The physics of Martian weather and climate: a review

How to cite:


© 2015 IOP Publishing Ltd

Version: Version of Record

Link(s) to article on publisher’s website:
http://dx.doi.org/doi:10.1088/0034-4885/78/12/125901

oro.open.ac.uk
The physics of Martian weather and climate: a review

This content has been downloaded from IOPscience. Please scroll down to see the full text.
2015 Rep. Prog. Phys. 78 125901
(http://iopscience.iop.org/0034-4885/78/12/125901)

View the table of contents for this issue, or go to the journal homepage for more

Download details:

IP Address: 137.108.145.21
This content was downloaded on 06/11/2015 at 09:37

Please note that terms and conditions apply.
Review

The physics of Martian weather and climate: a review

P L Read\textsuperscript{1}, S R Lewis\textsuperscript{2} and D P Mulholland\textsuperscript{1,3}

\textsuperscript{1} Atmospheric, Oceanic \& Planetary Physics, University of Oxford, Clarendon Laboratory, Parks Road, Oxford, OX1 3PU, UK
\textsuperscript{2} Department of Physical Sciences, The Open University, Walton Hall, Milton Keynes, MK7 6AA, UK
\textsuperscript{3} Department of Meteorology, Reading University, Earley Gate, PO Box 243, Reading, RG6 6BB, UK

E-mail: p.read1@physics.ox.ac.uk, stephen.lewis@open.ac.uk and d.p.mulholland@reading.ac.uk

Received 17 April 2015, revised 26 August 2015
Accepted for publication 27 August 2015
Published 5 November 2015

Abstract

The planet Mars hosts an atmosphere that is perhaps the closest in terms of its meteorology and climate to that of the Earth. But Mars differs from Earth in its greater distance from the Sun, its smaller size, its lack of liquid oceans and its thinner atmosphere, composed mainly of CO\textsubscript{2}. These factors give Mars a rather different climate to that of the Earth. In this article we review various aspects of the martian climate system from a physicist’s viewpoint, focusing on the processes that control the martian environment and comparing these with corresponding processes on Earth. These include the radiative and thermodynamical processes that determine the surface temperature and vertical structure of the atmosphere, the fluid dynamics of its atmospheric motions, and the key cycles of mineral dust and volatile transport. In many ways, the climate of Mars is as complicated and diverse as that of the Earth, with complex nonlinear feedbacks that affect its response to variations in external forcing. Recent work has shown that the martian climate is anything but static, but is almost certainly in a continual state of transient response to slowly varying insolation associated with cyclic variations in its orbit and rotation. We conclude with a discussion of the physical processes underlying these long-term climate variations on Mars, and an overview of some of the most intriguing outstanding problems that should be a focus for future observational and theoretical studies.

Keywords: atmospheric physics, Mars, weather, climate

(Some figures may appear in colour only in the online journal)
under, the martian surface has actually been known for many years, albeit almost entirely in the form of ice or water vapour, although the precise extent and amount are still being quantified as more sophisticated instrumentation is deployed to measure it. In addition, there is abundant evidence, both from geomorphological features on the martian surface and recent geochemical discoveries by \textit{in situ} instruments at the surface, that liquid water (or a water-like substance) has flowed and persisted on the surface in vast quantities in the distant geological past. Speculation continues unabated as to whether liquid water might be sustainable, even if only intermittently in localized regions, under present conditions. Such possibilities raise the prospect of Mars (past or present) as a potential habitat for living organisms, and the possible habitability (and even the actual habitation) of the martian environment continues to be a prime motivation for its continued exploration.

In developing an understanding of such questions relating to conditions in the martian environment, past, present or future, a knowledge of how its climate system works is crucial. As for the Earth, such an endeavor may entail the development of ever more complex and sophisticated numerical models of the martian environment, using techniques that directly parallel those used for simulating and predicting the Earth’s weather and climate. But to constitute a real understanding of the system, it is vitally important also to focus on a quantitative understanding of the physics and chemistry that underpin the formulation of such models and govern the behaviour of the atmosphere and near-surface environment, determining the structure of the atmosphere and the conditions of temperature, humidity, wind and composition close to the surface.

In many respects, the key processes that determine such conditions on Earth are also important on Mars. The temperature structure on both planets is largely determined by the balance of radiative heating by the Sun and cooling by thermal radiation to space from the surface and atmosphere, with temperature variations moderated by the transport of heat, momentum and other tracers in a global-scale atmospheric circulation system. The similar rotation rate and obliquity of both planets also means that the styles of global circulation encountered on both planets are somewhat similar, being dominated by direct overturning motions at low latitudes and cyclone–anticyclone weather systems at higher latitudes, with a similar pattern of seasonal variations between hemispheres during the year. But the differing sizes of the planets, atmospheric masses and compositions and distances from the Sun (among other things) lead to a number of key differences that need to be understood from first principles. Moreover, both planets have experienced quite different evolutionary paths, such that conditions early in their respective histories (and the consequent states of their climate) may have been very different from those encountered today.

In setting out to write this review, we have chosen to focus on recent developments in our understanding of the fundamental physical processes that determine the climate on Mars, rather than on describing the detailed phenomenology. The subject is a very broad one, and our coverage is of necessity somewhat selective. We have chosen to treat those processes that play the strongest role in affecting the present martian climate system, and to compare and contrast with similar processes on Earth. Since we are writing for physicists, we have emphasized the physical principles involved, if necessary at the expense of the technical details associated with the formulation and implementation of complex numerical models, or of some of the geographic and phenomenological detail that characterizes many current discussions of martian weather and climate.

The paper is organized as follows. Section 2 introduces the key energetic processes that determine the gross thermal structure of the martian atmosphere. These are combined in simple models which describe a straightforward equilibrium climate that neglects global transport processes associated with large-scale atmospheric motion. The latter is then discussed in some detail in section 3, in which we focus on classifying and examining the main coherent circulation patterns within the martian global circulation and discuss their origin and behaviour. Sections 4 and 5 respectively discuss the key material transport cycles of mineral dust and volatiles, including the formation of clouds and the possible interaction of condensibles with suspended dust. Finally, in section 6 we draw together the threads and discuss some outstanding unresolved issues that confront, and are challenging, the Mars science community, both now and for the future.

2. Atmospheric thermal structure

As with other externally irradiated atmospheres, the main source of buoyancy in the martian atmosphere is due to the differential illumination across the planet by the Sun, which is only partly compensated by direct thermal radiation from the surface and atmosphere to space. Together with the other thermal energetic processes in the martian atmosphere and surface, this leads to an observed atmospheric structure with a number of characteristic features. Some of these features are illustrated in figure 1, which shows a sequence of zonally averaged temperature fields as a function of latitude and height (log-pressure) for the cardinal seasons of Mars during 2007–9 (Mars Year 29\textsuperscript{4}). These are derived directly from recent infrared remote sounding observations from the Mars Climate Sounder (MCS) instrument on board NASA’s Mars Reconnaissance Orbiter spacecraft, currently in a low circular orbit around Mars (McCleese \textit{et al} 2010).

The main notable features are (a) the general tendency for temperature to decrease with height at most latitudes, especially in the tropics, (b) the trend for temperatures to decrease towards the poles in both hemispheres during equinoxes in the lower atmosphere, but for temperatures to reach a maximum near the summer pole and a minimum at the winter pole during solstice seasons, and (c) the occurrence of temperature inversions and local maxima in temperature at high latitudes in the middle and upper atmosphere. A particular point to note is that Mars does not possess a stratosphere in its middle atmosphere, where temperature generally increases in height,

\textsuperscript{4} By convention (Clancy \textit{et al} 2000), Mars Year 1 is taken to start on 11 April 1955 and subsequent years are numbered accordingly.
unlike on Earth between altitudes of around 10 and 50 km. This is mainly because, unlike on Earth, Mars does not have a photochemically induced ozone layer in its middle atmosphere, largely because there is little or no molecular oxygen from which to generate ozone. In the following subsections, therefore, we consider what are the key physical processes and factors that determine the large-scale thermal structure of the martian atmosphere. This will lay the basic foundations for understanding how Mars’s global atmospheric circulation is driven and why it adopts the form as observed and modelled.

2.1 Radiative forcing

As intimated above, many of the basic features of the thermal structure of the martian atmosphere are associated with how and where the atmosphere is heated and cooled. To a large extent, this is down to how the atmosphere interacts with radiation, both in the short-wave visible, near-infrared and ultraviolet wavelength bands received from the Sun, and the intrinsic thermal radiation emitted (mostly in the mid- and far-infrared) by both atmosphere and surface.

2.1.1 Sources of gaseous opacity and greenhouse warming. The current gaseous atmosphere of Mars is primarily composed of CO$_2$ (around 96% by volume), with smaller amounts of N$_2$ (1.89%), Ar (~1.93%) and trace amounts of other species such as O$_2$, H$_2$O and O$_3$ (Owen 1982, Mahaffy et al 2013). Such a mixture of gases is relatively transparent in the visible and soft UV, and only slightly absorbing in the near-infrared, due to weak absorption bands of CO$_2$ between 1 and 2 $\mu$m wavelength (e.g. see figure 2). As a result, only about 1% of the incoming solar radiation is directly absorbed by the gaseous atmosphere (Savijärvi et al 2005). CO$_2$, however, has a strong vibration–rotation absorption band centered around 15 $\mu$m (between 11 and 20 $\mu$m; see figure 2), together with other strong bands at around 2.7 $\mu$m and 4.3 $\mu$m, which makes the atmosphere act as a fairly strong absorber in the thermal infrared at the temperatures encountered in the martian environment, absorbing around 20% of the thermal radiation emitted from the surface. This disparity in opacity between the visible and mid-infrared leads to a significant greenhouse effect on Mars (amounting to a net warming of around 5–6 K; e.g. Pierrehumbert (2010)), though the overall greenhouse effect on Mars is not as strong as on the Earth (where net warming is ~30 K; e.g. Le Treut et al (2007) and Pierrehumbert (2010)). This is mainly because of the absence on Mars of significant amounts of other greenhouse gases found on Earth, such as H$_2$O and CH$_4$, as evident in figure 2, which shows only weak features attributable to H$_2$O except in the far infrared where thermal emission from the surface is quite weak.

In terms of column amounts of CO$_2$, Mars’s atmosphere actually contains much more CO$_2$ than the Earth’s. At the
current CO$_2$ concentration, $c$, in the Earth’s atmosphere of around 400 ppmv (e.g. Le Treut et al 2007), this corresponds to a vertical column mass $M_{CO_2} = cp/g \approx 4$ kg m$^{-2}$ (where $p_s$ is the surface atmospheric pressure), compared with around 150 kg m$^{-2}$ for Mars. A column of the Earth’s atmosphere, however, also typically contains around 40 kg m$^{-2}$ of water vapour (compared to around 10–100 g m$^{-2}$ on Mars (Smith 2002)) as well as smaller amounts of other greenhouse gases, all of which absorb heavily in regions of the spectrum that are not affected by CO$_2$. Together with the effects of pressure broadening, which are much greater on Earth than on Mars (e.g. Pierrehumbert 2010), this leads to a much greater overall opacity in the infrared than on Mars, with the atmosphere absorbing around 70–85% of the thermal radiation emitted from the surface. This is partly because at the low surface pressures on Mars, pure CO$_2$ absorbs relatively little except within the main 15 $\mu$m absorption band, which only cuts out a modest fraction of the complete blackbody spectrum emitted from the surface (see figure 2). But also, the main 15 $\mu$m absorption band of CO$_2$ at the low column densities and pressures encountered in the martian atmosphere becomes saturated and hence is less effective as an infrared absorber for a given column mass than if it were optically thin. Nevertheless, absorption and emission of infrared radiation in the main 15 $\mu$m band is one of the dominant means by which the atmosphere is heated and cooled, in the former case by exchanging upwelling thermal radiation with the warm planetary surface and in the latter by radiating to space.

The near-infrared bands of CO$_2$ around 1.4 $\mu$m provide relatively little direct solar heating in the lower atmosphere (around a few K d$^{-1}$; Savijärvi et al (2005)), where infrared optical depths are relatively high. But at altitudes above around 30 km absorption in these bands is quite substantial compared to other mechanisms, leading to dayside heating rates in excess of 20 K d$^{-1}$ (Savijärvi et al 2005). The processes involved are actually quite complex and, at even higher altitudes (above 60 km), can involve non-LTE (local thermodynamic equilibrium) effects due to direct radiative perturbations to the populations of key energy levels within the CO$_2$ molecule (e.g. Lopez-Puertas and Lopez-Valverde 1995, Lopez-Puertas and Taylor 2001). Such effects need to be taken into account for accurate computations of the thermal structure of the upper-middle atmosphere on Mars, e.g. in detailed climate models (Lopez-Valverde et al 1998, Forget et al 1999), but are negligible in the lower atmosphere.

2.1.2. Mineral dust. Although the gaseous components of the martian atmosphere are almost transparent to visible sunlight (at least in the lower atmosphere), some (variable) absorption and/or scattering of sunlight does typically occur due to the presence of suspended particulates, which consist either of mineral dust particles with radii of a few $\mu$m lifted from the surface or condensate cloud particles (made of H$_2$O or CO$_2$ ice). The radiative effects of suspended aerosol particles depend quite strongly on various factors, including the properties of the material (its complex refractive index as a function of wavelength), the distribution of particle sizes, the particle shapes, the bulk distribution of the aerosol within the atmosphere and the direction of illumination. Such effects are usually quantified with reference to the classical two-stream model of radiative transfer (e.g. Pierrehumbert 2010) in terms of the three basic scattering parameters, namely, the extinction coefficient, $Q_{ext}$, single scattering albedo $\omega$ and asymmetry parameter $g$, $Q_{ext}$ represents the effectiveness of the particles in removing photons from the direct beam, $\omega$ determines how reflective the particles are and $g$ how anisotropically light is scattered, such that $g = \pm 1$ corresponds to completely forward or backward scattering respectively and $g = 0$ represents perfectly isotropic scattering.

