the form of nearly repeatable cycles with a period close to 2 years. The earliest investigators necessarily had an incomplete understanding of the nature and variability of the oscillation period. Some early papers referred to a “biennial oscillation,” others to a “26-month oscillation.” The now-standard terminology “quasi-biennial oscillation” was introduced by Angell and Korshover (1964).

Figure 27-13 shows a height–time section of the monthly averaged, zonally averaged zonal wind at the equator from the upper troposphere to the middle stratosphere for the period since 1980. The wind values are computed from modern reanalysis data, but a very similar picture emerges when rawinsonde observations at even just a single near-equatorial station are used and when the record is extended back to 1953 (Naujokat 1986). Throughout the record (at least until 2016) an oscillation between prevailing easterlies and westerlies with transitions occurring very roughly once each year is apparent at all levels from near the tropopause (∼100 hPa) to the highest level shown (3 hPa). The zonal wind transitions appear to originate at high levels and propagate downward. In fact, the apparent downward propagation of features is almost omnipresent in the observed height–time zonal wind record, with the most notable exception occurring in 2016.

The existence of the QBO in the prevailing equatorial zonal wind was soon followed by observations showing
that the QBO has an effect on the temperature and ozone concentrations in low latitudes as well (e.g., Ebdon and Veryard 1961; Reed 1962; Funk and Garnham 1962; Ramanathan 1963; Angell and Korshover 1964). The QBO signal in the winds was shown to be equatorially centered with a half-width in the amplitude of roughly 12° of latitude (Reed 1965).

The discovery of the QBO was a complete surprise to the meteorologists of the time, and for some years the basic dynamics of the phenomenon eluded understanding. The key breakthrough was made by LH68, who posited that the mean-flow accelerations in the stratospheric QBO were driven by interactions between the zonal-mean flow and a spectrum of zonally and vertically propagating gravity waves generated in the troposphere. Just before this, Booker and Bretherton (1967) had shown that vertically propagating gravity waves will be strongly absorbed near critical levels where the mean flow equals the wave horizontal phase speed. LH68 showed that this implied that waves with eastward (westward) phase speed should produce westerly (easterly) mean flow accelerations in regions of westerly (easterly) vertical shear. This mechanism is thus able to produce equatorial mean flow accelerations in both zonal directions. However, the mechanism is self-limiting in the sense that once a region of strong westerlies (easterlies) has been formed it will effectively filter out waves with eastward (westward) phase speed from reaching higher altitudes.

LH68 incorporated this basic idea into a simplified numerical model of the height and time dependence of the equatorial mean flow and showed that the model could explain the slow QBO mean flow changes (and their downward propagation) through the impact of high-frequency upward-propagating waves that were expected to be generated by convection and other sources in the tropical troposphere. This striking result led to the general acceptance of the LH68 mechanism as the basis for understanding the QBO dynamics. This view was reinforced by the elegant “QBO-analogue”
laboratory experiment of Plumb and McEwan (1978), who investigated the interaction of the zonal flow around an annulus filled with salt-stratified fluid with propagating gravity waves forced at the lower boundary.

One limitation of the LH68 model is that mean flow accelerations are simply proportional to the mean flow shear and so there is no mechanism to account for observed accelerations in regions with weak shear (such as right at the jet maxima and also at times in the lowermost stratosphere). Holton and Lindzen (1972, hereafter HL72) generalized the LH68 model to incorporate the effects of planetary-scale equatorial waves with sufficiently slow vertical group velocity that they are significantly damped in the stratosphere by dissipative processes, including radiative transfer. The development of the HL72 theory followed the discovery in stratospheric data of prominent equatorial waves, including Kelvin waves (Wallace and Kousky 1968) and Rossby–gravity waves (Yanai and Maruyama 1966). HL72 showed that a simple model of the effects of large-scale Kelvin and Rossby–gravity waves could account for a reasonably realistic looking QBO of the equatorial mean zonal winds when including wave forcing with specified parameters close to those determined from observations. However, further analysis of the observed wave field suggests that the large-scale planetary waves are actually not strong enough to explain all the observed QBO accelerations (Lindzen and Tsay 1975), and it seems likely that both planetary waves and high-frequency gravity waves contribute significantly to the driving of the QBO (Dunkerton 1997).

As noted earlier, the QBO displays some cycle-to-cycle variability (e.g., Pascoe et al. 2005). The tendency for the descending easterly shear zone to “stall” near 20 hPa is apparent in the observed record (see Fig. 27-13), and the duration of this stalling appears to account for the most obvious differences among cycles. The QBO is not a subharmonic of the annual cycle, but there is evidently some subtle connection of the observed QBO with the seasonal cycle. Notably those QBO cycles that show almost no “stalling” appear to have a near 2-yr duration (e.g., 1998–99, 2006–07) while other cycles with the longest periods of stalling are almost 3 years in length (e.g., 2000–02). It has been hypothesized that the QBO is affected by the forcing from quasi-stationary planetary Rossby waves excited largely in the winter NH that can propagate into the tropics and that this could help account for the apparent “synchronization” with the annual cycle seen at times in both the tropical wind signal (Dunkerton 1983) and the corresponding ozone signal (Gray and Dunkerton 1990). The QBO is expected to modulate the meridional penetration of quasi-stationary planetary Rossby waves into the tropical stratosphere—specifically, one expects the propagation to be much more effective through mean westerlies than easterlies. This effect should lead to a modest zonal inhomogeneity in the QBO wind oscillation. Indeed, a roughly 10% zonal asymmetry in QBO amplitude around the equator at some levels has been observed in station rawinsonde data as well as in reanalysis data and comprehensive model simulations (Hamilton et al. 2004). An additional possibility is that the annual modulation of mean upwelling in the equatorial lower stratosphere could also contribute to the tendency for the QBO to synchronize with the annual cycle (Kinnersley and Pawson 1996; Hampson and Haynes 2004; Rajendran et al. 2018).

While the QBO has been notable for its overall regularity at least over the first ~27 cycles since regular observations began in 1953, around the beginning of 2016 the regular pattern was rather obviously disrupted (Osprey et al. 2016; Newman et al. 2016). As can be seen here in Fig. 27-13, the descending easterly shear zone completely stalled near 20 hPa and the usual easterly phase was aborted as the westerlies near 20–30 hPa persisted for almost 2 years (mid-2015 to mid-2017). Wind features displayed upward apparent propagation in the 20–50-hPa layer over the first half of 2016. It seems that more typical QBO behavior was restored by 2017, but the wind evolution through 2016 was unprecedented in the era of detailed observations. The prediction and predictability of this very anomalous event is a topic of ongoing research (Watanabe et al. 2018).

Reed (1965) noted that the QBO temperature perturbations should lead to a QBO in radiative heating and hence a QBO secondary circulation in the meridional plane. This effect was incorporated into a numerical model by Plumb and Bell (1982), who noted that the advection associated with this circulation will contribute to the strengthening (weakening) of the westerly (easterly) shear zones, at least near the equator, and could thus account for the obvious asymmetry between the strength of westerly and easterly shear zones (see Fig. 27-13). This effect should also lead the westerly mean flow accelerations to be concentrated close to the equator over much of the QBO cycle, and indeed this feature was later observed to be characteristic of the observed evolution of the mean zonal winds (Hamilton 1984; Dunkerton and Delisi 1985).

Advection by the QBO in equatorial upwelling also gives rise to a QBO in column amounts of ozone (Reed 1964) and the secondary meridional circulation produces a latitudinal structure as the induced circulation descends across the transition height from the middle and upper stratosphere where ozone is controlled by fast temperature-dependent chemical reactions to the lower stratosphere where its lifetime is much longer and is
hence dynamically controlled (Gray and Pyle 1989). Corresponding QBO distributions were subsequently predicted in many other trace constituents (Gray and Chipperfield 1990) and volcanic aerosol distributions (Trepte and Hitchman 1992).

While the equatorially trapped nature of the low-latitude QBO was established in early observational studies and the QBO-induced meridional circulation is essentially confined to subtropical latitudes, notable QBO signals have been observed at higher latitudes in both dynamical quantities and ozone. A key development was the study of Holton and Tan (1980, 1982), who showed that the NH polar stratosphere tends to be warmer (and the polar vortex weaker) on average in winters when the equatorial QBO near 50 hPa is in its easterly phase. Quasi-stationary planetary Rossby waves (see section 3) tend to be stronger in the polar winter stratosphere in the easterly QBO phase as well. Following the original Holton and Tan work, a number of observational and model studies of the effects of the QBO in the extratropical winter stratospheric circulation in both hemispheres have been performed (e.g., Baldwin and Dunkerton 1998; Anstey et al. 2010).

It appears that there are also teleconnections of the stratospheric QBO with aspects of tropospheric circulation. Ebdon (1975) and Holton and Tan (1980) documented a systematic effect of QBO phase on the sea level pressure in the extratropical NH winter. This issue has been examined in many subsequent studies, motivated, in part by a desire to use information about the state of the QBO in extended-range weather forecasts (see section 15; also Coughlin and Tung 2001; Thompson et al. 2002; Marshall and Scaife 2009; Garfinkel et al. 2018; Gray et al. 2018). Additionally, direct QBO influences at the surface in tropical latitudes, such as via tropical convection, have also been investigated (e.g., Collimore et al. 2003; Gray et al. 2018). This is discussed further in section 15.

Early comprehensive global atmospheric models (and coupled GCMs)—even those with considerable numerical resolution in the stratosphere—produced simulations that completely lacked the QBO (e.g., Manabe and Hunt 1968; Fels et al. 1980). Free-running GCMs typically simulated quite steady zonal-mean easterlies in the stratosphere and the winds in experiments initialized with realistic mean flow profiles were found to relax quickly to steady easterlies (Hamilton and Yuan 1992). Takahashi (1996) first reported a large QBO-like interannual variation in the simulated tropical stratosphere of a comprehensive atmospheric GCM. Takahashi obtained this result by running a fairly standard spectral GCM without any subgrid-scale gravity wave parameterization (see section 6), but with quite fine vertical resolution (vertical level spacing ~500 m) through the stratosphere and subgrid-scale diffusion coefficients reduced substantially over the values typically used. Takahashi’s discovery was soon followed by studies showing that spontaneous long-period mean wind oscillations (QBO-like phenomena) did develop in at least two other simplified atmospheric GCMs (Horinouchi and Yoden 1998; Hamilton et al. 1999). Since then (starting with Scaife et al. 2000) there have been a number of investigators who have simulated QBOs of various degrees of verisimilitude by incorporating parameterizations of nonorographic gravity waves into their GCMs [see section 6 for more discussion of gravity wave parameterizations, and see Butchart et al. (2018) for a list and brief description of 17 state-of-the-art global models that simulate the stratospheric QBOs].

While a great deal of progress has been made in understanding and modeling the QBO, many issues remain under active research. At the most basic level it is still somewhat unclear how the QBO wind reversals occur in the lowermost stratosphere where the QBO amplitude is very weak. In the original LH68 and HL72 models the transitions were crucially dependent on the effects of an assumed mean flow momentum mixing acting near the tropopause where the mean flow was assumed to be zero. Saravanan (1990) generalized the HL72 model to include a representation of a mean vertical advection. Comprehensive model studies suggest the advection by the mean equatorial upwelling of the BDC does indeed play an important role in the evolution of the QBO (e.g., Kawatani and Hamilton 2013), but this is an issue that will continue to attract attention (e.g., the recent work of Bui et al. 2019), particularly as changes to the strength of the BDC appear to be a key aspect of the global response to increased greenhouse forcing (Butchart 2014; see section 2).

Understanding the cycle-to-cycle variability in the QBO record is also a topic of great current interest. There have been attempts to determine if cycle-to-cycle variation in the QBO may have systematic physical causes, notably including the changes in tropospheric circulation and convection during different phases of the Southern Oscillation (e.g., Taguchi 2010). The very anomalous behavior seen in 2016 has raised interest in understanding the range of possible extreme departures from normal behavior and the implications for the predictability of the QBO. Proxy data are being applied in attempts to extend the record of QBO phases further back into the past (Brönnimann et al. 2016) and even to investigate the possibility that the QBO may have been absent at times (Hamilton and Garcia 1984).

In recent years an extensive effort to analyze and intercompare the QBO dynamics in many state-of-the-art GCMs has begun under the aegis of the World Climate
Research Programme (WCRP) Stratosphere–Troposphere Processes and their Role in Climate (SPARC) project (Butchart et al. 2018). This initiative has the potential to improve the representation of the QBO within GCMs and to help answer remaining questions about QBO dynamics and predictability.

8. The tropopause

The tropopause (Fig. 27-14) represents the interface (or perhaps more appropriately, the interfacial layer) that couples the troposphere to the stratosphere. This coupling takes place both upward and downward. For example, as discussed in section 3, much of the wave driving of the middle atmosphere originates in the troposphere; these waves therefore have to propagate through the tropopause. On the other hand, dynamical processes in the tropopause region couple to dynamics near the surface. Downward ozone fluxes through the tropopause influence near-surface air quality and affect radiative forcing. In the tropics, the tropopause represents the “gateway to the stratosphere” (via the upwelling branch of the BDC) and processes in the tropopause layer ultimately determine the water vapor content of the entire stratosphere, among other things.

The discovery of the stratosphere, and with it the tropopause, represents a prime example of discovery due to scientific curiosity and adventurism. As discussed in section 1, the temperature difference between mountain tops and valleys (roughly 7 K km⁻¹) had long been recognized. If the temperature decreases with height at this rate, this would lead to absolute zero at an altitude of about 40 km (assuming a surface temperature of 280 K). This suggests that the temperature lapse rate either reduces severely at high altitudes or that the top of atmosphere is reached below about 40 km altitude. Early explorers in the nineteenth century traveled on board hot air balloons and confirmed the steady drop-off of temperature up to around 10 km (Glaisher 1871). Measurements at higher altitudes could only be obtained using unmanned balloons, and this became possible toward the end of the nineteenth century. On 28 April 1902, Leon Teisserenc de Bort announced the discovery of an “isothermal layer” to the French Academy of Science (“zone isotherme” in his original French report; Teisserenc de Bort 1902) at altitudes between 8 and 13 km. Three days later, Richard Assmann announced the discovery of “a warmer air flow at heights from 10 to 15 km” to the Prussian Academy of Science (Assmann 1902). Both scientists stood in close collaboration and had agreed to announce their discovery at the same time. The layer they jointly discovered represents the first few kilometers of what we now call the stratosphere (a term that was introduced by Teisserenc de Bort around 1908). Sir Napier Shaw later introduced the term tropopause for the interface between troposphere and stratosphere; for example, in his 1920 *Manual of Meteorology* he refers to the tropopause as the “layer of the atmosphere which marks the outer limit of the troposphere and the lower limit of the stratosphere. Subject to reservations, the tropopause may be regarded as a surface; but the transition is not always so abrupt as to produce real discontinuity, and it is therefore convenient to use the word tropopause to connote the phenomena of the region of transition from the troposphere to the stratosphere. The phenomena may include a sudden transition to nearly isothermal conditions, a counterlapse leading to isothermal conditions, or a gradual transition from a lapse-rate which is near the adiabatic to a condition approximately isothermal” (Shaw 1936, p. xxxvii).

Modern high-resolution balloon observations, as well as other high-vertical-resolution temperature data such as from global positioning satellite radio occultations, have shown that the troposphere–stratosphere transition is extremely sharp on average (Birner et al. 2002; Randel et al. 2007). Furthermore, Assmann’s “upper inversion” or Shaw’s “counterlapse” case, which corresponds to a layer of enhanced thermal stratification (i.e., a tropopause inversion layer; Birner 2006), turns out to be the climatological behavior on a global scale (Grise et al. 2010).

---

Fundamentally, the tropopause exists because of the combined effects of dynamical lapse rate control in the troposphere and radiative lapse rate control in the stratosphere (Held 1982). For example, the tropopause height can be viewed as the result of a given tropospheric lapse rate (e.g., set by convection in the tropics and large-scale eddy fluxes in the extratropics), a given surface temperature, and a stratosphere in radiative equilibrium (e.g., Manabe and Strickler 1964). However, the stratosphere is clearly not in radiative equilibrium: upwelling by the BDC provides adiabatic cooling in the tropics, likewise downwelling by the BDC provides adiabatic warming in the extratropics. These tendencies lead to an elevated tropopause in the tropics and a lowered tropopause in the extratropics, compared to a stratosphere in radiative equilibrium, and thereby significantly increase the equator-to-pole contrast in tropopause height (Birner 2010).

The discussion so far has concentrated on the thermal structure of the atmosphere and the resulting thermal tropopause. For transport studies a more appropriate tropopause definition is based on the PV field (see section 3); specifically, a particular PV isosurface can be defined as a dynamical tropopause, which represents a material surface for adiabatic, frictionless flows. Typical values used for the dynamical tropopause are 1.5–4 PVU (1 PVU = 10^{-5} K kg^{-1} m^{2} s^{-1}), with the 2 PVU isosurface perhaps used most commonly. The differences between these PV isosurfaces are small in most situations due to the near-discontinuous PV contrast between the troposphere and stratosphere (e.g., Kunz et al. 2011). This tropopause definition has proven very useful for stratosphere–troposphere exchange studies (e.g., Gettelman et al. 2011 and references therein). In addition, tropopause definitions based on quasi-conserved tracers have been employed [e.g., ozone (Bethan et al. 1996) or idealized tracers such as E90 in models (Prather et al. 2011)]. The above three tropopause definitions are illustrated in Fig. 27-14 using the zonal-mean atmospheric structure for an example month (April 2012).

9. Stratosphere–troposphere exchange

Stratosphere–troposphere exchange (STE) describes an air mass or constituent flux across the tropopause and is primarily upward in the tropics and downward in the extratropics. The magnitude of exchange of STE was long thought to be dependent on small-scale processes (such as tropopause folding events), but a new theoretical framework was put forward in the seminal review on stratosphere–troposphere exchange by Holton et al. (1995). They showed that “wave-induced forces drive a kind of ‘fluid dynamical suction pump,’ which withdraws air upward and poleward from the tropical lower stratosphere and pushes it downward into the extratropical troposphere.” STE is thereby important for the chemical composition of both the stratosphere and the troposphere. Many long-lived trace gases that are inert in the troposphere are photolyzed in the stratosphere, where they can cause ozone depletion [in particular CFCs, water vapor, and nitrous oxide (N_{2}O)]. On the other hand (as mentioned above), the transport of stratospheric ozone into the troposphere can affect radiative forcing and air quality. However, the realization of the importance of STE came about only gradually and the theoretical framework for STE was long disputed.

The tropical tropopause can be seen as the gateway into the stratosphere, an understanding that developed along with the discovery of the stratospheric general circulation. As described briefly in section 1, early progress was made by Brewer, whose investigation of cirrus cloud formation from contrails during World War II led to the discovery of the dryness of the stratosphere. Brewer (1949) inferred from these measurements that the air must have passed into the stratosphere through the very cold tropical tropopause, where water vapor concentrations could be reduced to observed values by the process of ice crystal formation. Subsequent aircraft, balloon, and rocket observations of stratospheric water vapor using different measurement techniques between the 1950s and 1970s were hard to interpret due to both representativeness (time of year, altitude, and latitude) and accuracy issues [see reviews by Harries (1976) and Robinson (1980)]. However, Mastenbrook (1974) and Kley et al. (1979) developed Brewer’s hypothesis further, using concomitant water vapor and temperature observations, and correctly proposed that dehydration occurs primarily in specific geographical regions. Newell and Gould-Stewart (1981) postulated that observed stratospheric water vapor values could be explained if the region over the Indonesian continent, where the very lowest tropopause temperatures were measured, acts as a “stratospheric fountain” through which air enters the stratosphere from the troposphere.

In the 1990s, it was recognized that overshooting convection could affect lower stratospheric water vapor locally (e.g., Kelly et al. 1993). However, Highwood and Hoskins (1998) showed that such penetration would be relatively limited. It also became clear that the seasonal cycle in cold-point tropopause temperatures, determined by the seasonally varying strength of the stratospheric circulation (Yulaeva et al. 1994), leads to the seasonal cycle observed in water vapor (Rosenlof 1995). This cycle is imprinted in air as it slowly rises upward through the tropical tropopause, as revealed in the “tropical tape recorder” derived from satellite measurements (Holton...
et al. 1995; Mote et al. 1996; Fig. 27-15). Holton and Gettelman (2001) further highlighted that slow horizontal (and not vertical) advection through regions of very low temperatures over the tropical western Pacific could explain the overall dryness of the stratosphere, rebutting the idea of a localized pathway into the stratosphere in this region. This idea was further corroborated by detailed trajectory analyses from within the tropical tropopause layer (TTL), which could explain not only the minimum value in stratospheric water vapor, but also its interannual variability (Fueglistaler et al. 2005; Fueglistaler and Haynes 2005). The TTL is thereby defined as a transition region that exhibits typical characteristics of both the tropical troposphere and the stratosphere and encompasses roughly the region between 150 and 70 hPa (Fueglistaler et al. 2009).