These properties are typically a strong function of wavelength. Their retrieval from remote sensing observations is a difficult and ill-constrained inverse problem and may depend sensitively on the assumed (or inferred) particle size distribution. Approaches adopted for determining the radiative
The properties of martian dust have varied over the years, starting in earnest with analyses of observations of planet-encircling dust storms during the Mariner 9 mission in 1971 (Toon et al. 1977) and the Viking Lander missions (Pollack et al. 1979). The properties obtained from optical depth and scattering measurements over a range of wavelengths and phase angles were compared with a number of possible terrestrial mineral analogue materials (such as granite, basalt, andesite) or mineralogical weathering products (such as montmorillonite or palagonite) but none were found to match the observations closely. This kind of approach was subsequently extended to use combinations of data from various spacecraft. Clancy et al. (1995) combined measurements from Mariner 9, the Viking Orbiter and Landers and the Phobos mission, while Ockert-Bell et al. (1997) focussed on Viking Lander and orbiter measurements alone, but both studies reached somewhat different conclusions, especially concerning the derived single-scattering albedo $\omega$ in the visible range. Such discrepancies turn out to have a significant impact on the implied solar heating rates due to the presence of suspended dust, so this quantity in particular is important to constrain accurately, especially for including in predictive climate models.

The most recent determinations (Wolff et al. 2006, 2009) make combined use of Mars Global Surveyor (MGS) and Mars Exploration Rover measurements in the thermal infrared and the Compact Reconnaissance Imaging Spectrometer (CRISM) in the visible and near-infrared to derive what are currently regarded as the best estimates so far of the size distributions and complex refractive indices for martian dust. The corresponding scattering parameters are shown in figure 3 and compared with those derived by Ockert-Bell et al. (1997). This clearly shows a reduction in $Q_{\text{ext}}$ across the visible spectrum in the new data and a slight increase in $\omega$, effectively ‘brightening’ the dust compared with the earlier estimate. Evidence for the improved accuracy of the new estimates comes from using them (in spectrally averaged forms) in parameterizations implemented in numerical climate models (e.g. Madeleine et al. 2011). Madeleine et al. (2011) obtained the best agreement with observations in the temperatures predicted in their model when they used the dust parameters derived by Wolff et al. (2006, 2009) and allowed the effective particle radius $r_{\text{eff}}$ to vary as dust was transported within the model. The agreement obtained was impressive, although it relies on constraining a number of parameters that are difficult to measure directly, so it is still not yet clear that this is a ‘solved’ problem.

2.1.3. Water and CO$_2$ ice clouds. Water ice clouds are observed to form in various places across the planet, often appearing as bluish hazes in visible images, though also forming as early morning fog at low levels (see e.g. Read and Lewis 2004). They are most commonly observed in association with large-scale topographic features, such as the Tharsis plateau, at mid-low latitudes during the aphelion season when the atmosphere is relatively cool and close to saturation (Smith

![Figure 3. Dust single scattering parameters for martian dust, as deduced by Ockert-Bell et al. (1997) (grey line) and by Wolff et al. (2006) (black line), showing (top to bottom) extinction efficiency factor $Q_{\text{ext}}$, single scattering albedo $\omega$ and asymmetry parameter $g$. Dashed curves in the background represent the normalized blackbody emission spectra for temperatures of 5870 K and 210 K, respectively (in an area-preserving representation). Adapted from Madeleine et al. (2011) figure 1.](image-url)
The clouds are typically quite tenuous and optically thin, with visible optical depths rarely exceeding 1–2 and typical particle sizes around 4 μm (Wolff and Clancy 2003). Relatively dense water ice clouds are also observed to form over the polar regions during autumn and winter, known as the ‘polar hood’ clouds (Benson et al. 2010, 2011), and some have even been observed to produce transient precipitation (snowfall!) with particle sizes up to 40 μm (Whiteway et al. 2009).

Until relatively recently, it was thought that such tenuous clouds could not have a significant impact on the thermal structure of the martian atmosphere. A number of recent studies, however, have indicated that water ice clouds can lead to discernible changes in temperature over the course of a martian day at around the 10 K level or even a little more at certain times and places. Variations in heating and cooling rates associated with such changes in temperature can actually lead to significant dynamical impacts, most notably on the thermal tides (Wilson et al. 2007), boundary layer and lower atmosphere (Wilson et al. 2008, Madeleine et al. 2012). Clouds in a Mars GCM have also been modelled in combination with data assimilation using thermal and dust opacity data as well as total column water vapour (Steele et al. 2014b) or just water ice opacity (Steele et al. 2014a). These simulations confirm the importance of clouds in the martian climate cycle.

Water ice cloud particles are quite different from mineral dust and are typically highly reflective in the visible part of the spectrum with a single scattering albedo close to 1. As a result, although they may scatter incoming sunlight they absorb relatively little and so do not lead to significant direct heating. They are, however, significantly absorbing in the infrared. This leads to local atmospheric heating during the day by absorption of upwelling infrared radiation from the planetary surface, and net local cooling during the night from cloudy regions by emission to space and the cooler ground below. Because clouds tend to form lower down at night than during the daytime, the net effect averaged over the day is to cool the atmosphere at low levels (below around 15 km altitude) and to heat it at higher levels (Madeleine et al. 2012).

Some of the detailed effects of radiatively active water ice clouds are illustrated with some insight in figure 4, which is taken from a series of numerical model simulations by Madeleine et al. (2012) that include the representation of a full hydrological cycle under martian conditions and permit the representation (or not) of the radiative effects of ice clouds. Each panel shows contours of the difference in temperature ΔTactive between a simulation that includes a full representation of radiatively active clouds Tactive and a corresponding simulation in which the clouds’ radiative effects are omitted Tinactive. The overall impact of including fully radiative clouds is illustrated in figure 4(a), which shows a time–height section at a location over the Tharsis plateau during the course of a day in northern summer. The diurnal formation and dissipation of clouds is shown by the light and dark shading, which represents relative concentration of cloud aerosol. Contours of ΔTactive illustrate the general cooling of the atmosphere below around 15 km altitude, a general warming between 15 and 40 km and a traveling wave-like disturbance at higher altitudes with a nominal period of 1 sol (defined here as the mean solar day). The latter indicates that the diurnal variations in cloud radiative forcing have the dynamical effect of enhancing the diurnal thermal tide in the upper atmosphere, even though there may be no discernible clouds at those altitudes. Figure 4(b) indicates the purely radiative impact of the clouds in the model, sufficient to produce local warming in excess of 10 K at certain times of day, that derives partly from the direct radiative effect of the clouds (figure 4(c)) and an indirect effect which is at least as large due to the redistribution of dust by perturbations to the dynamical wind fields. The impact of water ice clouds on the thermal structure of the martian atmosphere is therefore seen to be quite substantial, but also quite subtle and complex.

CO2 also acts as a condensable in the martian atmosphere, especially during polar night at high latitudes in winter. Clouds of CO2 ice crystals can form spontaneously when the atmospheric temperature falls below about 145 K, developing a polar hood cloud from which CO2 snowfall takes place (Pettengill and Ford 2000, Hayne et al. 2012) during autumn and winter seasons. The radiative properties of CO2 ice crystals are quite different from those of H2O ice, in that the former are relatively reflective at both visible and infrared wavelengths, except in the vicinity of the 15 μm absorption band (Pierrehumbert 2010). With particle sizes of order 10–100 μm, typical of well developed clouds, CO2 ice crystals are very efficient scatterers across almost the entire visible and

![Figure 4](image-url)
infrared spectrum. This can allow dense CO₂ clouds to backscatter upwelling thermal radiation from the surface, effectively increasing the greenhouse warming of the surface and lower atmosphere (Hunt 1980, Forget and Pierrehumbert 1997, Pierrehumbert 2010). Such effects have been suggested to be important for early martian climate (Forget and Pierrehumbert 1997), though are perhaps less significant in the present martian climate apart from during polar winter. These radiative effects on atmospheric energetics are further supplemented by latent heat effects, as discussed further below.

2.2. Thermal structure in radiative–convective equilibrium

Given the relevant sources of opacity in the atmosphere, convolved with the spectral distribution of insoluation and thermal emission within the atmosphere and surface, it is straightforward in principle to compute the thermal structure that would exist in radiative equilibrium (such that the divergence of net radiative flux is zero). Such a radiative equilibrium represents a pivotal state of the atmosphere towards which its thermal structure will tend, all other things being equal and given sufficient time. The latter depends on the radiative relaxation timescale, which represents the time taken for atmospheric temperature to adjust by local absorption and emission of radiation alone. This can be readily estimated for a simple gray atmosphere (e.g. see Houghton 1986, James 1994) as

\[
\tau_{\text{rad}} \approx \frac{c_p \rho h}{4(2 - \eta)\sigma T^4} = \frac{\rho}{4(2 - \eta)\Gamma \sigma T^4}, \tag{1}
\]

where \(c_p\) is the specific heat capacity of air at constant pressure, \(\rho\) the density, \(h\) the atmospheric scale height, \(\sigma\) the Stefan–Boltzmann constant, \(\eta\) the infrared opacity of the atmospheric layer (= 1 − \(T\), where \(T\) is the transmissivity of the layer and 0 ≤ \(T\) ≤ 1) and \(\Gamma\) the adiabatic lapse rate (\(= g/c_p\)). For the Earth’s troposphere, typical values suggest \(\tau_{\text{rad}}\) is of order 20–30 d or so, but for Mars, for which \(\rho\) is much smaller than in the Earth’s troposphere, this is around 0.5 sols. Such a value is consistent with more accurate and sophisticated approaches (e.g. see Eckermann et al 2000, Nayvelt et al 1997) which also indicate timescales ranging from ~1 sol on large spatial scales to a few hours for small scale gravity waves. For the Earth’s troposphere, \(\tau_{\text{rad}}\) is rather longer than typical advective timescales (\(LU\), where \(U\) is a typical horizontal velocity and \(L\) a typical length scale), suggesting that motions in the Earth’s lower atmosphere are relatively unconstrained by radiative effects on moderate timescales. But on Mars \(\tau_{\text{rad}}\) is typically short compared with \(LU\) and so radiative effects may be expected to be much more important on short timescales, with motions being acted on by a much tauter ‘radiative spring’ (see the Earth’s upper stratosphere, where radiative timescales may be as short as ~5 d; e.g. Dickinson (1973)).

2.2.1. Semi-gray radiative–convective equilibrium. The previous subsection indicated that the spectral distribution of opacity in the martian atmosphere is quite complicated. But in fact many of the gross features of the atmospheric structure on Mars are reasonably well captured by a model that assumes the atmosphere is composed of a relatively simple, ‘semi-gray’ absorber. In the absence of dust, the martian atmosphere is reasonably well approximated by an absorber that is completely transparent in the visible part of the spectrum and acts as a simple, non-scattering gray absorber in the infrared. This is a classical problem to which the solution is readily obtainable analytically (e.g. Houghton 1986) in the form

\[
\sigma T^4(z) = \frac{S_0}{2} (1 + D \mu \chi_L(z)), \tag{2}
\]

where \(S_0\) is the solar irradiance at the top of the atmosphere, \(\mu = \cos \theta\) (where \(\theta\) is the solar zenith angle of incidence), \(D\) is a diffusivity factor and \(\chi_L(z)\) is the infrared optical depth from height \(z\) to the top of the atmosphere. This represents the simplest situation in which the surface is a pure blackbody with surface albedo \(A_s = 0\).

This approach can be generalized fairly straightforwardly (e.g. Guillot 2010, Pierrehumbert 2010) to represent an atmosphere that overlies a surface of non-zero \(A_s\) and which also has a non-zero opacity in the visible part of the spectrum due to an absorber which is uniformly mixed throughout the atmosphere with optical depth \(\chi_S(z) = \gamma \chi_L(z)\), to give

\[
\sigma T^4(z) = \frac{S_0}{2} \left[ 1 + \frac{D u}{\gamma} + \left( \frac{\gamma}{D \mu} - \frac{D u}{\gamma} \right) \exp \left\{ -\frac{\gamma \chi_L(z)}{\mu} \right\} \right] + (1 + D \chi_L(z))(\gamma - 1) A_s \exp \left[ -\frac{\gamma \chi_L(0)}{\mu} \right], \tag{3}
\]

which assumes a constant ratio between \(\chi_L\) and \(\chi_S\) for simplicity.