As highlighted already in Robinson (1980), the tropical gateway from the troposphere into the stratosphere also allows for other trace gas species to enter the stratosphere. Importantly, very short-lived halogenated substances, which contribute to ozone depletion, and precursors of aerosol, which affect the radiative budget of the region, are also brought into the stratosphere through this pathway.

We now turn to the extratropics, where STE was (historically) mostly important for the impacts on the troposphere. Between the late 1950s and early 1960s, the United States and the Soviet Union conducted high-altitude tests of nuclear weapons, assuming that radioactive contamination of the higher atmosphere would not affect the troposphere (see also the discussion in section 2 on the BDC). However, radioactive isotopes were detected in milk at NH high-latitude locations soon after the explosions (e.g., Telegadas and List 1964). The findings spurred research using rhodium, cadmium, and plutonium isotope measurements to derive transport pathways and time scales within the stratosphere (Kalkstein 1962). Key aspects of the BDC as we know them today were inferred, such as the ascending motion over the tropics, a strong descending motion within the polar vortex during winter, and strong mixing between the tropics and extratropics at 18–25 km (List and Telegadas 1969; see section 2 for subsequent developments). The ultimate transport mechanisms that brought the radioactive debris into the troposphere were thought to be associated with intense baroclinic zones in the vicinity of the jet stream (Reed and Sanders 1953; Reed 1955; Danielsen 1959; Reiter 1962, 1963; Mahlman 1965). These were characterized as tropopause folding events by Reed and Danielsen (1959). Additional mechanisms for STE put forward were the seasonal change in tropopause altitude (Staley 1962) and turbulent mixing (e.g., Libby 1956), as summarized in the review by Reiter (1975).

In the 1980s and 1990s, dedicated aircraft campaigns making observations of ozone, water vapor, other long-lived trace gases, and turbulence revealed more detailed evidence that turbulent mixing processes in the vicinity of tropopause folds and cutoff lows were of first-order importance as a mechanism of STE (Shapiro 1980; Ebel et al. 1991; Vaughan et al. 1994; Browell et al. 1998), and also that mesoscale convective complexes and thunderstorms could lead to downward mixing of stratospheric air into the upper troposphere (Poulida et al. 1996). Importantly, it was also realized that STE has a substantial upward, that is, troposphere-to-stratosphere, component at midlatitudes as well as in the tropics. Dessler et al. (1995) inferred from aircraft measurements of water vapor that not all air located in the lowermost stratosphere could have entered the region via the tropical cold-point tropopause, a finding corroborated by satellite observations (Pan et al. 1997).

Model-driven approaches helped conclude that stirring

![Figure 27-15](attachment:figure2715.png)
of tropospheric and stratospheric air masses—ultimately accomplished by Rossby wave breaking events—must be responsible for the exchange (Chen 1995; Appenzeller et al. 1996; Peters and Waugh 1996).

Research in the 2000s and 2010s was aimed at obtaining a more comprehensive understanding of the impact of two-way mixing on trace gas distributions in the tropopause region (Fischer et al. 2000; Hoor et al. 2002) and its variability on seasonal and interannual time scales. Dedicated STE aircraft campaigns (e.g., Zahn et al. 2004; Engel et al. 2006; Pan et al. 2007) led to the understanding that individual mixing processes lead to a mixing layer across the tropopause, now generally referred to as the extratropical tropopause transition layer (ExTL). Satellite observations have confirmed the ExTL to be a global phenomenon (Hegglin et al. 2009) and revealed interhemispheric differences, with a deeper ExTL in the NH than in the SH, consistent with the more frequent Rossby wave breaking events (Hitchman and Huesmann 2007).

10. Stratospheric composition

Distributions of trace gases in the stratosphere reflect the combined effects of atmospheric transport and photochemistry. Stratospheric composition measurements are key tools in exploring the variability and change in the dynamics of the stratosphere, for which direct measurements are often not available. In fact, many of the theoretical advances—including the BDC (see section 2; Brewer 1949; Dobson et al. 1929; Dobson 1956), the tropical pipe (Plumb 1996), and the leaky pipe (Neu and Plumb 1999)—were driven by peculiarities observed in chemical constituent distributions in the stratosphere (see also sections 2 and 9). The need for detailed knowledge of stratospheric composition also grew along with the realization of the potential harmful impacts of human-made substances on the ozone layer (Robinson 1980). This was driven early on by theoretical consideration that human-made substances found at Earth’s surface around the globe (Lovelock 1972) could adversely affect the stratospheric ozone layer (see section 11; Molina and Rowland 1974). Additionally, the importance of stratospheric trace gases, particularly ozone, water vapor, and aerosol, for the radiative forcing of climate led to renewed interest in measuring changes in their stratospheric composition distributions (e.g., Solomon et al. 2010, 2011; Nowack et al. 2017).

Some of the first global observations of chemical constituents in the stratosphere were obtained by pioneering satellite instruments such as the Limb Infrared Monitor of the Stratosphere (LIMS; Gille and Russell 1984) and the Stratospheric and Mesospheric Sounder (SAMS; Jones et al. 1986). These observations revealed for the first time the full effects of transport and chemistry on the stratospheric distributions of trace gases (Jones and Pyle 1984). The general structure of long-lived trace gases [such as nitrous oxide (N₂O) and methane (CH₄)] as obtained from limited in situ measurements was confirmed, with concentrations generally decreasing with height, but also along constant pressure levels toward the poles (see left panel in Fig. 27-16). This reflects that the sources of these gases are found in the troposphere and their main photochemical sinks in the stratosphere. On the basis of combined monthly mean observations of CH₄ and H₂O from the SAMS and LIMS satellites, Jones et al. (1986) confirmed the hypothesized source of H₂O from oxidation of CH₄ in the stratosphere (Robinson 1980), along with the robustness of the feature of a minimum in H₂O, referred to as the hygropause, found previously just above the tropical tropopause (Russell et al. 1984) (cf. middle panel Fig. 27-16). Indications for additional dynamical influences such as the SAO on tracer transport were also obtained by these early measurements (Gray and Pyle 1986), and they were used to make early quantifications of the BDC (Solomon et al. 1986; Holton and Choi 1988).
In the 1990s and 2000s, evaluation of satellite limb sounder observations profited from the rapidly growing theoretical knowledge of stratospheric dynamics and vice versa (see also section 2). Gray and Pyle (1989) identified the QBO to be a dominant source of interannual variability in the vertical ozone structure (see Fig. 27-17), with separate contributions to the total column ozone QBO signal coming from the upper stratosphere, where ozone is controlled by temperature-dependent chemical reactions, and from the lower stratosphere, where it is controlled by dynamical transport processes (see also sections 7 and 11). The QBO was also identified as one of the main drivers of variability in the stratospheric trace gas distributions of N2O and CH4, apart from a distinct seasonal cycle that is driven by the BDC (Gray and Chipperfield 1990; Randel et al. 1998). Variability in the SH winter polar vortex was linked to springtime ozone depletion (Schoeberl and Hartmann 1991). Leovy et al. (1985) and later Randel (1993) provided direct observational evidence of planetary wave breaking in the surf zone from trace gas observations. Neu et al. (2003) used probability density functions (PDFs) of satellite measurements to identify the boundaries between tropical and extratropical air and identified this subtropical “edge” as transport barrier (as discussed in section 3; see Fig. 27-2). The first recorded SH SSW was immediately revealed by total column ozone observations (Varotsos 2002). All these examples demonstrate that long-lived trace gas observations are valuable indicators of transport processes and as a consequence they are widely used to benchmark the transport (and chemistry) in CCMs (e.g., Prather and Remsberg 1993; Garcia et al. 1992; Eyring et al. 2006; Heglin et al. 2010; Strahan et al. 2011).

11. Stratospheric ozone

Following the discovery of ozone in 1839 by Schönbein at the University of Basel Switzerland, Houzeau in Rouen, France, showed in 1858 that ozone is prevalent in the atmosphere. In 1880 Hartley concluded that the strong atmospheric absorption of solar UV radiation

![Fig. 27-17. The effect of the QBO on ozone anomalies as derived from multiple instrument monthly averaged, zonally averaged ozone climatologies from the SPARC Data Initiative.](From SPARC (2017).)
between 200 and 320 nm observed by Cornu (1879) was associated with ozone and proposed therefore the presence of large amounts of ozone in the upper atmosphere. His findings were supported later by the UV measurements of Fabry and Buisson in 1913 in Marseilles, France. As described in section 1, a further milestone was the network of UV spectrophotometers by Dobson in the 1920s (Dobson 1931) that allowed, in conjunction with the Umkehr method developed by Götz, the retrieval of information on the vertical ozone profile. The altitude of the ozone maximum at about 22 km derived from the Umkehr method (Götz et al. 1934) was lower than previously assumed, but supported by the first in situ spectroscopic measurements from balloons in Germany by E. and V. H. Regener in 1934 and by the U.S. Explorer II mission in 1935.

Stratospheric ozone is formed naturally by photochemical reactions that require ultraviolet sunlight (Chapman 1930). In the first step, an oxygen molecule (O₂) is broken into two oxygen atoms (O) by the absorption of solar UV radiation [Eq. (27-6)]. In the second step, each of the oxygen atoms combines with an oxygen molecule in a three-body-reaction to form an ozone molecule [Eq. (27-7)]. Ozone is photolyzed into an oxygen molecule and oxygen atom by absorption of solar UV radiation [Eq. (27-8)]. The oxygen atom can recombine with an oxygen molecule to reform ozone [Eq. (27-7)], or react with an ozone molecule to produce two oxygen molecules [Eq. (27-9)]:

\[
\begin{align*}
    &\text{O}_2 + h\nu \rightarrow \text{O} + \text{O} \quad \lambda \leq 242 \text{ nm}, \tag{27-6} \\
    &\text{O}_2 + \text{O} + \text{M} \rightarrow \text{O}_3 + \text{M}, \tag{27-7} \\
    &\text{O}_3 + h\nu \rightarrow \text{O}_2 + \text{O} \quad \lambda \leq 1180 \text{ nm}, \tag{27-8} \\
    &\text{O}_3 + \text{O} \rightarrow 2\text{O}_2. \tag{27-9}
\end{align*}
\]

Reaction (27-9) constitutes a destruction of stratospheric ozone. However, the simple Chapman mechanism does not explain the observed concentrations in the stratosphere and mesosphere. Studies conducted after 1950 showed that ozone destruction can be catalyzed by different chemical species present in the atmosphere. One of these species is the hydroxyl radical (OH), produced in the upper atmosphere by the photolysis of water vapor (Bates and Nicolet 1950) and in the stratosphere by the chemical reaction of water vapor with the electronically excited oxygen atom (O¹D). Other species that destroy ozone in the stratosphere are nitric oxide (NO), which forms in the stratosphere by the reaction of nitrous oxide (N₂O) with O¹D (Crutzen 1970), and chlorine monoxide (ClO) and bromine monoxide (BrO), which are produced primarily by the photolysis of halocarbons including anthropogenic chlorofluorocarbons and halons (Stolarski and Cicerone 1974; Wofsy et al. 1975). Each catalytic molecule can destroy thousands of ozone molecules before it is removed from the stratosphere.

The seasonal and latitudinal behavior of ozone is well documented from ground-based and space observations (e.g., Dütsch 1970, 1978; see also section 10). Figure 27-18 shows the climatological annual mean distribution of ozone derived from the Solar Backscatter Ultraviolet Radiometer (SBUV) on the Nimbus-7 satellite (Bhartia et al. 1996) between 1980 and 1989. The formation of ozone by the photolysis of molecular oxygen occurs primarily in the tropical middle stratosphere between about 25 and 30 km height. Ozone is then transported by the BDC toward the polar lower stratosphere, primarily during the winter season in each hemisphere. As a result, the vertically integrated total ozone column reaches its highest values at mid- to polar latitudes, as already observed by Dobson in the 1930s (Dobson 1963).

In the early 1970s, a potential threat to the ozone layer was recognized from the nitrogen oxides that would be released in the stratosphere by a projected fleet of supersonic aircraft (Johnston 1971). Three years later, Molina and Rowland (1974) suggested that the increasing consumption and related release in the atmosphere of industrially manufactured chlorofluorocarbons, in particular CFC-11 (CCl₃F) and CFC-12 (CCl₂F₂), provides the major source of reactive chlorine in the stratosphere, and therefore might lead to a substantial erosion of the ozone layer. These anthropogenic halocarbons have a
sufficiently long lifetime to be transported into the stratosphere, where they are converted to reactive halogen gases (Cl, ClO, ClONO₂, HCl). In the lower stratosphere, these gases reside primarily in the form of inactive reservoir gases (chlorine and bromine nitrate ClONO₂, BrONO₂, and hydrogen chloride HCl) but as they reach the mid- and upper stratosphere, they are converted to Cl and ClO radicals and catalytically destroy ozone molecules.

While the effects of halogen and nitrogen source gases on the ozone layer were soon recognized by the scientific community (and to some extent taken into account in policy), a completely unexpected, severe ozone decline was reported over the Antarctic continent in the mid-1980s based on ground-based measurements at the Antarctic research stations of Halley Bay (Farman et al. 1985) and Syowa (Chubachi 1984) and later from satellite observations (Bhartia et al. 1985). This continent-wide ozone depletion, soon named the Antarctic ozone hole, started to develop in the mid-1970s, and has remained a regular annual phenomenon appearing in late winter and early spring at southern polar latitudes (Fig. 27-19). The ozone loss over Antarctica is largest in the 10–20 km altitude range, where ozone is nearly totally depleted. This phenomenon could not be explained by the catalytic ozone depletion cycles that dominate higher up in the stratosphere. A sustained period of extensive research, including dedicated observational missions, subsequently demonstrated that the ozone hole is produced over Antarctica because of a unique combination of meteorological and chemical conditions that increase the effectiveness of ozone destruction by reactive halogens. Rapid ozone destruction as observed over Antarctica requires low temperatures to be present for an extended period of time so that large quantities of solid and liquid polar stratospheric clouds (PSCs) can be formed. As discussed in section 3, conditions with sufficiently low temperatures are rarely found in the Arctic, but they are frequent during the austral winter inside the polar vortex where the polar vortex can act as a “containment vessel” (see section 3) so that large areas containing nitric acid (HNO₃) or ice PSCs are observed and provide the conditions for the formation of an ozone hole (Crutzen and Arnold 1986; Toon et al. 1986). Heterogeneous chemical reactions on the PSC surfaces activate the chlorine and bromine provided by halogen reservoir gases. The release of the reactive radicals ClO and BrO leads to dramatic ozone destruction as soon as sunlight becomes available at the beginning of spring (e.g., Solomon et al. 1986; McElroy et al. 1986; Tung et al. 1986; Molina and Molina 1987; Anderson et al. 1989).

In the period 1979–97, when stratospheric reactive halogens, also known as ozone depleting substances (ODSs), were increasing to their highest concentrations, satellite instruments recorded a substantial decline of NH midlatitude ozone in the middle and upper stratosphere by about −7% decade⁻¹ (WMO 2014; Fig. 27-20, left). Atmospheric models, considering the chemical effects of ODSs on ozone, were able to reproduce the observed ozone decline in the upper stratosphere, thus confirming the causal relationship between ODSs and the observed ozone decline (e.g., Oman et al. 2010; WMO 2014) (Fig. 27-20, left). Global annual mean total ozone in the mid-1990s was about 5% below the 1964–80 average, with the ozone depletion being additionally enhanced by the effects of the Mt. Pinatubo volcanic eruption in 1991 (Fig. 27-21; see also section 13).

With growing evidence of the harmful effects of anthropogenic halogens on the ozone layer and the associated risks for life and human health resulting from enhanced surface-UV radiation, the World Meteorological Organization (WMO) developed in 1977 a World
Plan of Action on the Ozone Layer and conducted a series of international scientific ozone assessments. Following the Vienna Convention for the Protection of the Ozone Layer in 1985, the Montreal Protocol on Substances that Deplete the Ozone Layer was signed in 1987. The Montreal Protocol, which has been ratified by all 197 United Nations members, and subsequent amendments and adjustments (London, 1990; Nairobi, 1991; Copenhagen, 1992; Bangkok, 1993; Vienna, 1995; Montreal, 1997; Australia, 1998; Beijing, 1999; Kigali, 2016) successfully established legally binding controls for developed and developing nations on the production and consumption of halogen source gases.

As a result of the Montreal Protocol and its amendments, the overall abundance of ODSs in the atmosphere has been gradually decreasing since the late 1990s (Fig. 27-22), and upper stratospheric ozone has increased since the turn of the century (WMO 2014; Fig. 27-20, right). Global total ozone values are still lower than in the pre-1980 era, but they have ceased to decline and are slowly returning toward their mid-twentieth-century values (WMO 2018; Fig. 27-22). Through the implementation of the Montreal Protocol, much larger ozone depletion than currently observed has been avoided, specifically in the polar regions of both hemispheres (Chipperfield et al. 2015). Climate models with interactive ozone chemistry assuming compliance with the provisions of the Montreal Protocol project a return of global annual mean total ozone to values of 1980 shortly before 2050, while over Antarctica, total ozone is projected to reach the 1980 benchmark about 10 years later (Langematz and Tully 2018; also see Fig. 27-22). With declining ODS concentrations, the rising concentrations of greenhouse gases will have an increasingly important effect on the future evolution of the ozone layer.

![Fig. 27-20. Observed and modeled ozone trend profiles in northern midlatitudes for the periods (left) 1979–97 and (right) 2000–13 from observations (black lines) and the mean of CCMVal-2 model simulations (gray line with uncertainty range shaded). Red lines indicate the trend attributed to ODSs alone. [From WMO (2014).]](image)

![Fig. 27-21. Satellite observations showing the depletion of global total ozone beginning in the 1980s. Annual averages of global ozone are compared with the climatological averages from the period 1964 to 1980 before the ozone hole appeared. Seasonal and solar effects have been removed from the observational dataset. On average, global ozone decreased each year between 1980 and 1990. The depletion worsened for a few years after 1991 due to the effect of volcanic aerosol from the Mt. Pinatubo eruption. [From WMO (2015).]](image)
12. Solar variability and climate

When the American Meteorological Society (AMS) was founded in 1919, a substantial literature already existed on the topic of solar activity affecting Earth’s climate. At that time, it was well understood that the sun is the source of energy for Earth’s climate system, and observations showed that the sun’s output is variable. Over the past several decades, satellite and ground-based observations, together with advances in theory and modeling, have greatly advanced our knowledge of the sun and the importance of the stratosphere.
Observations have indicated that electromagnetic radiation from the sun varies with the 11-yr solar cycle (SC) so that the sun emits more radiation at sunspot maximum when, paradoxically, it is most covered with dark sunspots. A review paper by Gray et al. (2010) provides an overview of solar variability, observational evidence for solar variability affecting the climate, mechanisms for solar impacts on the climate, and advances in climate modeling of solar influences.

Herschel (1801) documented changes in features at the sun’s surface, and conjectured how the sun’s variation might affect climate and the price of wheat. Schwabe (1844) published a paper suggesting that the number of sunspots he observed varied periodically on a decadal time scale. It seemed only logical to inquire how much the sun’s energy output varied and to ask how this might affect our weather and climate. Langley (1884) attempted such measurements, and Abbot (1910) continued these efforts. Abbot (1910) claimed to observe a relationship between the sun’s total irradiance and climate, but later studies showed that the evidence for this was inconclusive. The beginning of the AMS coincided with greatly increased amounts of meteorological data, and the availability of these data along with the long record of sunspot variations led to a number of papers suggesting that variations in the sun’s activity were related to variations in a number of meteorological parameters. In Pittock’s (1978) critical review, he notes that already in 1920, Helland-Hansen and Nansen (1920) reviewed the literature on sun–weather relationships for the period 1826–1914, and cited 149 references. In this early period, the papers were of a statistical nature, correlating sunspot activity with variables such as lake levels, surface pressure, surface temperature, and precipitation, but later as upper-air data became more available, correlations were noted with a wider variety of variables, including temperature, tropopause pressure, and ozone. Pittock’s (1978) review paper contained about 170 references, most of which were statistical studies published during the 1970s, with a lesser number from earlier decades. This proliferation of papers during the 1970s probably motivated his critical review.