Figure 5(b) shows an early calculation of a radiative–convective equilibrium profile for a pure CO₂ atmosphere under martian conditions by Gierasch and Goody (1972). This shows the clear trend for temperature to decrease monotonically with height from the ground during daytime, with a sharp jump in temperature at the surface itself. In the lowest 10–15 km of the atmosphere this temperature structure is actually convectively unstable during this time and so the thermal gradient follows closely the adiabatic lapse rate \(\Gamma = g/c_p\) until the optically thin region above 15 km, where the profile becomes almost isothermal. During the night the surface cools rapidly, forming a shallow temperature inversion. This basic structure is also evident in the analytical profiles shown as the dashed-dotted/ dotted lines in figure 5(a), which assume \(A_s = 0.2, \chi_S = 0\) and \(\chi_L(0) = 1.25\). Temperature values in the analytical profile do not accurately reflect the pure CO₂ profile, however, because this analytical profile assumes a fixed absorption coefficient and an optical depth that scales linearly with pressure \(p\). This is not particularly accurate for CO₂ because the opacity structure in the infrared consists of discrete bands separated by a relatively transparent continuum (see figure 2) so the pressure dependence of \(\chi_L\) is more complicated. An important implication of these simplified solutions for thermal profiles, however, is that neither account fully for the observed structure of the atmosphere under typical martian conditions (e.g. see figures 5(c) and (d)), which is typically significantly warmer at low-mid levels compared with the profiles computed for a clear atmosphere.
If an additional source of opacity in the visible region is added to the atmosphere, however, the thermal profile will change significantly. Figure 5(a) illustrates a profile in which a second gray absorber has been added with an optical depth $\chi^{S_2}$ in the visible and $\chi^{S} = 0.65$ in the infrared, which roughly emulates the effect of adding a well mixed dust component. As one might anticipate, increasing $\chi^{S_2}$ has the effect of warming the mid-level atmosphere. For higher values of $\chi^{S_2}$ (not shown) the solution even produces a significant temperature inversion, such as might be encountered during a major dust storm on Mars. Such thermal inversions during dust storms are commonly encountered in practice (e.g. Smith et al. 2002) and may have an important role in stabilizing the atmosphere and bringing about the decay of the dust storm by shutting off the radiatively driven near-surface wind forcing that raises dust. Further evidence that this simple approach to modeling captures the gross features of the thermal structure of the martian atmosphere via the interaction of solar and thermal radiation with idealized, well-mixed gray absorbers can be seen in figures 5(c) and (d), which show some typical observations of vertical temperature profiles on Mars obtained from either radio-occultation techniques or during entry, descent and landing of the two Viking Lander spacecraft, Mars Pathfinder, the Mars Exploration Rovers and the Phoenix Lander.

Figure 5. (a) Example temperature profiles in radiative (solid or dash-dotted lines) and radiative–convective (dashed or dotted lines) equilibrium for a semi-gray atmosphere with the same mean solar irradiance as for Mars, but for solar absorption optical depth $\tau = 0$ (dust free, dash–dotted line) or 0.1 (notionally due to dust in the atmosphere, solid line) and solar zenith angle of 0°. These profiles assume a gaseous optical depth $\chi_L^{S_1}$ of 1.25 in the infrared, a total infrared optical depth as given by equation (4), and a surface albedo $A_s = 0.2$ (see text). (b) Temperature profiles as a function of height as derived for a pure CO$_2$ atmosphere by Gierasch and Goody (1972) at 40° latitude during summer conditions at local times of 06:00 and 16:00. The stippled region indicates the range of observed temperatures at this latitude. (c) Temperature profiles as a function of height as derived from radio occultation data obtained from NASA’s Mars Global Surveyor spacecraft (Hinson 2006). Late-afternoon temperatures are shown from southern mid-latitudes during summer. Nighttime temperatures are also shown from mid-latitudes during summer. (d) Temperature profiles observed during the entry, descent and landing of the two Viking Lander spacecraft, Mars Pathfinder, the Mars Exploration Rovers and the Phoenix Lander.
though temperatures in the upper atmosphere tend to reach rather lower values (110–160 K) than the simple gray model profiles would suggest. This almost certainly reflects various shortcomings of the gray atmosphere approximation, either because of the existence of spectral ‘windows’ where the atmosphere is relatively transparent at thermal wavelengths, or because key absorbers (especially dust) are not well mixed in altitude.

Despite these shortcomings, the gray atmosphere approach can be followed through to take account of differing solar zenith angles of insolation, due either to differing latitudes, time of day and/or differing seasonal conditions. Figure 6 shows a number of examples of radiative–convective equilibrium temperature distributions as a function of latitude and height representative of northern spring equinox ((a), (d) and (g)), summer solstice ((b), (e) and (h)) and winter solstice ((c), (e) and (i)), as computed by Zalucha et al (2010). In these examples, profiles were initially computed in local radiative equilibrium with diurnally averaged sunlight incident at the appropriate solar zenith angle, and then (as in figure 5(a)) adjusted to eliminate convectively unstable parts of the profile by conservatively mixing specific entropy, $s$, or potential temperature, $\theta$. The latter is defined for a parcel of dry air, currently at temperature $T$ and pressure $p$, as

$$\theta \equiv T(p_0/p)^{R/c_p}$$

with $p_0$ the reference pressure, $R$ the specific gas constant and $c_p$ the specific heat at constant pressure of the appropriate air mixture (such that $s = c_p \ln \theta$). Potential temperature is of great significance since profiles of $\theta$ indicate the static stability of the air: if $d\theta/dz < 0$ then air is unstable to convection, which would act rapidly to bring the profile back towards neutral stability, $\theta = \theta_0$, as implemented as an adjustment process in radiative–convective models such as by Zalucha et al (2010).

The model results in figure 6 broadly indicate the thermal structures towards which the atmosphere is forced by purely radiative processes. The second and third rows further take into account the effects of the zonally averaged surface topography, which exhibits some substantial hemispheric asymmetries on Mars and can significantly affect the intensity of mean insolation incident on surfaces with substantial slopes compared to those that are completely flat. A number of features are clearly seen in these calculations. At equinox the atmosphere is radiatively forced towards a thermal state that is more or less symmetric about the equator, with strong vertical temperature gradients near the ground in the tropics, substantial horizontal thermal gradients at mid-high latitudes.
and almost isothermal tropics at altitudes greater than around 15 km. At the solstices, atmospheric and surface temperatures increase monotonically from the winter pole to the summer pole at almost all heights, with strong vertical temperature gradients in the summer hemisphere at low levels and strong horizontal temperature gradients at mid-latitude in the winter hemisphere at all altitudes. The region of atmosphere overlying the winter polar region becomes almost isothermal, with temperatures decreasing with height to <135 K. Such temperatures are actually colder than the frost point of CO$_2$, which would therefore lead to condensation of dry ice throughout the polar atmosphere in winter, much as found in observations.

Such trends are broadly reflected in the observed temperature structures at these seasons (see figure 1), though with some important differences. In particular, polar winter temperatures are not quite as low as these gray radiative–convective equilibrium calculations would suggest, though temperatures below the CO$_2$ frost point have been measured (Colaprete et al. 1997). Also, there is no evidence in the radiative–convective temperature distributions for localized temperature maxima over the polar regions in the middle atmosphere, at altitudes of 40–80 km. Such features are clearly seen in figure 1 and appear to be robust structures that are regularly found in observations at almost all seasons. These aspects cannot be accounted for by radiative processes alone, and imply that dynamical motions are playing an important role through the effects of adiabatic compression or rarefaction through upwelling or downwelling. More subtle features over the tropics at high altitude are also not consistent with pure radiative forcing, clearly demonstrating the need to take dynamics into account (e.g. see McCleese et al. 2008).

2.3. Convective adjustment

In deducing thermal structures in the vertical based on radiative transfer computations, it is common practice to take into account the possibility of the creation of convectively unstable sections of a vertical profile by allowing for the effects of rapid convective mixing. On Earth, this typically entails an assumption that an initially super-adiabatic lapse rate (by which is meant a thermal gradient steeper than the dry adiabatic lapse rate—e.g. Andrews (2001), Holton (2004), Houghton (1986), James (1994)) will rapidly induce vigorous convective overturning, leading to a well-mixed state with a uniform vertical distribution of specific entropy and other conserved tracers. Atmospheric models do not usually attempt to represent the convective mixing explicitly, but rather assume that such mixing takes place rapidly compared to other processes, and immediately adjust the vertical profile towards a state that approximates closely to a neutrally stable profile, a process often referred to as convective adjustment. Such a procedure was used, for example, to generate the radiative–convective equilibrium profiles shown in figures 5(a), (b) and 6 above.

Near the martian surface, however, super-adiabatic layers with steeply decreasing temperature profiles with height are frequently observed (Schofield et al. 1997, Smith et al. 2006), indicating that convective mixing is often incomplete. This is largely due to three major factors: (a) the low density of the near-surface atmosphere, which limits the heat-carrying capacity of the air; (b) the increased relative importance (compared with the Earth’s atmosphere) of direct radiative heating of the boundary layer through absorption of thermal radiation by CO$_2$ and dust; and (c) the low thermal inertia of the surface, which results in large excursions of surface temperature during the diurnal cycle (Haberle et al. 1993, Savijärv 1999, Petrosyan et al. 2011). The overall effect on the typical thermal structure of the atmosphere is to produce sustained regions of super-adiabatic thermal gradients and the development of a very deep, convective boundary layer (Hinson et al. 2008, Spiga 2011). Large-eddy model simulations (LES) suggest that, throughout most daylight hours, the convection takes the form of intense upwelling plumes of buoyant air with powerful updrafts of 10–20 m s$^{-1}$, interspersed with broader and weaker regions of downwelling with vertical velocities of 5–10 m s$^{-1}$, often organized into polygonal cellular patterns (Michaels and Rafkin 2004, Spiga and Forget 2009).

Figure 7 illustrates an example of such plume-dominated convection, taken from an LES model simulation of boundary flow by Colatis et al. (2013) under martian conditions. The asymmetric structure of upward plumes and broader downwelling is clearly apparent up to the level of the capping inversion at an altitude of around 7 km (level 100). Similar plume-like features are also seen, for example, in mini-TES observations of thermal structures in the lower boundary layer (Smith et al. 2006). This pattern of flow is associated with a mean thermal profile of the form illustrated in figure 8, which clearly shows a super-adiabatic lapse rate from the surface to around 2–3 km altitude, topped by a near-neutral entrainment layer beneath the stable capping inversion. The profile corresponding to simple adiabatic convective adjustment is shown as a thick solid line, from which it is clear that such an assumed profile leads to an underestimate of low-level atmospheric temperatures by about 5 K.

3. Large-scale dynamics and transport

Previous sections have reviewed the martian atmosphere as a whole, its structure and the radiative processes that drive it. Attention is now turned to the large-scale atmospheric motions that result and which give Mars its weather patterns. Winds are an essential component of the martian climate system, not only for their ability to transport heat and momentum, but also constituent gases and aerosols that can alter the radiative balance of the atmosphere and its thermal structure. This, in turn, changes the winds and leads to feedback. As will be discussed in later sections, the atmospheric winds also lift dust from the surface and can enhance the transfer of volatiles from near the surface into the free atmosphere. No component of the weather and climate system can be considered in isolation.

3.1. The global circulation

The seasonal mean circulation observed on Mars and Earth is determined by the response of the atmosphere and surface to...
the strong forcing by solar radiation, both through direct heating of the surface and internal heating, especially when the aerosol opacity becomes thick enough to absorb significant visible and infrared radiation, as is often the case on Mars. The circulation responds to the astronomical variations of the solar heating as the sub-solar point moves over a similar range in latitude. This is a consequence of the presently similar (non-zero) values of the obliquities of both planets, 25.2° for Mars versus 23.9° for Earth. The absence of oceans on Mars means that the thermal inertia of the surface, and so its response time to changes in heating, is much smaller than that of the Earth on average. This leads to the warmest part of Mars’ surface moving to very high latitudes in the summer hemisphere and to a much more extreme asymmetry between the summer and winter hemispheres. The martian summer hemisphere is typically characterized by a statically stable atmosphere (above...
the daytime super-adiabatic surface layer; see figure 8), weak horizontal temperature gradients and relatively weak winds in the lower atmosphere. In direct contrast, the winter hemisphere is characterized by almost neutral stability, a strong equator-to-pole thermal gradient and a strong westerly jet. This pattern is much more strongly asymmetric than that on Earth, although winter hemisphere westerlies do tend to be somewhat stronger and the summer atmosphere tends to be a little more stable.

Another seasonal effect arises as a result of Mars’s orbital eccentricity: the eccentricity of Mars’s orbit is 9.3% versus 1.7% for Earth. The increased eccentricity of Mars’s orbit results in significantly greater heating, and consequently a stronger circulation, near perihelion (which occurs just before northern winter solstice at $L_S = 251^\circ$, an effect that is much harder to detect in Earth’s atmosphere.

3.1.1. Longitudinally averaged flows and seasonal change. We first consider the circulation at equinox, a time when the martian atmosphere is most symmetric between the hemispheres. The martian atmosphere, averaged over all longitudes and 30 degrees of areocentric longitude (about 60 days of time) just following the northern hemisphere spring equinox, is shown in figure 9. This figure is taken from global circulation model results, but using a model that has been tuned very closely to match the observed climatology and has then had martian dust and thermal observations assimilated into the lower atmosphere, below about 40 km (Lewis et al 2007, Lewis 2010). The advantage of using model-derived results is not only a uniform coverage in time and space, but a fully balanced set of winds whereas remote observations tend to be retrievals of temperature from which winds may only be inferred, often by applying a simplified balance condition (e.g. Conrath et al 2000, Smith 2004, McCleese et al 2010).

The state shown in figure 9 is somewhat similar to the Earth at most times of year, with a warm equator and westerly winds in each hemisphere. A similar equinoctial state is repeated near northern hemisphere autumn equinox. The increasingly westerly (eastward, or positive in the right-handed coordinate convention) winds with height at middle to high latitudes in both hemispheres (a positive vertical gradient in the zonal wind in the lower and middle atmosphere) are linked to the horizontal thermal gradients from equator to pole. This is a result of the dominant balances in the vertical and horizontal.

In the vertical, there is a hydrostatic balance between gravity and the pressure gradient. In the horizontal, there is an approximate geostrophic balance between pressure gradients and the Coriolis force, which is a result of the planet’s rotation. The geostrophic approximation holds to an accuracy of about 10% on both Earth and Mars for typical large-scale winds at mid-latitudes, with the result that horizontal temperature gradients are linked to vertical wind gradients, known as thermal wind balance (e.g. Andrews 2001, Holton 2004, Vallis 2006). The reduction in the westerly jet magnitudes in the upper part of

Figure 9. Zonal-mean (longitudinally averaged) atmospheric temperature (shaded colours) and winds (positive/solid contours show eastward or westerly winds, negative/dashed contours show easterly winds), at northern spring equinox ($L_s = 0^\circ$). The plot is on a log-pressure vertical axis, but a pseudo-altitude axis based on a 10 km scale height is shown for comparison on the right-hand side; this is less accurate at high altitudes where the scale height can fall to roughly 7 km at the colder temperatures. The black line and masked region at the bottom of the figure indicates the zonal-mean surface pressure.
Figure 9 is matched by a reversal in the horizontal temperature gradient, with warmer air high over the poles around 80 km altitude than over the equator at the same altitude.

An equinoctial state on Mars is only a relatively brief, transient phenomenon, however, as the atmosphere moves between solstitial states. Figure 10 shows zonal-mean temperatures and winds for a similar period of 30 degrees of areocentric longitude immediately following northern hemisphere winter solstice. This is close to the time that Mars is nearest to the Sun, and is the period of the most extreme contrast between hemispheres with a warm southern hemisphere, cold northern polar regions, with the exception of the more pronounced warming that occurs in the middle atmosphere at 50–80 km, and strong westerly winds in the winter hemisphere.

3.1.2. Hadley circulations and meridional overturning. The previous section showed longitudinally averaged eastward wind and temperature. It is also worthwhile to consider the mean meridional mass streamfunction, which shows vertical and latitudinal transport averaged over all longitudes. Figure 11 shows this circulation for an equinoctial period corresponding to figure 9. This reveals a roughly Earth-like situation with rising flow roughly at the equator and descending flow at about 30°–60° latitude away from the equator at both equinoxes, reminiscent of the well-known Hadley Cells on Earth. As on Earth, the cells are not quite of equal strength, depending on the exact averaging period chosen. The most striking contrast between the planets is that there is no strong tropopause and stratosphere on Mars and so the cells are free to extend higher in the atmosphere without clearly defined tops (which occur at about 15 km on Earth). A counter-rotating cell at about 50° N is really an artefact of the Eulerian averaging process, sometimes known as the ‘Ferrel Cell’ on Earth, and is balanced by wave transports at these latitudes (see the following subsection). It would not appear in a true Lagrangian average.

The mean meridional circulation reveals the reason for the westerly jets at middle and high latitudes in figure 9. Air from near the surface at the equator has a very small wind speed, close to zero. When transported poleward in both cells, the air moves nearer to the rotation axis of the planet. The air must conserve its angular momentum and so it tends to rotate about the axis more rapidly than the planet’s surface below. As seen from the reference frame of the planet, this is a westerly wind. This argument has assumed that all the motion is longitudinally symmetric, which is certainly not true, as will be seen, but it explains some of the broad features of the circulation.

Figure 12 shows the same circulation for the solstitial period corresponding to figure 9. This solstitial circulation is dominated by a single, equator-crossing cell on Mars that covers at least 90° of latitude and is typical of much of the year, although this is the most extreme example, thanks to the strong heating in the southern summer hemisphere, arising through a combination of peak insolation and an increase in atmospheric dust loading. It should be noted that the cell transports much more mass than the equinoctial cells in figure 11. It is also worth noting that, averaged over the course of a martian year, this cell tends to dominate even over the oppositely

![Figure 10. Zonal-mean atmospheric temperature and winds, as figure 9, but at northern winter solstice (Ls = 270°), when the subsolar latitude is ≃25° S.](image)
rotating northern hemisphere summer cell that lasts for longer in the year but which is somewhat weaker. This is a result of the hemispheric asymmetry in topographic height, visible as the zonal-mean surface pressure line in figures 9–12, with a slope downwards on average from the southern to the northern hemisphere.

This extensive southern summer solsticial overturning cell, stretching from southern to northern mid-latitudes, enables transport of tracers between hemispheres, which in the case of dust can lead to the rapid expansion of localised lifting events in the southern hemisphere into major dust storms that encompass almost the entire planet. Dust storm growth is boosted
further by the existence of a positive radiative feedback, in which atmospheric heating by lifted dust strengthens and expands the large-scale circulation, increasing near-surface wind speeds and encouraging further dust lifting (Haberle et al. 1982).

3.1.3. Super-rotating jets. The origin of the westerly jets in an angular-momentum-conserving circulation was briefly discussed in the preceding section. It is of interest to consider this a little further since sometimes the angular momentum of the martian (and the terrestrial) atmosphere may exceed that of air at rest at the equator. This is known as local super-rotation (Read 1986). Lewis and Read (2003) demonstrated that the martian atmosphere super-rotates particularly when in a dusty state over the equator. This does not occur at the latitude of the highest westerly jet speeds, which are still sub-rotating compared to air at rest at the equator brought without loss of momentum to their latitude, but at low latitudes and even over the equator itself where positive, westerly winds immediately indicate super-rotation.

Figure 13 shows the zonal mean winds and the local super-rotation index, $s$, again from an assimilated dataset (Montabone et al. 2014), from a period in the autumn of Mars Year 25, just after the June 2001 planet-encircling dust storm had commenced and when the opacity of the martian atmosphere was very high. Local super-rotation, $s$, is defined as the ratio of the excess angular momentum of the atmosphere about the rotation axis of the planet to the angular momentum of air at rest relative to the surface at the equator. For an atmosphere in solid-body rotation with a planet, $s = 0$ at the equator, falling to $s = -1$ at the poles. Positive values of $s$ indicate atmospheric super-rotation, which is not possible to achieve by rearranging an atmosphere initially at rest by axisymmetric processes alone (Hide 1969). This is an extreme example that clearly shows a super-rotating equatorial jet that was absent in the previous year when dust loading was much lower at this time.

As noted, this state could not be achieved by axisymmetric transport in the mean meridional circulation. Waves or eddies are required to transport angular momentum up-gradient into the jet. Lewis and Read (2003) showed, using idealised model experiments, that a state of this type could be achieved on Mars as a result of forcing by the generation of thermal tides, to be discussed in more detail in a following section, and linked, as expected, to the amount of dust suspended in the atmosphere.

3.2. Baroclinic weather systems

3.2.1. Observations. Since the time of the early Viking missions to Mars (Ryan et al. 1978, Tillman et al. 1979, Barnes 1980, 1981) it has been observed that active weather systems occur at northern mid-latitudes during the autumn, winter and spring seasons. These are evident in a variety of ways, including strong, travelling storm-like perturbations to surface wind, pressure and temperature (e.g. Barnes 1980, 1981), the appearance in images from Mars orbit or the Hubble Space Telescope of tenuous water ice clouds in spiral fronts (Briggs and Leovy 1974, Gierasch et al. 1979, Hunt and James 1979, Malin et al. 2008) and the organization of circumpolar dust storms (James et al. 1999, Malin et al. 2008, Cantor et al. 2010). An example of the latter is graphically illustrated in figure 14, which shows a sequence of images from the Mars Orbiter Camera (MOC) on board the Mars Global Surveyor spacecraft taken roughly every 2 h during the development of an intense baroclinic cyclone wave near Mars’ north polar ice cap. Low-level winds are concentrated into narrow fronts (much like the cloudy frontal structures seen in satellite images of
mid-latitude cyclones on the Earth) where dust is lifted from
the surface to form a pattern of dust storms that visualise the
large-scale flow.

More detailed information on the structure of these sys-
tems has been provided in atmospheric remote sounding
observations from Mars orbital platforms such as MGS
show them to be large-scale eastward-travelling waves that
extend upwards from the surface. Figure 15 illustrates a typi-
cal structure observed during northern hemisphere spring in
radio- occultation measurements (Hinson 2006), showing the
flow to be dominated by a zonal wavenumber 3 pattern in geo-
potential height that slopes towards the west with height, but
with a temperature structure with two amplitude peaks, one
close to the surface and another at higher altitudes (around
30 km). Infrared remote sounding techniques (Banfield
et al 2004) provide more complete spatio-temporal coverage
and confirm the presence of such traveling waves with typi-
cal zonal wavenumbers of $m = 1$–3 and comparable vertical
structures to those shown in figure 15 at this season.
The occurrence of these strong traveling waves varies substantially on seasonal timescales, being largely present around winter and adjacent seasons in their respective hemisphere. The waves in the southern hemisphere, however, seem to be somewhat weaker than in the north, and do not develop across all longitudes (Wilson et al. 2002, Banfield et al. 2004). This may be related to hemispheric differences in surface topography, as well as to the weaker solar irradiance and therefore smaller thermal gradients in southern winter, although the phenomenon is still not well understood in detail.

3.2.2. Origins as baroclinic and/or barotropic instability? These traveling wave disturbances are most probably owe their origin to a natural instability of the basic, thermally-driven zonal flow, either as a barotropic shear instability, a potential energy–releasing baroclinic instability, or a mixture of the two.

Barotropic instability occurs when eddy-like perturbations are able to organize themselves with respect to a laterally-sheared zonal flow so as to transport zonal momentum away from the peak of the zonal jet, thereby gaining kinetic energy at the expense of the zonal flow. In practice, this is achieved by wave-like perturbations that ‘tilt’ in the north–south direction to ‘lean into’ the shear of the zonal flow so that the covariance $\overline{u'v'}$ results in a divergent meridional flux of zonal momentum and a pattern of Reynolds stresses that decelerates the zonal flow. A necessary (though not sufficient) condition for barotropic instability is the so-called Rayleigh or Rayleigh–Kuo criterion (e.g. Holton 2004, Vallis 2006), which states that $\partial \Phi / \partial y$ (the lateral gradient of total absolute vorticity),

$$\bar{q} = f − \partial \Pi / \partial y,$$

where $f = 2\Omega \sin \phi$, $\phi$ is latitude and the overline indicates a zonal mean) changes sign within the domain of interest, i.e.

$$\frac{\partial \pi}{\partial y} = \beta - \frac{\partial^2 \bar{u}}{\partial y^2} < 0,$$

where $y$ is the northward coordinate, $\bar{u}$ is the zonally averaged zonal flow and $\beta$ is the planetary vorticity gradient, i.e. the rate of change of Coriolis parameter with latitude that occurs in a spherical planetary atmosphere (e.g. Holton 2004, Vallis 2006).

The selection of the dominant scale of the instability involves a number of factors, including the degree of supercriticality and the intrinsic lateral scale of the unstable jet. The simplest inviscid, linear instability theory (e.g. Howard and Drazin 1964) predicts a wavelength of maximum growth rate $2\pi \sim L \sqrt{2}$, where $L$ is the lateral width of the jet. This works reasonably well as a lower limit for the wavelength of the dominant fully developed instabilities, but more sophisticated theories (e.g. Niino and Misawa 1984) indicate that other factors will modify this simple result, such as jet curvature and internal viscosity. Pure barotropic instability, however, is favoured only when the basic shear flow does not depend on altitude, which is seldom the case in a planetary atmosphere such as on Mars, where thermal effects result in more complicated vertical structures and more mixed forms of instability.

As on the Earth, the imbalance in solar irradiation and thermal infrared emission between the tropics and higher latitudes on Mars leads to a systematic decrease in temperature from equator towards the poles in both hemispheres near equinox and in the winter hemisphere otherwise. In combination with a broadly stable stratification in the vertical direction (promoted in part by buoyant transports in the low-latitude Hadley circulation) and rapid planetary rotation, this implies conditions favourable for the development of so-called baroclinic instability (e.g. Hide and Mason 1975). This is a wave-like instability of a usually axisymmetric flow that has the capacity to release potential energy associated with sloping isopycnals (isotherms or isentropes) that are statically stable in the vertical direction, provided some key vorticity constraints are satisfied. Although the quantitative details depend to a significant extent on the precise configuration of the initial flow (background vertical and horizontal wind shear, static stability structure, boundary conditions etc.), in many circumstances the simplest possible theoretical model for baroclinic instability due to Eady (1949) is sufficient to estimate the likely favoured length scales and growth rates.

Assuming infinitesimal distortions to a uniform (linear) vertical shear flow in a zonal channel with uniform static stability (characterized by buoyancy frequency $N$, zonal velocity scale $U$ and vertical length scale $D$), this model predicts the horizontal wavelength of the most rapidly growing wave mode to be $\lambda_{\text{max}} \sim 3.9ND / f_0$ (where $f_0 = 2\Omega \sin \phi_0$ and $\phi_0$ is the latitude of the centre of the channel). The corresponding exponential growth rate $kC \sim 0.31(f_0 / N)U / Dz$. For the Earth’s mid-latitudes, for example, $N \sim 10^{-2} \text{s}^{-1}$, $D \sim 10 \text{km}$, $U \sim 50 \text{m s}^{-1}$ and $f_0 \sim 10^{-4} \text{s}^{-1}$, suggesting $\lambda_{\text{max}} \sim 4000 \text{ km}$ and a growth timescale $1/(kC) \sim 0.75 \text{ d}$. Such figures are remarkably close to the scales for developing baroclinic weather systems on Earth, and this success of Eady’s model (and that of his contemporary (Charney 1947)) was a major factor in enabling the baroclinic interpretation of the origin of mid-latitude weather systems to gain acceptance in the 1950s and 1960s.

Hess (1950) used similar considerations (and some ingenious ways of estimating $N$ and other parameters) to suggest that the martian atmosphere might exhibit traveling wave-like weather systems, in many ways similar to Earth, on similar length and time scales. Later, Leovy (1969) calculated that baroclinic waves with zonal wavenumbers of 2–4 would have the highest growth rates at mid-latitudes in a simple, two-layer quasi-geostrophic model. He also suggested that the short radiative damping timescale on Mars would not significantly affect the length scales of the waves or their growth rates. Around the same time, Leovy and Mintz (1969) adapted an early (primitive equation) Global Circulation Model to martian conditions (but without topography). They found baroclinic traveling waves in their model simulations with length scales similar to those predicted by Leovy (1969) that were responsible for significant transports of angular momentum and heat in the winter hemisphere near solstice, and in mid-latitudes in both hemispheres at equinox.