Several crucial advances in observations of the sun and new ways of analyzing the atmospheric data occurred during the period from mid-1970s to early 1980s. The most significant was the start of direct measurements by satellites in 1978 of the total amount of irradiance across the frequency spectrum, referred to as the total solar irradiance (TSI). These satellite observations were able to eliminate the interference from atmospheric effects that had plagued Langley and Abbot. Figure 27-23 shows a reconstruction of the TSI over four decades (from http://spot.colorado.edu/~kopp/TSI/). Note that it is maximum when the sunspot numbers maximize, and the amplitude of the smoothed TSI variations is only about 0.1%. It should also be noted that the construction of Fig. 27-23 from several different satellite instruments requires careful
treatment to account for differences in the absolute calibration of the different instruments.

The search for sun–climate relationships was further advanced by Karin Labitzke in her 1987 paper (Labitzke 1987). Several years earlier, Holton and Tan (1980) had suggested that the phase of the QBO in equatorial zonal winds had a great influence on northern polar (NP) winter temperatures (see sections 5 and 7), so that polar temperatures were higher in winters when the QBO was in its easterly phase. Labitzke (1987) linked this with solar cycle variability and showed that there was a strong tendency for higher NP temperatures to occur during solar maximum in QBO west phase years, while there was a very weak tendency for lower NP temperatures to occur during solar maximum in QBO east phase years. Figure 27-24 shows an update of van Loon and Labitzke’s (1994) analysis (using geopotential height instead of NP temperature) that illustrates that for the extended period 1942–2016 the relationship between the Holton–Tan effect and the phase of the 11-yr SC is still evident (Fig. 27-24). Model simulations with internally generated QBO and prescribed 11-yr SC (e.g., Schmidt et al. 2010; Kren et al. 2014) have been able to reproduce aspects of the observed relationship, but only over limited periods of the simulations and with limited statistical significance. This shows the complexity of the topic and the need for further research.

Another significant advance was the advent of proposed mechanisms for SC influence that could be clearly tested in GCMs. It was known, even before the direct measurements of TSI, that the very small, expected 11-yr (and 22-yr) SC modulations of solar output could have only a limited direct impact on temperatures at the surface and there had to be some manner in which these small modulations could tap into (and thus be amplified by) the very large atmospheric energy cycle to produce significant effects. One possible amplification route, known as the “bottom-up mechanism” involves the direct impact of TSI variations on sea surface temperatures that would then influence the evaporation of water vapor and produce regional-scale feedbacks either via cloud formation that can influence the nature of ENSO (Meehl et al. 2008) or by directly influencing the large-scale east–west dynamical circulation of the tropical troposphere, that is, the Walker circulation (Misios et al. 2019).
A second amplification mechanism is via the upper atmosphere. It was well known that this region showed very large variations during the SC, and that variability in the very short UV wavelengths was responsible for those changes. This led Hines (1974) to suggest that solar activity changes in the upper atmosphere could modulate atmospheric planetary wave propagation and structures, which could then feed back to lower levels, producing significant changes in weather and climate resulting from solar activity changes.

It is well known that there is much greater variability at the shorter solar spectral wavelengths than in the TSI [e.g., see the bottom panel of Fig. 3 in Gray et al. (2010)]. Wavelengths between 100 and 240 nm dissociate oxygen molecules, thereby leading to ozone formation, and wavelengths between 240 and 350 nm are effective in dissociating ozone and heating the stratosphere. Thus, there is a clear path for solar variability to influence not only heating of the stratosphere directly via changes in UV but also via changes in ozone that also lead to changes in ozone heating (Haigh 1994). It is also known that solar activity modulates energetic particle precipitation (EPP) that can produce reactive nitrogen in the thermosphere, and this reactive nitrogen can then be transported downward to the stratosphere to influence ozone amounts.

Changes in ozone and temperature of the stratosphere will also change the background winds and circulation (see section 3), providing the possibility for dynamical feedback in which the background wind structure affects wave propagation, thus amplifying the SC signal and enabling it to extend from the stratosphere downward into the troposphere (e.g., Kodera and Kuroda 2002). Both UV and EPP effects are generally referred to as the “top-down mechanism” since they involve modulation of the temperature, ozone, and circulation in the upper stratosphere that then penetrates to lower levels, including the troposphere and surface (e.g., Gray et al. 2013, 2016).

Dickinson (1975) suggested another possibility that involved solar-variability-induced variations in galactic cosmic rays (GCRs) affecting aerosol ionization. GCR fluxes are known to vary inversely with solar activity and can penetrate into the troposphere. Dickinson (1975) speculated that the ionization of aerosols might affect their effectiveness to act as condensation nuclei that spawn high-level clouds. Any such modulation in cloudiness would directly affect the atmospheric energy cycle and could influence weather and climate. This suggestion has been followed up by more recent studies showing solar influences on cloudiness, but the results of these have been questioned [see section 3.2.4 of Gray et al. (2010) for more details].

A number of GCM and CCM simulations have been carried out to study SC impacts on climate, including direct TSI impacts (e.g., Meehl et al. 2008; Misios et al. 2019) and indirect impacts via UV changes (e.g., Matthes et al. 2006; Ineson et al. 2011) and EPP changes (e.g., Baumgaertner et al. 2011; Arsenovic et al. 2016). No comparable climate models have yet examined the influence of solar modulation of GCRs on climate. Much more research is required before a clear understanding of the relative impacts of these mechanisms on weather and climate can be assessed.

In summary, solar influences on weather and climate had been active areas of research for many years before the AMS was founded. Enormous progress has been achieved in our knowledge and understanding since then, especially over the past few decades. The topic has emerged from its beginnings of almost purely investigations of statistical relationships that were subject to substantial criticism to become a solid scientific field that involves both solar physicists and climate scientists.

13. Volcanic eruptions, the stratosphere, and climate

In 1919 when the AMS was founded, we already knew that the 1883 Krakatau eruption had produced major environmental impacts (Symons 1888) and that volcanic eruptions were an important natural cause of climate change (Humphreys 1913). But because there were no major eruptions for more than 40 years, until the 1963 Agung eruption in Bali, little attention was paid to the topic in this period, except by Humphreys (1940) and Mitchell (1961). Mitchell (1961) was the first to conduct a superposed epoch analysis of the cooling impacts of volcanic eruptions at the surface, averaging the effects of several eruptions to isolate the volcanic effect from other presumably random fluctuations. He showed clear volcanic signals using 5-yr average periods, but did not have a very long temperature record.

In the years since the Agung eruption we have learned a lot. We now know that in addition to hemispheric or global cooling at the surface, the differential impacts of volcanic eruptions at the surface and in the stratosphere produce temperature and pressure gradients and dynamical responses, in the atmosphere and ocean, that affect surface air temperature patterns, El Niño, precipitation, and monsoons. Ozone is affected because of changed transport as well as chemistry; sulfate aerosols from volcanic eruptions serve as surfaces in the stratosphere for heterogeneous chemistry and ozone depletion. New modeling and observing capabilities, including satellites, balloons, and ice cores, have allowed much deeper understanding of the impacts of volcanic eruptions on climate and allowed us to tackle the issues involved in separating natural and anthropogenic impacts on climate change. Volcanic eruptions have been shown
to be an important cause of a “human genetic bottleneck” (a sharp reduction in the size of the human population) after the massive Toba eruption 74,000 years ago (Ambrose 1998; Robock et al. 2009), and of the Little Ice Age starting in about 1250 CE. Reviews by Robock (2000, 2013) and Timmreck (2012) have summarized our current understanding, and this chapter only has room to touch on some of the most important advances.

It is well known that volcanic eruptions with large sulfur dioxide injections into the stratosphere produce stratospheric sulfate aerosol clouds with an e-folding lifetime of about 1 year for tropical eruptions and several months for high-latitude eruptions, and that these clouds scatter some of the incoming sunlight back to space, cooling Earth. The aerosol layer also warms the stratosphere through absorption of infrared radiation, as shown by a climate model simulation of the impacts of the 1963 Agung eruption by Hansen et al. (1978). This results in a change to Earth’s energy balance. In addition, there are atmospheric and oceanic dynamical responses to large eruptions, producing characteristic regional and seasonal patterns of climate response. For example, following a large tropical volcanic eruption, the resulting latitudinal gradient of stratospheric heating, ozone depletion, and surface temperature patterns are observed to produce a stronger polar vortex in the NH, with a positive mode of the Arctic Oscillation in the winter, and winter warming of NH continents (e.g., Robock 2000). In fact, evidence of a warm winter in Europe from 1257 to 1258 CE was used to help determine the timing of the largest eruption of the past millennium, the 1257 Samalas eruption in Indonesia (Lavigne et al. 2013). The exact mechanism by which this “winter warming” is produced by volcanic eruptions, and whether volcanic eruptions are even involved, is still a matter of ongoing research (e.g., Polvani et al. 2019, and references therein), although climate models routinely produce this response to large tropical volcanic eruptions (e.g., Zambri and Robock 2016; Bittner et al. 2016a, 2016b).

Insolation reductions cool the land more than the oceans. Summer monsoons, which are driven by the land–ocean temperature gradient, are observed to be weaker following volcanic eruptions. The cooling reduces evapotranspiration and slows the hydrological cycle to some degree (Tilmes et al. 2013). The reduction in precipitation following the large 1783–84 Laki eruption in Iceland, likely due to the weaker summer monsoon circulation and lower water content of the advected air, produced famine in Africa, India, China, and Japan (Oman et al. 2006). A similar pattern was observed following the 1991 Mt. Pinatubo eruption, with widespread drought and reduced streamflow, but without quite such devastating impacts (Trenberth and Dai 2007).