Blumsack and Gierasch (1972) examined the effects of the substantial topographic slopes found on Mars at planetary
scales on the traveling waves. They found that the slopes could reduce both the length scales and growth rates of the most unstable waves. Pollack et al. (1981) extended the general circulation model of Leovy and Mintz (1969) to include topography (as well as it was known at the time) but found few differences in the transient waves, somewhat in conflict with the predictions of Blumsack and Gierasch (1972). More recent modeling work, however, has demonstrated substantial topographic impacts, including the ability of topographic features to steer and focus baroclinic instabilities into distinct ‘storm zones’ (Hollingsworth et al. 1996), and differences in the amplitudes of these transient waves between the northern and southern hemispheres that can be attributed to variations in topography (Mulholland et al. 2015).

The realistic background zonally averaged flow departs significantly from the idealized form assumed in this class of simple models. Figure 16 illustrates the typical structure of such a flow during northern hemisphere winter, showing a very intense, eastward jet at upper levels with strong shear in both the vertical and horizontal directions. This results in a complex pattern of potential vorticity that strongly influences the style and character of the resulting instabilities, including the possibility of both baroclinic and barotropic instabilities or some combination thereof. Barnes (1984), for example, used a more accurate representation of the zonal mean state in a quasi-geostrophic numerical model to confirm that zonal wavenumbers of 2–4 are favoured. In an alternative approach, Michelangeli et al. (1987) considered the barotropic instability of the winter polar jets for Mars, finding that waves with zonal wavenumbers of 1 or 2 and periods of 0.75–2.5 sols are favoured. Such barotropic perturbations would be mainly located well above the surface, where an orbital atmospheric sounder would be the ideal instrument to detect and measure them.

These length and timescales, and more detailed wave structures, were confirmed in the most sophisticated linear instability calculations by Tanaka and Arai (1999) of essentially mixed baroclinic–barotropic instabilities of zonal jets with realistic lateral and vertical shear. The results (see figure 17) clearly show the longest wavelength unstable modes penetrating most deeply through the atmosphere with a secondary maximum in amplitude near the surface, but with the strongest zonal phase tilt with height close to the surface and relatively barotropic (height-independent) structure at higher altitudes. Shorter wavelength disturbances are found to be much shallower, with a surface-intensified, strongly baroclinic structure (vertically tilted) and no upper level maximum. The wave-number of maximum overall growth rate turned out to be around $m = 5–6$, although all modes with wavenumbers $< 9$ were typically found to grow.

---

**Figure 16.** Latitude height sections of the zonally averaged temperature (upper left), zonal velocity (lower left), potential vorticity (a generalization of absolute vorticity; see equation (6) and Holton (2004), Vallis (2006)) (upper right) and northward gradient of potential vorticity (lower right), derived by Banfield et al. (2004) from MGS TES observations.
In practice, however, the fully developed flow is most commonly found to be dominated by $m = 1\text{--}3$ disturbances, which almost certainly results from the nonlinear development of the flow when the unstable waves grow to large amplitude (the theory above is only valid for perturbations of infinitesimal amplitude).

This behaviour is clearly apparent in fully nonlinear model simulations in numerical circulation models (e.g. Barnes et al. (1993), Hourdin et al. (1995), Collins et al. (1996) and Wilson et al. (2002)). Barnes et al. (1993) found that zonal wavenumbers 1 and 2 dominated the higher altitudes, with $m = 1\text{--}4$ having significant amplitude near the surface. The southern hemisphere winter’s transient waves were weaker than those of the north by a factor of 2 or more. Hourdin et al. (1995) noted a seasonal dependence in the variance of surface pressure in the Viking Lander data on timescales of a few sols, and had some success reproducing this dependence with the LMD global climate model (GCM). Northern transient wave surface pressure variance showed a two-peak structure with a relative minimum at northern winter solstice, the depth of which Hourdin et al. showed to increase with the atmospheric dust opacity (see also recent work by Lewis et al. (2015) and Mooring and Wilson (2015)).

3.2.3. Regular or chaotic waves and nonlinearity? Motivated by the cyclic patterns in Viking Lander observations (Barnes 1980, 1981), Leovy (1981a) noted the apparent regularity of martian baroclinic waves compared to those on Earth. In particular, time series of measurements of surface wind, pressure and temperature from the higher latitude Viking Lander (Lander 2, at $\phi \sim 48^\circ$ N) appeared to show evidence for remarkably periodic perturbations that indicated the almost regular passage of coherent waves past the spacecraft (see figure 18). From analyses of the phase relationships between $u$, $v$ and $p$, and the phase speed, Barnes (1981) deduced that the two main periodicities in this record corresponded to $m \approx 2$ and $m \approx 3\text{--}4$, indicating a tendency for the martian atmosphere to favour the equilibration towards low zonal wavenumber flows with amplitudes that may remain approximately constant for several tens of sols.

This tendency to favour low wavenumber, coherent, regular traveling wave structures is quite different to what is usually encountered in the Earth’s atmosphere, raising questions relating to:

(i) why does Mars typically favour the dominance of such low wavenumber baroclinic wave modes?
(ii) why are the equilibrated flows dominated by single wavenumber patterns?
(iii) is the resulting flow quasi-periodic or chaotic in nature?
(iv) is the martian mid-latitude circulation consistent with a low-dimensional attractor?

Leovy (1981a, 1985) suggested that, by analogy with rotating annulus laboratory experiments (Hide and Mason 1975), the greater atmospheric stability and smaller size of Mars might cause its transient baroclinic waves to be far more regular than those on Earth. Such regular, steady and stable traveling wave patterns are frequently observed in rotating annulus experiments under conditions for which the Rossby deformation radius is comparable in scale to the width of the annular channel, for wavenumbers smaller than a limiting value determined by the ratio \( b/a \) of the outer radius \( b \) to the inner radius \( a \) of the annulus itself (Hide and Mason 1975). Such semi-regular wave flows may be quasi-periodic or chaotic in behaviour, and seem to be consistent with the existence of an attractor of quite low dimension (Guckenheimer and Buzyna 1983, Read \textit{et al} 1992, 1998).

The mechanisms by which rotating annulus waves select which wavenumber to favour at fully nonlinear equilibration are still not fully understood, but seem likely to share some aspects in common with mechanisms identified in simple, so-called \textit{weakly nonlinear} spectrally truncated models of baroclinic instability. Such models (e.g. see Drazin 1970, Pedlosky 1970, 1971) represent only the leading order nonlinear interactions between a single mixed baroclinic–barotropic traveling wave and the background \((m = 0)\) zonal flow (i.e. suppressing quadratic and higher order wave–wave interactions). The nonlinear self-interaction of a growing, linearly unstable wave generates a correction to the \( m = 0 \) zonal flow (at second order in wave amplitude) that feeds back on the growth rate, eventually reducing it to zero (a steady wave state, for which the modal amplitude equations may asymptotically reduce to a set of Landau equations in the presence of some frictional damping) or with a more complicated, quasi-periodic or chaotic time dependence (e.g. see Lovegrove \textit{et al} 2001, 2002), for which the modal amplitude equations may reduce asymptotically to the classical real or complex Lorenz equations.

When more than one distinct wavenumber mode is able to grow from infinitesimal amplitude on a given zonal flow, weakly nonlinear models do not provide a unique answer as to what mechanism will act to select the dominant mode. However, one commonly found factor is for the flow to select the mode that is capable of releasing the most available potential energy (APE) from the initial flow (Hart 1981). In practice, this may correspond to the mode that can reach the largest barotropic amplitude (Hart 1981, Appleby 1988), provided nonlinear wave–wave interactions are absent. Where wave–wave interactions are permitted, the mode selection may become hysteretic, such that a non-optimal wave mode (i.e.
one that does not release the maximum possible APE) may persist as the dominant mode if it was previously dominant under more favourable conditions at an earlier time. In a situation where the forcing that maintains the background zonal state is varying cyclically with time over a range that crosses the boundary between two or more optimal modes, this can lead to complex and chaotic behaviour as the flow pattern flips erratically from one dominant mode to another (e.g. see Buzyna et al (1978), for examples in the laboratory setting).

The presence of wave–wave and higher order nonlinear interactions may also permit the possibility of secondary instabilities of the primary dominant wave mode, thereby preventing the sustained dominance of a single baroclinic wave mode. The prototypical idealized model for such a situation in the context of a planetary atmosphere considers the stability of the basic Rossby–Haurwitz (RH) mode on the sphere to wave-like barotropic perturbations (Baines 1976, Hoskins 1973, Lorenz 1972), although this has also been generalized to investigate baroclinic perturbations and instabilities of the basic RH wave (e.g. see Kim 1978, Grotjahn 1984a, 1984b). The basic criterion for barotropic stability of the RH wave can be interpreted in relation to Fjørtoft’s theorem for energy transfer in a quasi-geostrophic flow (Fjørtoft 1953), for which both energy and squared vorticity must be conserved in non-dissipative nonlinear interactions. This essentially requires that a given wave mode must lose energy simultaneously to both a higher and a lower wavenumber mode. Thus, the longest wavenumber modes capable of fitting into the domain tend to be relatively stable because of the unavailability of longer wavelength modes to which they can lose energy in an instability.

Such considerations seem to apply reasonably well to Mars, where the most regular wave modes reaching largest amplitude tend to favour relatively low wavenumber states with $m = 1$–$4$ (in contrast to the Earth, for which dominant wavenumbers are around $m = 4$–$8$; e.g. see Randall and Stanford (1985)). This would seem to be qualitatively consistent with the wave-number selection mechanisms highlighted by Hart (1981) and Appleby (1988), and with the results of Lorenz (1972) and subsequent work that suggested that Rossby–Haurwitz wave modes with $m < 5$ could be barotropically stable. But it is too much of an oversimplification to regard martian weather systems as purely quasi-periodic. Collins et al (1996) used an early version of the Oxford Mars GCM and Viking Lander data to show that baroclinic waves in Mars’s middle atmosphere actually appear to flip chaotically between two dominant transient wave modes, one with period near 2–4 sols and the other near 5–7 sols, at intervals of several tens of sols. Their model simulations included cases both with and without diurnal variations in forcing. Without the diurnal cycle, the waves were remarkably regular in behaviour, but would change wavenumber abruptly (and hysteretically) at various points in the seasonal cycle, as their background state slowly changed. With the diurnal cycle active, however, the flow would flip erratically back and forth between different wavenumber states, much as appeared in the Viking observations.

This would appear to indicate that the diurnal forcing actually stimulates chaotic mode flipping in a way that emulates the chaotic intermittency found in the vicinity of a noise-induced crisis in a low-dimensional dynamical system (Sommerer et al 1991a, 1991b), with the thermal tide playing at least part of the role of the perturbing noise. This is probably an incomplete analogy, since observations (e.g. Banfield et al 2004) also often suggest the simultaneous presence of moderate wavenumber ($m \gtrless 3$–$4$) shallow, surface-intensified baroclinic modes in addition to the low wavenumber, deep, mixed baroclinic–barotropic modes that exhibit this chaotic flipping. More recent simulations using a more realistic model with dust opacities derived from satellite observations (Mulholland 2012) show rather less qualitative difference between runs with and without an active diurnal cycle. Even with diurnally averaged forcing, the flow seems to adopt a more uniformly chaotic behaviour, with both modes present for much of the time, though there is still some evidence of interchanges of dominance between the two main modes, similar to what was seen by Collins et al (1996) with diurnal forcing. In the absence of perturbations to the baroclinic wave train due to the diurnal cycle, it seems that there are other perturbations present in these simulations, e.g. due to the more complex atmospheric dust field used and the presence of radiatively active water ice clouds. The anticorrelation between the amplitudes of the two most dominant modes appears stronger still when the diurnal cycle is activated, however, as shown in figure 19.

The analogy between this and erratic baroclinic mode-flipping in the martian atmosphere, even in the simplified configurations considered by Collins et al (1996), has yet to be investigated in detail (such as for its scaling behaviour (Sommerer et al 1991b)), but if true, would imply that this aspect of martian atmospheric dynamics is governed by a relatively low-dimensional attractor, rather than a turbulent one, in radical contrast to the Earth (Read et al 1998, Martínez-Alvarado et al 2009a).

### 3.3. Diurnal variations

To a large extent the martian atmosphere is dominated by the diurnal cycle of solar heating, especially near the surface and away from the winter polar night. As with a desert on Earth, there is a large surface temperature difference between day and night (on Mars as much as 100 K is possible and more than 50 K is commonly observed). This heating pattern, and the fast response time of the low-thermal inertia atmosphere and strong thermal forcing per unit mass (e.g. Conrath 1976, Zurek 1976, Zurek et al 1992), forces the development of atmospheric thermal tides. A theoretical linear partial differential equation framework for tides, primarily for application to the Earth’s oceans, was developed by Laplace in 1775 and the problem of ocean tides on Earth has been revisited by a variety of physicists since, including Lord Kelvin in the 1870s.

Thermal tides are atmospheric modes forced by solar heating, with periods that are harmonics of the solar day, and can equally well be described by classical theory (Chapman and Lindzen 1970, Lindzen 1970). They are a class of waves with periods equal to or less than one day, in contrast to planetary waves such as the baroclinic waves described in the previous subsection that have periods longer than one day, but the tidal modes have large spatial structures. The primary diurnal tide
is a wavenumber 1 disturbance around a latitude circle with a wavelength of about 35 km in the vertical, and the semi-diurnal (12 h period) mode is zonal wavenumber 2 and has a vertical wavelength of over 100 km (Zurek 1981, Zurek and Leovy 1981, Wilson and Hamilton 1996). A series of harmonics is generated, known as sun-synchronous tides if they move westwards with the Sun, and the situation is further complicated by interaction with the surface, either through varying thermal properties or topographic height (see the following subsection), which can result in modes that propagate eastwards, in the opposite direction to most tides rather than being phase-locked to the sub-solar point (Conrath 1976, Zurek 1976, Forbes et al 2002). One important example is the combination of the diurnal tide and zonal wavenumber 2 surface variations, which gives rise to an eastward-propagating wave number 1 with a period of one day. On Mars this mode primarily takes the form of a wavenumber 1 ‘Kelvin wave’, which may be resonantly excited (Zurek 1976, 1988), and which has been identified in spacecraft data (Wilson 2000, Wilson et al 2002, Banfield et al 2003).

As well as the thermal signature, tides redistribute mass and so produce a surface pressure signal. According to classical tidal theory (Chapman and Lindzen 1970), the temperature and pressure response is in quadrature with the local thermal forcing, in the absence of other dynamics. On Earth, this signal can just be detected in a surface pressure time series, although it is not readily apparent in the higher latitudes compared to typical variations resulting from high- and low-pressure weather systems. On Mars, the atmospheric tides have a broadly similar absolute amplitude, but against the background of a surface pressure that is typically two orders of magnitude lower than on Earth, tidal surface pressure signals can dominate the surface pressure and produce a repeatable diurnal cycle, varying with time of year and atmospheric dust loading (i.e. with thermal forcing by absorption of solar radiation).

There is now considerable observational and modelling evidence for atmospheric tides on Mars. Hess et al (1977) and Leovy (1981b) have analysed Viking Lander surface pressure observations to show the presence of tidal modes on Mars. Wilson and Hamilton (1996), Bridger and Murphy (1998) and Lewis and Read (2003) describe major tidal modes found in various Mars global circulation models (MGCMs). More recently, the thermal signature of tides in the atmosphere has been documented from Mars Global Surveyor (Banfield et al 2000, Guzewich et al 2014) and Mars Reconnaissance Orbiter observations (Lee et al 2009), supplemented by surface pressure measurements from the Curiosity Rover (Harri et al 2014). Lewis and Barker (2005) used data assimilation, combining MGS thermal and dust opacity measurements with an MGCM (see Lewis et al 2007 for details), to diagnose the sun-synchronous diurnal and semi-diurnal tides and the first Kelvin mode and to show their correlation with the atmospheric dust loading over three martian years.

For an atmosphere with low or moderate dust loadings, over most of the martian year, the largest response is in the diurnal mode. The semi-diurnal tide has about half the amplitude of the diurnal tide. Maxima in the Sun-synchronous tidal amplitudes occur just before northern winter solstice ($L_s = 270^\circ$) each year. At $L_s = 25^\circ$, Mars is closest to the Sun and receives most solar insolation, which, combined with the annual peak in dust levels, results in the strongest thermal forcing. The Kelvin mode has two maxima per year, with the largest occurring during the relatively clear and cold northern hemisphere summer, when the diurnal mode is small. This large response is most likely to be due to the first Kelvin mode frequency being close to atmospheric resonance under these conditions (Zurek 1988, Wilson and Hamilton 1996, Bridger and Murphy 1998).

Individual dust storms leave a signature in the tidal amplitudes. When the atmospheric dust loading grows, there is a...
clear correlation between the dust loading and the amplitude of the diurnal, Sun-synchronous tide, as might be expected, but the clearest linear relationship is with the semi-diurnal, Sun-synchronous tide. This is largely a result of the saturation of the diurnal tide at higher dust loadings. When the dust opacity increases to higher levels, such as during a global dust event, the solar forcing of the surface becomes less strong and heating is spread over a greater range of altitudes within the atmosphere as dust is mixed upwards. This can result in potentially destructive interference once significant heating extends over more than half a wavelength of the diurnal mode. The change in heating location can also alter the phase of the diurnal mode, with a phase retardation of about two hours at the surface under planet-encircling dust storm conditions (Lewis and Barker 2005).

In contrast, the large vertical wavelength of the semi-diurnal mode means that it effectively integrates the thermal response over the whole atmosphere and continues to become stronger with greater dust loading in a proportional manner (Lewis and Barker 2005). Thus the semi-diurnal Sun-synchronous tide is preferentially forced when the heating is distributed over a number of atmospheric scale heights (Leovy 1981a), as is observed on Mars (Zurek 1981). There is much less correlation between the Kelvin mode amplitude and the dust loading. In this case, higher dust loading may enhance the strength of the tidal forcing, but, under dustier conditions with increased atmospheric static stability, the period of the Kelvin mode is increased (Richardson and Wilson 2002b), though this phenomenon has not yet been quantified. In the present climate, the combination of the higher topography and stronger solar heating around perihelion in southern summer, when Mars is closest to the Sun in its orbit, means that the largest dust storms are generated almost exclusively in the southern hemisphere.

The role of topography in modulating and exciting different thermal tidal modes has been mentioned previously, but topography equally affects the more slowly varying planetary waves. Variations in topography cause the excitation of planetary-scale stationary wave patterns, which affect CO2 condensation in the polar regions (Colaprete et al. 2005).

Topographic features also alter the circulation at much smaller scales: slope winds, buoyancy-driven circulations caused by diurnal variation in solar heating on sloping terrain, which are stronger on Mars than on Earth owing to the lower atmospheric thermal inertia and so faster atmospheric response to solar heating of the surface (Savijärvi and Siilin 1993), are an important source of flows strong enough to lift dust from the surface (Rafkin et al. 2002, Spiga and Lewis 2010). During the day, anabatic winds blow up slopes, lifting dust from cliff edges and forming plumes above the largest volcanoes. At night, cold katabatic winds drain downslope and warm as they descend to higher pressures. This warming has a noticeable impact on day–night surface temperature contrasts in a ring around Olympus Mons (Spiga and Lewis 2010, Spiga et al. 2011). The process of downslope winds warming the surface is analogous to similar winds on Earth, although these are generally formed by deflection of larger scale winds over mountain ranges, such as the Chinook (or ‘snow-eater’) winds on the eastern side of the Rocky Mountains in North America. Diurnal katabatic winds on Earth tend to be on much smaller scales (insufficient to produce noticeable warming, except perhaps in Antarctica) and are often felt as cold, drainage flows in valleys at night.
4. The mineral dust cycle

Airborne dust is ubiquitous on Mars and plays a key role in affecting many aspects of its climate. This is in contrast to the Earth, where airborne dust and aerosols play a relatively minor role in terrestrial meteorology and the atmospheric radiative energy balance except in particular localized regions, such as in deserts and following large, explosive volcanic eruptions. However, variations, e.g., in anthropogenic sulphate aerosols are believed to play a significant role in affecting the global albedo of the planet, partially off-setting warming by increasing greenhouse gases (Denman et al 2007). Because the martian surface environment is essentially a global desert, covered in a variable mixture of dry boulders, sand and dust, meteorological phenomena can lift and deposit such friable surface material relatively easily, leading to the possibility of major storm events that raise dust over a large area, and even long-term systematic movements of such material across large distances. Such long-term movements are referred to as the martian dust transport cycle and are considered here in the context of the processes and meteorological phenomena that lift dust particles into the atmosphere, transport them across the planet and subsequently deposit them back to the surface.

As discussed above, the presence of mineral dust particles suspended in the atmosphere provides a substantial contribution to the opacity of the atmosphere in the visible and thermal infrared. Even under relatively clear conditions the optical depth of dust layers in the atmosphere at visible wavelengths, $\tau_v$, is typically around 0.1–0.2, and can rise to O(2–5) in the midst of a substantial regional or planet-encircling dust storm. Such opacities significantly attenuate the sunlight incident upon the ground, whilst also leading to significant local heating of the dust itself and subsequently the air in contact with the dust. This local heating has a substantial impact on the static stability above and below the suspended dust layer, and can even lead to nonlinear feedback effects through radiatively induced changes in the local winds affecting the raising of dust. Such effects, coupled with the highly active nature of martian meteorology, lead to substantial variability in atmospheric dust loading on a variety of space and time scales.

4.1. Phenomenology of martian atmospheric dust

Airborne dust (on Mars and the Earth) consists mainly of small particles of silicate minerals with a typical particle size of $\lesssim$1–2 $\mu$m. This size category is selected because atmospheric motions are sufficient to keep the particles suspended for long periods before sedimentation removes them under gravity. For particle sizes $\gtrsim$10 $\mu$m or so sedimentation rates are so fast that dust is unable to be transported very far before falling out to the surface. As we discuss below, fine-scale dust is commonly lofted into the air by a variety of processes that may be affected by the precise meteorological conditions close to the surface. Because of the nature of the martian surface, dust is available to be lifted by atmospheric motions close to the surface almost everywhere except where the surface is covered by ice. As a result, the dust loading at any one location may be highly variable. Some typical recent measurements are illustrated in figure 20, which shows a time series of near-infrared optical depth obtained at various locations from orbit by the OMEGA instrument on ESA’s Mars Express mission (Vincendon et al 2009), together with a local series of measurements from NASA’s Mars Exploration Rovers Spirit and Opportunity (Lemmon et al 2004, 2015).
This shows the typical range of variability in dust optical depth, from minimum values of around 0.2–0.4 during relatively quiescent periods in northern hemisphere spring and summer ($L_s \sim 45^\circ$–$130^\circ$) to values $\geq 1$ in sharp peaks that correspond to the passage of individual dust storm events. Most of these events occur during the period corresponding to southern hemisphere summer ($L_s \sim 180^\circ$–$300^\circ$) when the planet is close to periapsis and solar heating is at its strongest. However, they occur more or less at random during this period. Most of them are relatively small, regional events leading to short-lived peaks in $\tau$ of no more than around $\tau = 2$. In the second Mars year illustrated in figure 20 (Mars Year 28), however, there is a much larger event that began at $L_s \sim 265^\circ$ that produced a peak of around $\tau = 4$ and persisted for around 40$^\circ$ in $L_s$ (corresponding to around 80 d duration). This is an example of a major ‘global’ dust storm, more accurately referred to as a ‘planet-encircling’ storm, in which dust was persistently raised over a large area, feeding material into the large-scale circulation which transported the dust cloud over large tracts of the planet. The climatological incidence of significant dust storms has now been thoroughly documented from spacecraft observations over several Mars years (Montabone et al 2015, Wang and Richardson 2015), presenting some important challenges to understanding these processes.

Dust storms of any significant size are observed to be relatively rare during northern summer, raising the question of what events might sustain the low level of background dust suspended during this time. The answer to this is still somewhat uncertain, but both observations (Balme and Greeley 2006) and models (Newman et al 2002a, Toigo et al 2003, Basu et al 2004) suggest that ensembles of small-scale convective vortices (‘dust devils’) occur frequently over large areas of the summer hemisphere at this time. Such vortices seem capable of lifting a substantial column of suspended dust to the top of the convective boundary layer (at altitudes of 5–9 km above the surface (Petrosyan et al 2011)), where the dust is detrained and transported in diluted form across wide areas. Figure 21 shows imaging of some small dust devil vortices observed close to the MER Spirit rover. Such features have much in common with dust devil vortices frequently observed in terrestrial desert regions under convective conditions. On Mars, observations and models (Kahre et al 2006, Whelley and Greeley 2008) suggest that the dust lifting rates within these features may be enough to account for much of the residual dust opacity in the atmosphere during seasons when major dust storms are rare. However, models such as that of Kahre et al (2006) tend to underpredict the occurrence of cap-edge dust storms in the southern hemisphere during southern autumn and winter ($L_s = 0$–$180^\circ$), so the quantitative role played by dust devils in providing much of the observed dust loading outside the main dust storm season still remains somewhat uncertain.

At least some of the processes that can organize major dust lifting events appear to be related to particular classes of meteorological phenomena. During winter seasons in each hemisphere, for example, meteorological variability is dominated at mid-high latitudes by large-scale cyclonic weather systems resulting from baroclinic instability of the main circumpolar vortex. The growth and equilibration of such disturbances is observed to lead to the formation of intense weather fronts which, as on the Earth, are linear features within which horizontal thermal gradients are concentrated. Such regions are also often associated with enhanced and gusty near-surface winds, which evidently lead to the initiation of regional dust storms that entrain dust into the synoptic-scale circulation. Such baroclinically generated regional storms, however, seldom seem to develop into planet-encircling events. The larger, planet-encircling storms appear to have a more complex life cycle with origins that are often difficult to unravel from observations alone. Large-scale meteorological events undoubtedly play a role in their initiation, but apparently involve some complex interactions between different scales.

Once dust has been lifted by a series of dust storms it may be entrained into the global scale atmospheric circulation and transported across the planet as a quasi-conserved tracer. This seems to be largely consistent with the observed horizontal spreading of dust clouds lifted by discrete events (e.g. Martin 1974, Strausberg et al 2005) but recent satellite observations from MCS have revealed that the vertical distribution of suspended dust is sometimes concentrated into an elevated layer, centered at an altitude of up to 25 km. Figure 22 shows a recent example of a zonally averaged cross-section of dust mixing ratio inferred from MCS limb profiles (Heavens et al 2011) that clearly shows such an elevated layer over the martian tropics close to both solstices. How such a distribution is produced and maintained is not currently well understood.