High-latitude eruptions are different from low-latitude eruptions in several ways (Oman et al. 2005). For injections into the lower stratosphere, the aerosols have a shorter atmospheric residence time, on the order of 2–4 months, since the nature of the shallow part of the BDC ensures they remain in the high latitudes in a region of subsidence. However, cooling of Earth in only one hemisphere may shift the intertropical convergence zone toward the other hemisphere, potentially causing global-scale precipitation changes (Frierson and Hwang 2012; Haywood et al. 2013). Their impact on climate also depends on the time of year, with little impact in the fall and winter when there is little insolation (Kravitz and Robock 2011). Several small, high-latitude eruptions in the past decade had a smaller impact than small tropical eruptions during that time (Kravitz et al. 2010, 2011; Solomon et al. 2011; Bourassa et al. 2012). However, a very large high-latitude eruption, such as the 1783–84 Laki eruption, which included 10 explosive episodes with stratospheric injections at least as large as the 1812 El Chichón eruption (Thordarson and Self 2003), could have global impacts (Zambri et al. 2019a,b).

Climate model simulations have shown that a series of very large eruptions at the end of the thirteenth century, starting with the 1257 Samalas eruption, reduced North Atlantic oceanic heat flux into the Arctic so much that a feedback perpetuated this cool climate for centuries, starting the Little Ice Age (Zhong et al. 2011; Miller et al. 2012; Zambri et al. 2017; Slawinska and Robock 2018). In model simulations, large eruptions produce decadal-scale shifts in the North Atlantic circulation, with impacts during the next decade (Otterà et al. 2010; Booth et al. 2012; Zanchettin et al. 2012, 2013; Slawinska and Robock 2018).

Volcanic eruptions are known to significantly change stratospheric ozone via changes in dynamics and chemistry (e.g., Tie and Brasseur 1995; Tilmes et al. 2008a,b; WMO 2011), with important radiative impacts. In model simulations, the climate response to volcanic eruptions depends on accurate treatment of stratospheric ozone (Muthers et al. 2014; Mills et al. 2017). Simulations suggest that dynamical changes induced by large tropical volcanic eruptions are characterized by increased stratospheric upwelling in the tropics and enhanced extratropical downwelling. This can result in an increase of ozone in higher latitudes (Aquila et al. 2013). In addition, eruptions can strengthen the polar vortices, isolating and cooling the air within them, and inducing more chemical ozone depletion at high latitudes (Tilmes et al. 2009). Different factors may also change the frequency of SSWs (see section 5) and the associated rate of ozone depletion.
Volcanic eruptions also affect stratospheric chemistry by altering chemical reaction rates. Volcanic aerosols radiatively warm the stratosphere and in low and mid-latitudes, which can accelerate the rates of several important ozone-destroying cycles, including the Chapman cycle (see section 11). The enhanced aerosol surface area in the stratosphere increases the rate of heterogeneous and photolytic reactions.

The potential impacts of supervolcano eruptions, like the Toba volcano on the island of Sumatra, Indonesia, 74,000 years ago, may start ice ages and produce extinctions as well as a human genetic bottleneck (e.g., Ambrose 1998). Current work suggests that glacial advances were not observed after the Toba eruption (Robock et al. 2009; Haslam and Petraglia 2010; Svensson et al. 2013). Whether the decadal climate change was large enough to have large biological impacts is not settled—climate modeling gives different amplitudes depending on modeling assumptions (Robock et al. 2009; Timmreck et al. 2010) and paleoclimate observations are not detailed enough (e.g., Lane et al. 2013) to resolve annual signals.

Volcanic eruptions provide analogs for potential anthropogenic injection of stratospheric aerosols, either inadvertently, as a by-product of nuclear war (Toon et al. 2008, 2017), or advertently, in suggestions of using stratospheric geoengineering to reduce global warming (e.g., Crutzen 2006; Robock et al. 2008, 2013). In both cases, the transport of stratospheric aerosols and their impacts on climate and ozone as observed following volcanic eruptions can help evaluate climate model simulations of the effects of smoke from burning cities and industrial areas that might be targeted in a nuclear war and of sulfate aerosols from geoengineering. Clearly nuclear war must be avoided because of the horrendous direct effects of nuclear weapons, but also because of the potential for nuclear winter if the current arsenals were used in a war between Russia and the United States (Robock et al. 2007a; Toon et al. 2008), or catastrophic climate change from even a nuclear war between new nuclear states such as India and Pakistan (Robock et al. 2007b). While volcanic eruptions teach us that a geoengineered stratospheric sulfate cloud, if technically possible, would indeed cool Earth and reduce many impacts of global warming, there are likely to be many unintended consequences; it could also produce ozone depletion, with enhanced surface ultraviolet radiation; reduce summer monsoon precipitation (e.g., Trenberth and Dai 2007); and affect remote sensing and astronomical observations (Robock et al. 2013).

There are still a number of remaining research questions with respect to the impacts of volcanic eruptions on climate (e.g., Robock 2002). These include how volcanic SO$_2$ and ash emissions into the stratosphere interact to produce a stratospheric aerosol cloud (Is there quick removal of sulfur on the ash? What is the resulting aerosol size distribution?), how well the ice core record represents the past volcanic forcing of climate, and how the QBO interacts with volcanic aerosol clouds to produce climate responses. NASA (2018) now has plans to better observe the next large volcanic eruption.

In addition, we need to know how global warming will change future responses to volcanic eruptions. Aubry et al. (2016) showed that because of a rising tropopause in the future as a result of global warming, the same strength volcanic eruption plumes as now would produce fewer injections that made it into the stratosphere, and those that made it to the stratosphere would be closer to the tropopause and have shorter lifetimes. Hopcroft et al. (2018) pointed out that in a warmer future climate there would be less snow and ice and their positive albedo feedbacks would be weaker, and that with a more polluted troposphere in the future, the radiative forcing from stratospheric volcanic clouds would be smaller. All of these effects would lessen the impact of volcanic eruptions. On the other hand, Fasullo et al. (2017) showed that with a more stratified ocean in the future, the oceanic response to volcanic eruptions would be stronger.

While we have learned a lot about the effects of volcanic eruptions on climate in the 100 years since the founding of the AMS, the next 100 years promise many new findings. Volcanic eruptions cannot be predicted, but we can look forward to much better predictions of its impacts following the next big eruption, which is sure to occur.


Through the first half of the twentieth century the stratosphere was generally regarded as quiescent, free of weather, and not a significant influence on surface weather and climate. Scherhag (1952), who discovered SSWs (see section 5), was the first to suggest—based on very few observations—that sudden warming effects may descend to the surface and affect weather. More than 30 years later, Boville (1984) showed GCM simulations in which imposed variations in the strength of the stratospheric vortex resulted in circulation changes in the troposphere. Also, Quiroz (1986) examined observations of blocking before and after SSWs, but his results were inconclusive. Hines (1974; also Geller and Alpert 1980) suggested downward reflection of planetary waves as a sun-weather mechanism, but a clear statistical signal was only found a few decades later (Perlwitz and Graf 2001; Perlwitz and Harnik, 2003, 2004) and a clear effect on surface fields in observations.
was found by Shaw and Perlwitz (2013) and Lubis et al (2018).

During the rapid advances in studies of the ozone hole in the 1980s, it became clear that ozone loss had large effects on the stratospheric circulation, but no publications suggested the idea that SH surface weather and climate would be affected. That the surface circulation was affected was first shown by Thompson and Solomon (2002). Similarly, Kodera et al. (1990) and Kodera (1995) showed descending atmospheric signals associated with stratospheric variability but did not consider surface climate.

The idea that PV anomalies (see section 3) in the stratosphere, such as an anomalously weak vortex during an SSW, could have some effect on the surface circulation is implicit in the work of Hoskins et al. (1985). However, the magnitude of such effects was not clear, and there was no compelling observational basis. Theoretical advances, involving studies of perturbed axisymmetric vortices, were hampered by the practice of imposing a simplified lower boundary condition that did not allow surface pressure change. It was not until Haynes and Shepherd (1989) that surface pressure was allowed to vary in theoretical studies of axisymmetric vortices.

The principle of “downward control” was proposed by Haynes et al. (1991), in which sustained anomalies in wave driving (Eliassen–Palm flux divergence) in the stratosphere affect the zonal-mean meridional circulation below. In general, stratospheric Eliassen–Palm flux convergence decelerates the zonal-mean westerly flow and should therefore have an effect on the tropospheric circulation below.

“PV inversion” was used by Hartley at al. (1998), and later Black (2002), to calculate the near-surface effects of stratospheric PV anomalies. Their approach was successful, but they did not find large changes to the tropospheric circulation from stratospheric PV anomalies. The Arctic Oscillation (now the NAM) was defined by Thompson and Wallace (1998), who showed that it was strongly connected to the stratosphere. Baldwin and Dunkerton (1999) pursued this idea, extending the NAM definition from the surface through the stratosphere. Their discovery that anomalies in NAM appeared to propagate downward to the surface was unexpected, since prior research did not indicate a significant lower-tropospheric response to stratospheric changes.

Subsequent research (e.g., Baldwin and Dunkerton 2001) refined the observational picture (see Fig. 27-25 for a zonal-mean view of effects on temperature), which showed the following main tropospheric characteristics...
of stratosphere–troposphere coupling (Baldwin et al. 2019, manuscript submitted to Nat. Geosci.):

- The sea level pressure (SLP) response to stratospheric variability is similar to the NAM pattern.
- The surface climate impacts of stratospheric variability last for around 2 months on average.
- To leading order, surface effects are proportional to NAM anomalies in the lower stratosphere—the relationship is approximately linear.
- The stratospheric signal at the surface is slightly delayed, giving the appearance of downward propagation.
- Atlantic and Pacific jets and storm tracks shift systematically in response to stratospheric variability (Fig. 27-26). This is consistent with the NAM signal.

Observations show that, on average, a NAM response occurs within a few days of weak/strong vortex events (Baldwin and Dunkerton 2001). Modeling studies (Fig. 27-27) have shown that the SLP response—similar to the NAM—is qualitatively similar across all time scales from weekly to centennial (Hardiman et al. 2012; Ineson and Scaife 2009; Ineson et al. 2011; Scaife et al. 2012). This result holds for both hemispheres and suggests that the same mechanism dominates at all time scales. Perlwitz and Harnik (2004), however, noted that during winters for which there is strong downward reflection of planetary waves, the zonal-mean downward signal does not extend to the troposphere; instead, there is a strong downward zonal wave coupling (Perlwitz and Harnik 2004).

Unambiguous observations of surface effects of stratospheric variability, as well as numerical simulations, suggest an important role of tropospheric eddy feedbacks [Song and Robinson (2004) and Kunz and Greatbatch (2013) for NAM coupling; Lubis et al. (2018) for wave coupling]. However, a simple theoretical explanation of the tropospheric effects listed above has proven to be challenging.