4.2. How does dust enter the atmosphere?

4.2.1. Near surface wind stress. From our experience of being caught in a strong gust of wind, e.g. on a dry, sandy beach or dusty yard, there might seem to be no great difficulty
in lifting dust into the air. Strong winds on Earth seem to be able to lift clouds of dust easily into the air around us. However, the lifting of dust into the free atmosphere, beyond the lowest few cm–m of the boundary layer, is far from straightforward. This is because most of the dust particles which are caught in strong near-surface gusts are relatively large in size (≥20 μm) and, although they may blow around at low levels in local eddies for a while, they quite quickly fall to the ground as a result of their weight. For dust to access higher altitudes requires the emission of smaller, micron-sized particles, which require much stronger winds to mobilise directly.

The theory behind the lofting of dust particles up from the surface due to the action of near-surface winds has been developed over the past seventy years, starting with the work of Ralph Bagnold (Bagnold 1954), with the primary goal of understanding and accurately modelling dust emission on Earth. A combination of experimental measurements of dust flux, from both field campaigns and wind-tunnel experiments (e.g. Gillette 1978), and theoretical analysis of the physics involved in the process (e.g. Kok et al. 2012, Marticorena and Bergametti 1995) has contributed to the formulation of semi-empirical expressions for the calculation of emitted dust flux, as a function of local surface wind velocity and a variety of other quantities.

Based on his observations of wind-blown sand and dust in the north African desert during World War II, Bagnold (1954) suggested that the resistance of surface particles to being lifted by the wind was a threshold-dependent process which could be represented quantitatively by defining two threshold wind speeds: the fluid threshold and the impact threshold. The fluid threshold is the speed at which wind stress alone enables dust particles to be lifted directly from the surface. The impact threshold is then the minimum wind stress component of the total surface stress (including particle impacts) which leads to the lifting of small particles in the presence of impacting larger ones. The process is schematically illustrated in figure 23.

The fluid threshold is generally represented by defining a threshold drag velocity, \( u_\theta \), which must be exceeded by the actual drag velocity \( u_{\theta} \) for lifting to occur. \( u_\theta \) is related to the near-surface wind stress \( \zeta \) and atmospheric density \( \rho \) by

\[
 u_\theta = \frac{\zeta}{\sqrt{\rho}}, \tag{8}
\]

and may be determined from near-surface wind velocities in the surface boundary layer (e.g. Garratt 1994). Air velocities are generally very close to zero in a very thin layer next to the ground surface, and just above this layer, there is a sub-layer within which velocities vary approximately logarithmically with height,

\[
 u(z) = \frac{u_\theta}{k} \ln \left( \frac{z}{z_0} \right), \tag{9}
\]

where \( k \) is von Karman’s constant (≈0.4) and \( z_0 \) is the so-called roughness length. Hence

\[
 u_\theta = \frac{ku(z)}{\ln(z/z_0)}. \tag{10}
\]

The dependence of the threshold velocity \( u_\theta \) has been determined semi-empirically from a range of experiments carried out in the laboratory, covering conditions applicable to both the Earth and other planets including Mars (Bagnold 1954, Greeley and Iversen 1985), and found to take the form

\[
 u_\theta \approx A \sqrt{\frac{D_p \rho_d - \rho}{\rho}}, \tag{11}
\]

where \( \rho_d \) is the density of the material comprising the dust particles, and \( D_p \) is their diameter. \( u_\theta \) thus becomes large for large \( D_p \). However, some important details are concealed by the factor \( A \), which in turn depends on factors including interparticle cohesion, particle size \( D_p \) and the friction Reynolds number at threshold \( Re_\theta = u_\theta D_p \nu \), where \( \nu \) is the kinematic viscosity of the air. As particle size becomes small, the threshold becomes dominated by an increasing interparticle cohesion, which on Earth includes an important contribution from surface moisture, but on Mars arises primarily from other factors such as
electrostatic forces, intermolecular (van der Waals) forces (Castellanos 2005, Kok et al 2012) and even weak magnetic interactions (Madsen 1999).

The critical measure of these interactions is the cohesive granular Bond number, \( B_{og} \), defined by

\[
B_{og} = \frac{F_a}{mg},
\]

where \( F_a \) is the attractive interparticle force and \( mg \) is the particle weight. Particulate soils are cohesive if \( B_{og} \gg 1 \) and tend to clump together, whereas particles for which \( B_{og} \ll 1 \) will tend to ‘flow’ easily as particles can move readily over each other (compare trying to pour rice (non-cohesive) with pouring finely divided flour (a cohesive powder), for example). Because of the way in which particle weight and interparticle forces scale with particle diameter (\( \sim mg \sim D_p^3 \) whereas \( \sim F \sim D_p \) for spherical particles with van der Waals forces, smaller particles tend to be more cohesive than larger ones (Smalley 1970). These factors generally make it more difficult for the smaller particles (\( \sim 1 \mu m \)), which would be sufficiently small to remain in suspension once in the free atmosphere, to be lifted by surface winds than for larger ones.

The balance between particle cohesion and weight means that thresholds become large for both large and small \( D_p \). Numerical solutions of threshold equations similar to (11) (e.g. see Greeley and Iversen (1985) and Shao and Lu (2000)), with \( \rho_f = 2.7 \times 10^3 \text{ kg m}^{-3} \), predict an ‘optimum’ particle size for direct lifting by the wind, of \( D_p \sim 180 \mu m \) (shown in figure 24), requiring a wind at 5 m altitude of around 30–40 m s\(^{-1}\). For much smaller particles, however, closer to the size typically found in suspension in the free atmosphere on Mars (\( D_p \approx 2 \mu m \)), the threshold velocity is found to be extremely high, requiring enormous 5 m wind speeds which would virtually never occur. Therefore, a combination of saltation and sandblasting are generally assumed to be responsible for the lifting of very small dust particles both on Earth and Mars.

Bagnold (1954) and White (1979) found that the upward flux (\( F \)) of small particles lifted once the fluid threshold for large particles had been exceeded could be related to the horizontal flux of large (sand) particles, \( H \) (itself dependent on drag velocity \( u_0 \) and surface roughness \( z_0 \)), in the form

\[
F(u_0, z_0) = \alpha_N(u_0)E(z_0)H(u_0, z_0),
\]

where \( E \) is the fraction of the surface that is erodible (i.e. capable of emitting dust) and \( \alpha_N \) is a lifting efficiency parameter. The variables \( F, H \) and \( \alpha_N \) will also depend on the size distribution of the particles present on the surface, although on Mars this is often neglected in practice on the grounds that most lifting will take place for particle sizes close to that with the lowest \( u_0^{\ast} \); and since little information is available on surface particle size distributions.
Equation (13) offers the basis for parameterising the lifting of dust in numerical simulations of Mars’s atmospheric circulation, given formulae for $H$ and $E$. Both Newman et al (2002a) and Basu et al (2004), for example, have utilized a form of equation (13) with $E$ set to 1 (in the absence of information on spatial variations of $z_0$ and soil composition) and with $H$ given by an empirical formulation due to White (1979) and Seguro and Lambert (2000),

$$H(u_\ast) = \int_{u_\ast}^{\infty} H^*(u_\ast)w(u_\ast)du_\ast, \quad (14)$$

$$H^*(u_\ast) = 2.61 \frac{\rho}{g} (u_\ast)^3 \left(1 - \frac{u_\ast^3}{u_\ast^3}ight)^2 \left(1 + \frac{u_\ast^3}{u_\ast^3}ight)^2, \quad (15)$$

where $w(u_\ast)$ is the Weibull distribution, representing the statistical distribution of wind gusts (Seguro and Lambert 2000), given by

$$w(u) = \left(\frac{\kappa}{c} \right)^\kappa \left(\frac{u}{c}\right)^{\kappa-1} \exp\left[-\left(\frac{u}{c}\right)^\kappa\right] \quad (16)$$

where $u$ is the wind speed, $c$ is a speed scale and $\kappa$ is a dimensionless shape parameter (such that small values of $\kappa$ represent long-tailed distributions implying relatively gusty conditions). The latter is included to take some account of the variability of surface winds, allowing for the possibility of occasional gusts strong enough to lift dust via saltation, even if the mean wind strength is below the lifting threshold. Fenton and Michaels (2010) found a shape parameter of $\kappa \approx 2.5-3$ to adequately represent sub-hourly wind variability on Mars, but although both Newman et al (2002a) and Basu et al (2004) have previously considered the use of probability distributions, modellers have generally used a simpler, deterministic approach, giving a well-defined threshold and producing spatio-temporally intermittent lifting similar to that observed on Mars, in a broad sense. With such a threshold effect, equations (14) and (15) reduce to

$$H(u_\ast) = \max \left[0, 2.61 \frac{\rho}{g} (u_\ast)^3 \left(1 - \frac{u_\ast^3}{u_\ast^3}ight)^2 \left(1 + \frac{u_\ast^3}{u_\ast^3}\right)^2\right], \quad (17)$$

which implies essentially no lifting for surface friction velocities below $u_\ast^\text{th}$ and a saltation flux that increases roughly as $(u_\ast)^3$ once the threshold is exceeded. Such a dependence is at least qualitatively consistent with an almost explosive increase in dust lifting as the threshold is exceeded, which may explain the rapid development of dust storms in windy conditions.

More recent work has sought to take into account the difference in the effect of surface winds on stationary particles compared with their effect on particles in active saltation. The two relevant thresholds, introduced earlier, fluid threshold $u_\ast^a$ and impact threshold $u_\ast^b$, are fairly similar in magnitude on Earth ($u_\ast^b$ is around 80% of $u_\ast^a$ (Kok 2010b)). For Mars, however, some recent theoretical studies (Claudin and Andreotti 2006, Almeida et al 2008, Kok 2010a, 2010b) have suggested that these two thresholds may be much more different, with values of ratio $u_\ast^a/u_\ast^b$ ranging from 0.1 to $\approx 0.5$. This is a consequence of much longer trajectories of saltating sand particles on Mars, which form due to the combined effects of lower surface gravity than on Earth and the reduced effects of wind resistance (Kok 2010a, 2010b). This relatively large ratio suggests the possibility of significant hysteresis in threshold-limited dust lifting, in that a relatively large threshold wind may be necessary to initiate the first lifting events but a significantly lower wind speed may be sufficient to sustain saltation and lifting, suggesting that, once mobile, sand can be moved around relatively easily along the martian surface. Verification of this effect has yet to be demonstrated experimentally, however, because it would require activating saltating flows over much longer distances than are practical in suitable wind tunnel facilities. But observational indications from recent HiRISE images (Bridges et al 2012) of relatively mobile sand dune fields, at least in some regions of Mars, suggest that these effects are real and that large saltating dust fluxes, comparable to those found in dust storms on Earth, are fairly common on Mars despite the much thinner atmosphere.

Some authors have suggested that saltation transport on Mars may be further modified by other factors such as electrostatic effects (Kok and Renno 2008), where the very low humidity allows for the development of significant tribological electric fields, and interparticle collisions (e.g. Huang et al 2007, Kok et al 2012), potentially increasing the saltating mass flux (Zheng et al 2006, Kok and Renno 2008, Rasmussen et al 2009) and affecting the height of the saltating layer (Kok and Renno 2008). Electrostatic effects may also enable sedimenting fine dust particles to clump together to form loose aggregates, as apparently seen in some locations (Sullivan et al 2008) by the Mars Exploration Rovers (see figures 25(a) and (b)). Such aggregates may be readily dispersed during saltation (if not lifted directly by the wind as a result of their lower density), facilitating the release of finely divided dusty material into the air (Sullivan et al 2008). These aspects are still relatively poorly understood, however, so any firm conclusion as to the quantitative roles of electrostatic effects and interparticle collisions on saltating flows on Mars must await further research.

4.2.2. Dust devils. The other main phenomenon that seems likely to contribute to maintaining the atmospheric dust loading on Mars is that of the so-called ‘dust devil’ (see Balme and Greeley 2006) for a recent and comprehensive review). Dust devils are convective vortices that form under conditions where the ground is heated strongly by sunlight, forcing the atmospheric structure towards a super-adiabatic lapse rate. Buoyant convective plumes may form into vertically aligned vortical structures that are observed on Mars to range in diameter from a few metres up to more than 100 m. Their vertical extent is related to the height of the convective boundary layer, which can extend up to several km on Mars, owing to the strength of solar heating and the relatively low air density.

The source of vorticity for these features (on both Earth and Mars) is still somewhat controversial. Evidence on Earth seems to indicate no particular preference for either sign of vertical vorticity (Balme and Greeley 2006), suggesting that planetary rotation plays only a minor role. This is consistent
with the very large Rossby numbers found in such vortices, implying that Coriolis accelerations are small compared with local shears. Topographic features may influence the sense of rotation in some cases, but many dust devil vortices form in flat regions with weak ambient winds, showing that topographic sources are not essential. So it is more likely that the main source of vorticity in many cases is due to the tilting of horizontal vorticity within the frictional boundary layer by convection.

Figure 26 shows a schematic view of the likely circulation inside a typical dust devil. Local heating of the surface renders a patch of air positively buoyant. As it rises, surrounding air is drawn in along the ground towards the centre, tending to spin up as it approximately conserves angular momentum. But subsequent trajectories of the air are then affected by secondary circulations that lift and spread air vertically and radially. Stagnation points may occur close to the ground and sometimes at higher altitudes, leading to parts of the vortex core being associated with downward flow, well above the ground.

This form of circulation appears spontaneously in numerical simulations using 3D LES models, under both terrestrial (e.g. Gu et al. 2006) and martian conditions. Figure 27 shows an example of a simulated vortex on Mars from Toigo et al. (2003). Strong inward flow in the surface layer is implied by the negative pressure anomaly at the vortex core, while the strongest upwelling velocities are seen in the walls of the vortex, in agreement with the profiles marked in figure 26. The vertical and horizontal scales of this feature are much larger than terrestrial examples of the phenomenon, which are generally of a few tens of metres or less in size.