Observational and modeling studies have examined the occurrence of extreme surface weather. Following an SSW, the likelihood of Atlantic blocking increases (Scaife and Knight 2008; Woollings et al. 2010). Stratospheric variability is associated with extreme weather events in Europe (Kolstad et al. 2010). For example, Tomassini et al. (2012) found that 40% of extreme winter cold spells over Europe may be preceded by a weakening of the stratospheric polar vortex.

A striking example of stratospheric change affecting surface climate is SH ozone loss leading to marked changes in surface climate, through the radiative and dynamical effects of the Antarctic ozone hole [see Thompson et al. (2011) for a review and Kidston et al. (2015) for a summary]. The effects on surface climate are most pronounced during the austral summer season and strongly resemble the most prominent pattern of large-scale SH climate variability, the southern annular mode (SAM; Thompson and Solomon 2002). The anomalous tropospheric SAM is consistent with low summertime temperatures over east Antarctica and higher temperatures in Patagonia and the northern Antarctic Peninsula. Ozone loss is associated with increased summertime precipitation on the eastern side of the Great Dividing Range in southeastern Australia and the Southern Alps in New Zealand. Summertime temperatures are higher than normal throughout much of New Zealand and lower than normal over central and eastern subtropical.
Australia (Thompson et al. 2011). Over the next few decades, recovery of the ozone hole and increases in greenhouse gases are expected to have significant but opposing effects on the SAM and its attendant climate impacts during austral summer.

The stratosphere also appears to influence the ocean in both hemispheres (Lenton et al. 2009; Reichler et al. 2012; Scaife et al. 2013; Cagnazzo et al. 2013; O’Callaghan et al. 2014). SH ozone depletion has caused a poleward shift of the tropospheric jet (Arblaster and Meehl 2006; Son et al. 2010). This jet shift affects wind stress over the Southern Ocean (Cagnazzo et al. 2013), which affects the pattern of air–sea fluxes, and this is also likely to change ocean carbon uptake (Lenton et al. 2009).

In addition, low-frequency variation of the Atlantic thermohaline circulation is similar to variations to the stratosphere over the last 30 years, which model simulations suggest could be linked to persistent stratospheric circulation anomalies (Reichler et al. 2012). Coupling between the atmosphere and the ocean may also explain the lag of about 3 years in the North Atlantic climate response to solar variability (Scaife et al. 2013).

Tropospheric effects of stratospheric variability are not necessarily limited to mid- to high latitudes. The QBO phase may affect convection in some equatorial regions (e.g., Collimore et al. 2003) and the amount of MJO activity (Yoo and Son 2016). The phase of the QBO has been used to predict the number of Atlantic hurricanes (Gray 1984). However, using a longer data record, Camargo and Sobel (2010) conclude that the QBO does not exert a significant influence on tropical cyclones.

Although there are theoretical reasons to expect surface effects from stratospheric variability, quantifying the effects has been challenging. In general, the consensus has been that the observed surface effects are larger than can be justified without a tropospheric amplification mechanism such as transient eddy feedback—for example, a mechanism that would cause maximum SLP anomalies near the North Pole. Baldwin et al. (2019, manuscript submitted to Nat. Geosci.) quantified the amplification of stratospheric variability (Fig. 27-28) and proposed a mechanism to explain why, at least on short time scales, the SLP pattern looks like the NAM. They showed that the tropospheric heat flux into the Arctic is partly controlled by stratospheric variability, leading to anomalously cold/warm Arctic conditions, and suggested that through radiative-cooling-induced anticyclogenesis (e.g., as in the Siberian anticyclone) pressure anomalies similar to the NAM are formed. The net effect is that the stratospheric pressure signal is amplified.

15. Role of the stratosphere in weather and climate prediction

The importance of the stratosphere for weather and climate prediction has now matured to the point where, for some time scales at least, there is clear evidence of an important role for the stratosphere in contributing to the skill of predictions. In recent years, an increasing number
of leading numerical prediction systems have therefore improved their representation of the stratosphere.

Careful analysis of observational data (e.g., Kodera 1995; Baldwin and Dunkerton 1999) revealed that circulation anomalies appeared to descend through the stratosphere and change the troposphere. The observations of downward progression of wind anomalies do not necessarily imply that downward causality is actually present (Plumb and Semeniuk 2003), but climate model experiments in which only the stratosphere was perturbed demonstrated similar tropospheric effects (Kodera et al. 1990; Polvani and Kushner 2002; Scaife et al. 2005; Hardiman and Haynes 2008). Here we review the evidence for stratospheric impacts in predictions across time scales, from short-range weather forecasts out to seasonal and decadal climate predictions.

The evidence for stratospheric influence on weather forecasts of several days ahead is limited. Some direct evidence of downward influence in weather forecasts has been derived from experiments where the stratosphere was directly altered in forecasts (Jung and Barkmeijer 2006). Tropospheric effects were found to develop after a few days in these experiments, and there is some evidence that improved stratospheric representation in models could improve the skill of forecasts (Charron et al. 2012). One area where the stratosphere is important is in the assimilation of data to produce comprehensive initial conditions for forecasts. The fact that the satellite radiometer measurements are often significantly weighted in the stratosphere means that accurately resolving the stratosphere can improve atmospheric data assimilation. Furthermore, it is possible to see emerging impacts of stratospheric initial conditions on weather forecasts a few days ahead (Charlton et al. 2004, 2005). As weather forecast systems continue to become more accurate, the relative importance of accurate numerical representation of the stratosphere increases.

There is much more evidence of a stratospheric impact on forecasts out to a month ahead. Early experiments gave a statistically significant reduction in the quality of monthly forecasts when the stratosphere was degraded. This occurred through a clear change in surface conditions over high latitudes (Boville and Baumhefner 1990). Baldwin et al. (2003) showed that statistical forecasts of the troposphere were possible from prior knowledge of the stratospheric flow alone, and subsequent analysis demonstrated that this process could help generate forecasts that exceeded the skill of dynamical long-range forecasts at the time (Christiansen 2005). Armed with new knowledge of where to look for surface impacts, numerical studies with ensembles of monthly forecasts showed a clear role for the stratosphere, particularly during case studies of SSWs (Kuroda 2008), and in some cases the stratosphere has larger effects than the better-known impacts of ocean conditions (Kushnir et al. 2019; Scaife and Knight 2008). Most of the impact is seen in the North Atlantic Oscillation (NAO) or its hemispheric equivalent, the NAM, and associated blocking patterns, with evidence of two-way interactions between the troposphere and stratosphere (Martius et al. 2009; Kolstad et al. 2010). Other SSW events have since been analyzed, and the results indicate the same surface effects (Mukougawa et al. 2009; Marshall and Scaife 2010; Sigmond et al. 2013; Tripathi et al. 2015), adding to the evidence for an important role of the stratosphere on monthly surface weather forecasts.

There is now clear evidence for stratospheric impacts on seasonal forecasts out to a few months ahead. These are mostly for the NH extratropics (Gerber et al. 2012) where the occurrence of SSWs is important for forecast skill (Scaife et al. 2016). However, there is also evidence of stratospheric impact on seasonal forecasts for the SH (Seviour et al. 2014), and recent studies suggest potentially important effects on tropical predictability from the stratospheric QBO (Yoo and Son 2016; Marshall et al. 2017) via effects on the MJO. A cornerstone of seasonal prediction is the long-range predictability of ENSO, but studies have long shown that this has a significant effect on the stratosphere (e.g., Hamilton 1993) and this allows it to impact surface climate, not only in the Pacific sector but also in the Atlantic (Brönnimann et al. 2004; Ineson and Scaife 2009; Cagnazzo and Manzini 2009). Improved representation of the stratosphere is now leading to improved fidelity of this teleconnection in seasonal prediction systems (e.g., Butler et al. 2016). Even intraseasonal effects of ENSO from the stratosphere (Moron and Gourand 2003; Herceg-Bulić et al. 2017) are starting to be represented in climate models, albeit weakly (King et al. 2018; Ayarzagüena et al. 2018a). In long-range prediction systems that already include the stratosphere, the QBO stands out as an obvious source of atmospheric predictability out to very long time scales. Because it is fairly regular, the QBO itself can be predicted at seasonal to interannual range (Pohlmann et al. 2013; Scaife et al. 2014). It also affects the extratropical stratosphere (Holton and Tan 1980) and surface climate in the NH (Thompson et al. 2002; Gray et al. 2018) and therefore has the potential to improve seasonal predictions (Boer and Hamilton 2008; Marshall et al. 2009). However, this teleconnection is weaker in models than in the observational record (Scaife et al. 2014), and although it is possible that the observed teleconnection in the limited data record (one sample of ∼60 years) is stronger than would be observed in a much
longer data record, Andrews et al. (2019) show that this is unlikely to account for the difference.

Decadal forecasting using dynamical models is a relatively young field, premised on the effects of initial conditions, particularly in the ocean, that give effects lasting for years or even a decade into the predictions and adding to skill from external forcing (Smith et al. 2007; Meehl et al. 2014). It is only now that these predictions are moving from research into real-time operations (Smith et al. 2013; Kushnir et al. 2019). Although the memory of the stratosphere is generally shorter than these time scales, it still plays an important role, for example, in the forced atmospheric response to tropical volcanic eruptions (see section 13). Volcanoes have the potential to add multiyear predictability—but only in the years following volcanic eruptions (Swingedouw et al. 2017). A second source of decadal predictability involving the stratosphere is the quasi-regular variations in solar irradiance from the 11-yr SC. The top-down mechanism (see section 12) has now also been reproduced in general circulation models (Matthes et al. 2006; Ineson et al. 2011) where it generates predictable multiyear influences on the NAO and Arctic Oscillation (Gray et al. 2013, 2016; Dunstone et al. 2016), aided by persistent heat anomalies in the Atlantic Ocean (Kodera 2007; Scaife et al. 2013; Andrews et al. 2015). There is further evidence that this may even phase lock internal decadal variability of the NAO to the SC (Thiéblemont et al. 2015). Finally, Reichler et al. (2012) suggest an active role for the stratosphere on longer, interdecadal time scales through interactions with the Atlantic multidecadal oscillation, and the experiments of Omrani et al. (2016) provide further evidence that North Atlantic Ocean variability (NAV) impacts the coupled stratosphere–troposphere system. As NAV has been shown to be predictable on seasonal-to-decadal time scales, these results have important implications for the predictability of the extratropical atmospheric circulation on these time scales. It remains to be seen whether the stratosphere leads to more skillful predictions on these multidecadal time scales.

In summary, long-range forecasts are now produced operationally on all time scales from monthly to seasonal to decadal. Although operational predictions on the decadal time scale are only just being achieved (Kushnir et al. 2019), the stratosphere has been shown to be important and, in some cases, potentially crucial, to the skill of climate predictions on all of these time scales. The influence of the stratosphere on predicted surface climate is often due to downward propagating zonal wind anomalies and subsequent impacts on the NAO and the NAM. The QBO and solar variability impart low-frequency variability to the troposphere, while for ENSO, the stratosphere simply acts as a conduit for teleconnections. Skillful predictions on some of these time scales are recent achievements compared to the century of progress considered in this monograph, but we are in a period of rapid progress in this area and some of the effects of the stratosphere currently appear to be too weakly represented in climate models (Scaife and Smith 2018), suggesting further room for improvement. The stratosphere is now emerging as a key factor in long-range forecasting, and it will be exciting to see how this develops in the coming years.

16. Climate change and the stratosphere

Although it has been known since the nineteenth century that increasing the amount of CO₂ in the atmosphere will cause a warming at Earth's surface, Manabe and Wetherald (1967) were the first to predict a concomitant cooling of the stratosphere. Such a pattern of tropospheric warming and stratospheric cooling has been observed over the last 50 years (Fig. 27-29) and cannot be explained by natural changes in solar irradiance, volcanic eruptions, or natural climate phenomena such as El Niño and La Niña. Instead, this pattern is now recognized as a characteristic "fingerprint" of increasing amounts of CO₂ in the extratropics (e.g., Santer et al. 2013), with warming below ~200 hPa and cooling above that level.

The earliest evidence that increasing amounts of CO₂ could also change the amplitude of the tropospheric wave forcing of the stratosphere came from Rind et al. (1990). With doubled CO₂ concentration in their stratosphere-resolving model they found a strengthening of the BDC (see section 2). By the time of the first comprehensive survey of the role of the stratosphere in tropospheric climate change (Rind and Lacis 1993), it was already recognized that the troposphere responds to both changes to the circulation and temperature of the stratosphere, and to changes in stratospheric composition through perturbations to the radiative forcing.

Surface temperature is also radiatively affected by changes in stratospheric ozone, water vapor, and aerosol. Since the mid-1970s it has been known that the presence of stratospheric aerosol is likely to cool Earth’s surface (Harshvardhan and Cess 1976). Changes in the temperature structure of the tropical tropopause layer, methane oxidation, or emissions from high-flying aircraft may also change stratospheric water vapor concentrations. Ozone heats the stratosphere through the absorption of solar UV and visible radiation (see section 11). Manabe and Wetherald's (1967) original radiative study concluded that if ozone concentrations are reduced, then solar heating will be reduced and the
stratosphere will cool. Most of the observed cooling of the stratosphere toward the end of the twentieth century (Fig. 27-29) was due to ozone depletion (Langematz et al. 2003; Arblaster et al. 2014), with increasing CO₂ cooling being a secondary effect.

Some recent simulations with CCMs show significant impacts of interactive chemistry of stratospheric ozone on surface climates in a global warming simulation (Nowack et al. 2015) or paleoclimate simulations (Noda et al. 2017, 2018) compared with each corresponding simulation with prescribed chemistry, although the stratospheric ozone chemistry feedbacks are not critical for the climate sensitivity in another global warming model (Marsh et al. 2016).

Globally, stratospheric ozone began to decrease around 1970, reaching a minimum around 2000 (ODS concentrations began to decrease in the mid-1990s). Since approximately 2000, global ozone has slowly begun to increase (see section 11). Although the global average is expected to recover in the future, the latitudinal and seasonal distributions of ozone will not be the same as in 1970—because the stratosphere will have cooled significantly and the distribution of ODSs will not be the same. Future trends in stratospheric temperatures (Arblaster et al. 2014; WMO 2018) will mainly result from the opposing effects of increasing CO₂ (colder stratosphere) and increasing ozone (warmer stratosphere).

Surface climate impacts of ozone depletion in the SH are expected to reduce over the coming decades as stratospheric ozone levels recover. However, greenhouse gas (GHG) concentrations will continue to grow and will thus be a key driver of future SH climate change. The relative importance of ozone recovery for future SH climate will depend on the evolution of atmospheric GHG concentrations (WMO 2018).

Consistent with the smaller ozone loss observed in the Arctic, NH tropospheric and surface circulation changes cannot be robustly linked to Arctic stratospheric ozone depletion (Arblaster et al. 2014). Nonetheless, a demonstrable downward coupling between stratosphere
and troposphere (see section 14) exists on all time scales [Fig. 27 from Kidston et al. (2015)] and has been investigated since the beginning of the century (e.g., Black 2002). Model predictions of future trends in the strength of the NH polar vortex are uncertain. This is because the climate forcing in mid- to high latitudes is the small difference between GHG cooling and adiabatic warming from a projected faster BDC with additional uncertainties arising from ozone changes and from changes in tropospheric wave driving. For example, a robust trend in the simulated frequency of occurrences of SSWs (see section 5) is difficult to establish statistically because the events are infrequent (Nishizawa and Yoden 2005). Analyzing climate projections of 12 models participating in the Chemistry–Climate Model Initiative (CCMI), Ayarzagüena et al. (2018b) did not find a statistically significant change in the frequency of SSWs over the twenty-first century, irrespective of the metric used for the identification of the event.

Undoubtedly the uncertainty in the high-latitude stratospheric response contributes to the uncertainty in the corresponding tropospheric response (e.g., Simpson et al. 2018). However, perhaps more importantly, Sigmond et al. (2008) discovered that the response of the tropospheric storm tracks to a doubling of the CO2 concentration also depended on details of the representation of the stratosphere in their model. Scaife et al. (2012) found that, for an ensemble of models, stratosphere–troposphere interactions had a significant influence on twenty-first-century climate change projections of the Atlantic storm track and hence extreme rainfall. A significant consequence of these findings, supported by more recent studies (e.g., Karpechko and Manzini 2012; Manzini et al. 2014; Kidston et al. 2015), is a move toward a new generation of state-of-the-art climate and Earth system models with fully resolved stratospheres. This will then allow the full extent of stratospheric influence on the response of surface climate to anthropogenic forcing to be investigated.

17. Concluding remarks

For the past 100 years, we have come to understand many of the dynamical and chemical processes and phenomena in the stratosphere. We have learned that the stratosphere is very sensitive to small changes in concentrations of radiatively active gases—including ozone depleting substances and greenhouse gases. Arguably, the most important accomplishment of stratospheric science was solving the ozone crisis. Ozone loss has potential severe consequences for humanity and the environment, and we are indeed fortunate that 1) the observers and observing systems were in place to identify the problem, 2) we were able to find solutions, and 3) there was the political will to build and adhere to the Montreal Protocol and subsequent Adjustments. GCM simulations have been carried out to estimate what would have happened if we remained ignorant or if we chose to ignore the problem (Morgenstern et al. 2008; Newman et al. 2009; Garcia et al. 2012). These studies are commonly called “the world avoided.” It turns out that the Montreal Protocol provided a dual protection to ozone and climate. It not only helped to prevent damage to Earth’s ozone layer, it has also slowed global warming. Observations of ODSs and ozone have been largely consistent with model simulations supporting the Montreal Protocol and subsequent amendments. In simulations, the severe ozone loss that would have occurred without the Montreal Protocol couples downward to the surface, especially in the Antarctic and Arctic, resulting in large changes to temperatures, pressure, and winds. The Montreal Protocol has provided an enormous benefit not only to the stability of the stratospheric ozone layer but also to surface climate (Morgenstern et al. 2008; Garcia et al. 2012).

Future changes to the stratosphere—and to the stratospheric impact on surface weather and climate—depend on two main factors: maintaining worldwide agreements to greatly reduce ODSs, and the degree to which humanity continues to extract fossil carbon. Although the story of fixing the ozone layer appears to be over, environmental regulations cannot be taken for granted and must be safeguarded. Recently (Rigby et al. 2019; Montzka et al. 2018), it has been discovered that some countries are being less than truthful by producing, but not disclosing, ozone-depleting chemicals such as CFC-11. Emissions of the fully controlled CFC-11 are increasing, and it is likely that this increase comes from illegal production.

Although the term “greenhouse gas emissions” is frequently used, the essence of the problem is that the main driver of increasing CO2 is the extraction of fossil carbon for energy and cement production. Future stratospheric temperatures—say, in the year 2100—are not known, because we do not know what path humanity will choose. The stratospheric ozone layer is expected to not only recover back to its pre-1980 condition, but overshoot its previous levels, mainly because of increasing CO2. The precise recovery timeline is uncertain, and depends on which emissions scenario is assumed.

Acknowledgments. We thank Shigeo Yoden, Peter Haynes, and an anonymous reviewer for insightful comments and suggestions. We thank Paul Newman for comments and advice. This research has been funded by the National Science Foundation Grant AGS-1430051,
the Natural Environment Research Council Grant NE/M006123/1, the European Research Council under the European Union’s Seventh Framework Programme (FP7/2007-2013), ERC Grant Agreement 603557, and the European Commission Marie Curie Career Integration Grant FP7-PEOPLE-2013-CIG. AS and NB were supported by the Joint DECC/Defra Met Office Hadley Centre Climate Programme (GA01101). KS was supported by JST CREST Grant Number JPMJCR1663, Japan.

REFERENCES


Herceg-Bulic, I., B. Mezzina, F. Kucharski, P. Ruggieri, and M. P. King, 2017: Wintertime ENSO influence on late spring


Stiller, G. P., and Coauthors, 2012: Observed temporal evolution of 27.60 METEOROLOGICAL MONOGRAPHS V OLUME 59

Strahan, S. E., and Coauthors, 2011: Using transport diagnostics to under-


Talaat, E. R., and R. S. Lieberman, 2010: Direct observations of non-


Strahan, S. E., and Coauthors, 2011: Using transport diagnostics to un-


Strahan, S. E., and Coauthors, 2011: Using transport diagnostics to un-