If the vortex is large enough, dust particles may be lifted, through saltation or direct detachment by strong tangential winds, possibly assisted by an additional lift force applied to surface particles by the low pressure anomaly (Greeley et al. 2003). Particles are then carried by the winds and tend to concentrate into the upwelling conical sheet at the periphery of the vortex core. The result of this can be seen in figure 21, including, on the right of the image, a possible example of an ‘ejecta skirt’, formed by the larger of the lifted particles, which sediment too rapidly to be maintained in suspension higher up in the dust devil walls.

### 4.2.3. Roughness, erodibility and interannual variability

Another factor that may often affect the amount of dust lifted by near-surface winds is the availability of erodible material on the surface (essentially determining $E$ in equation (13)). Ruff and Christensen (2002) used thermal emission spectrometer (TES) albedo data from MGS to derive a map of the distribution of surface dust particles (smaller in size than around 100 $\mu$m). This showed the southern hemisphere of Mars to be more sparsely covered by dust grains than the northern plains. Episodic changes in surface albedo, notably in the wake of major dust storms, were detected by Szwast et al. (2006), indicating the redistribution of significant fractions of this surface dust by lifting events (changing the surface dust covering by several microns). These authors further noted that dust lifting in the Hellas region during the 2001 planet-encircling dust storm (where the storm appeared to originate) seemed to cease after around 20 d, even though the storm was still developing rapidly. Local surface winds were deduced by Montabone et al. (2005) to have remained strong at this location, suggesting that the Hellas region had become exhausted of available surface dust, preventing further lifting. The possibility that the occurrence of a major regional dust storm could deplete a local region of liftable dust was first considered by Newman et al. (2002b) in an attempt to explain why model simulations of spontaneously generated dust storms in a global circulation model were unable to reproduce the observed interannual variability of such storms. Some influence was noted, although this led to dust storms permanently dying out in some regions as dust reserves became
exhausted and were not spontaneously replenished; an effect that was largely confirmed in further investigations by Kahre et al (2005) using the NASA Ames Mars GCM. These results may point to large-scale biases in model dust transports, rather than suggest the presence of deeper reservoirs of surface dust at storm initiation sites.

The surface aerodynamic roughness length, $z_0$, a macroscopic quantity describing the distribution of non-erodible elements (stones and boulders) on the martian surface, is also of relevance to dust lifting. Recently mapped by Hébrard et al (2012), $z_0$ controls the efficiency with which near-surface momentum is converted into surface drag, through equation (10), and modifies the threshold drag velocity, due to the impinging of a fraction of the surface drag on the non-erodible elements rather than the dust itself. Variations in roughness length are routinely accounted for by terrestrial dust emission models (e.g. Laurent et al 2008), but their impact on martian dust lifting is yet to be investigated.

A combination of these two effects may play an important role in the interannual variability of major dust storms. The widespread presence of non-erodible elements on the surface makes it possible for some dust to be temporarily sheltered from the prevailing winds during a given storm (leading to the formation of so-called ‘desert pavements’), only to become available for another event at a later time. It has been observed (Greeley and Iversen 1985) that the rate of dust removal from an area containing a mixture of dust/sand and non-erodible features tends to decrease over time as the non-erodibles are uncovered and begin to shelter the remaining erodible material. This has been suggested (Pankine and Ingersoll 2002) to have the effect of increasing the effective threshold for saltation and lifting following a major storm, making it more difficult to lift dust subsequently. At later times, however, other dust-moving processes (such as dust devils) may release some of the dust previously hidden behind non-erodible elements, reducing the effective lifting threshold on timescales of months or years, and reactivating the region for lifting by later storms. Simple, low-order models have even suggested (Pankine and Ingersoll 2004) that this slow variation in effective saltation threshold can act as a form of self-organized criticality, holding large areas close to the critical threshold for dust lifting on interannual timescales and leading spontaneously to the chaotic occurrence of major storms. Attempts to implement a parameterization of this kind of variable threshold into more complex atmospheric models are the subject of current research, though preliminary results (Wilson and Kahre 2009, Wilson 2011, Mulholland et al 2013) are promising in terms of affecting simulated interannual variability of major storms. But this is clearly an area where more research is needed, in global modelling and in both field measurements (ideally on Mars, but also on terrestrial analogues of the martian environment) and laboratory measurements.

4.3. Convection and entrainment of lifted dust

Once lifted into the lowest region of the atmospheric boundary layer, turbulent eddies and other, more coherent air movements might then be expected to mix the dust upwards into the free atmosphere. Convective motions, enhanced by local heating due to absorption of solar radiation by the entrained dust, are often evident as cumulus cloud-like features at the tops of regional dust storms (e.g. Guzewich et al 2015, Malin et al

Figure 26. Sketch showing schematically the possible flow inside a dust devil, from Balme and Greeley (2006). Solid arrowed lines show the radial and vertical flow directions, while dotted lines represent the vertical wind speed profile that would be measured at that height. Stagnation points are indicated that could account for why downward flow is sometimes found in some dust devils, though not near the ground.
On larger horizontal scales (10–100 km), dust clouds are occasionally seen to organize into coherent spiral patterns (Wang and Ingersoll 2002), though are more commonly seen to take a more irregular and amorphous form. The spiral dust clouds are reminiscent of an early suggestion (Gierasch and Goody 1973) that dust heating by solar radiation would act like the release of latent heat in moist convection on Earth and enhance the formation of ‘hurricane-like’ vortical circulations. In fact an analog of total convective available potential energy (TCAPE) can be defined, even for this type of dry convection (Rennó and Ingersoll 1996, Holton 2004, Spiga et al 2013).

\[
C = \int_{\text{storm}} g \frac{\Delta T}{T_{\text{env}}} dz,
\]

where \(\Delta T\) is the temperature change experienced by individual air parcels via radiative heating within the dust storm, \(T_{\text{env}}\) is the mean environmental temperature and \(g\) is the acceleration due to gravity. An upper limit on the likely updraft velocities in such convective events can then be estimated as \(w \sim \sqrt{2C}\).

Given the rapid heating rates that may be possible for dust-laden air in summertime conditions on Mars (\(\sim 10\) K h\(^{-1}\)), and the relatively deep convective boundary layer (\(\sim 5\) km may be typical), this suggests the possibility of updrafts as strong as a few tens of m s\(^{-1}\). Such large values cannot be sustained over large areas but suggest the possibility of strong, localized updrafts that could entrain dust and lift it high into the atmosphere. Such jet-like injections of dust-laden air into the middle atmosphere constitute one possible explanation for the development of elevated layers of dusty air that have been observed in the martian atmosphere (e.g. see figure 22 and Heavens et al 2011). Once the upward jet of dust-laden air reaches its level of neutral buoyancy, e.g. upon reaching a thermal inversion at higher altitude, it is then likely to begin to detrain and spread horizontally, forming a detached dusty layer.

Such behaviour has been found in some model simulations, though coarse resolution global models seem to be reluctant to exhibit this form of dust transport except in unusual circumstances. Figure 28 shows one early example by Newman et al (2002b), in which a regional-scale dust lifting event was simulated over the Hellas basin in the martian southern hemisphere during southern summer. In this case, the dust tracer was made radiatively active, enhancing the local heating of dust-laden air during the daytime. This evidently led to enhanced lifting into a narrow plume near the equator that injected dust up to altitudes of 30–40 km, from which it then spread horizontally in latitude (and in longitude, though this was not shown in the figure) to produce a planet-encircling, elevated dust layer that subsequently gradually settled out due to sedimentation.

Such events almost certainly occur on smaller scales too, though this cannot be resolved in global models that are run at ‘climate’ resolution. But mesoscale numerical models are now being applied to Mars which allow much higher resolution simulations to be carried out, albeit over limited areas of the planet. Figure 29 shows a recent example of such a simulation of a ‘rocket-like’ dust storm event occurring on a regional scale (Spiga et al 2013). This also shows a column-like vertical jet of dust-laden air rising up out of the boundary layer to then detrain into a flat patch of dust that subsequently drifts horizontally downwind. This simulated event was run to emulate a dust-lifting event observed by the OMEGA instrument on Mars Express, where a patch of dust was also lifted to altitudes of 30–40 km over a period of 6–12 h before drifting downwind.

4.4. Removal of dust from the atmosphere

The ubiquitous presence of dust in the martian atmosphere is the result of a dynamic equilibrium between various processes lifting and injecting dust into the air and other processes that tend to remove it. The density of mineral dust particles is generally much larger than that of the atmosphere in which they are suspended, so the dominant process tending to remove dust is simple sedimentation under gravity.
4.4.1. Dry deposition. Where this takes place without the interaction with condensable species, it is known as dry deposition. For small particles (for which the sedimentation Reynolds number $\nu \approx 35$; e.g. Rossow (1978)), the rate of sedimentation is commonly estimated from the Stokes terminal velocity with the Cunningham slip correction (which applies when the Knudsen number, $Kn = \lambda / r_p \gg 1$, where $\lambda$ is the mean free path and $r_p$ the particle radius).

Figure 30 shows the times $\tau_{\text{sedim}}$ taken by particles of different radii to fall a distance of 1 km at various heights in the atmosphere, as calculated from equation (19). For particles of density 2500 kg m$^{-3}$, the fall times can be approximated by $\tau_{\text{sedim}} \approx 490 H r_p^2 / \eta$, where $H$ is the fall height (in km), $r_p$ is the particle radius (in $\mu$m) and $\eta$ is the atmospheric density. As can be seen, particles fall more quickly higher in the atmosphere, and submicron particles can remain in the atmosphere on timescales of years, as their fall times rapidly increase as they reach the lowest few kilometres.

An apparent decrease in fluid viscosity results at Knudsen numbers $Kn > 0.1$, i.e. for particle sizes $r_p < 10 \lambda$. On Earth at STP, where $\lambda \approx 0.1 \mu$m, this is the case for all particles smaller than around 1 $\mu$m. Under martian conditions, where the pressure is much lower and $\lambda \approx 30 \mu$m, this is true for particle sizes $<300 \mu$m. Thus, the fall speed $w_{\text{sed}}$ of a particle of radius $r_p$ and density $\rho_p$ is calculated via

$$ w_{\text{sed}}(r_p) = \frac{2 \rho_p r_p^2}{9 \eta} (1 + \alpha Kn), \quad (19) $$

$$ \alpha = 1.246 + 0.42 e^{-0.87/\lambda}, \quad (20) $$

where $\eta \approx 10^{-5}$ N s m$^{-2}$ is the dynamic viscosity of the atmosphere, and the $(1 + \alpha Kn)$ term is the Cunningham slip correction factor.

In addition, at the lowest pressures and for the sub-micron particle sizes considered, the Cunningham slip correction factor itself no longer applies and strictly Brownian motion would need to be taken into account.

Even for 1 $\mu$m particles with sedimentation speeds of order mm s$^{-1}$, model simulations suggest that typical updraft speeds averaged across large horizontal scales are not sufficient to keep particles suspended indefinitely (Newman et al 2002a). But small-scale turbulent diffusion seems to be both necessary and sufficient to mix weakly suspended small particles throughout the lower atmosphere on Mars, allowing small amounts of micron-sized dust particles to remain suspended indefinitely. This is what is commonly observed more or less all the time, even in the absence of significant dust storm activity.

4.4.2. Wet deposition and ‘scavenging’. The removal of aerosol or dust particles from the atmosphere through their
interaction with clouds and condensibles is known as scavenging (e.g. Pruppacher and Klett 1978), and the subsequent arrival of the particles at the surface is referred to terrestrially as wet deposition. Dust/aerosol scavenging has received considerable recent attention in terrestrial climate research due to its contribution to total aerosol deposition rates (e.g. Sportisse 2007). It can occur in two ways: in-cloud scavenging, as the result of heterogeneous cloud nucleation (discussed

**Figure 29.** Mesoscale model simulation of a regional ‘rocket-like’ dust storm, adapted from Spiga et al (2013) figure 4: longitude–altitude sections of density-scaled dust optical depth at 21.6 μm in 10^{-3} m^2 kg^{-1}. Sections are obtained at latitude 2.5° S, the season is late northern summer (L = 135°), and diagnostics are shown every two hours from local time 14:00–00:00.
in more detail in the next section), and below-cloud scavenging, which is the removal from the atmosphere of aerosol particles encountered by precipitating liquid drops or ice crystals. On Mars, the situation is simplified somewhat compared to Earth, as only ice cloud processes need be considered; however, these are less well understood than the liquid water equivalents, perhaps as they are generally of secondary importance on Earth. Martian dust scavenging can be further divided into removal respectively by H₂O ice and by CO₂ ice, both of which are potentially significant at different times and locations.

In-cloud scavenging by ice clouds can be described simply by a scavenging efficiency $R_s$, which is the fraction of available dust (or aerosol) particles that act as ice nuclei (IN). The mass of ‘interstitial’ (not cloud-borne) dust decreases through scavenging at a rate proportional to both $R_s$ and the rate of ice cloud nucleation. $R_s$ is expected to depend on the dust particle size, but values for the scavenging efficiency are not well known even on Earth (Stier et al 2005). Nevertheless, in-cloud scavenging by CO₂ ice seems likely to be a small effect even at winter high latitudes. Scavenging during H₂O ice cloud formation could, depending on $R_s$, significantly affect the vertical distribution of atmospheric dust, particularly at low latitudes during northern spring and summer, though direct evidence for this is limited at present. Based on observations of cloud and dust aerosols by the Spectroscopy for Investigation of Characteristics of the Atmosphere of Mars instrument (SPICAM), and earlier modelling results (Michelangeli et al 1993), Montmessin et al (2006) suggested that the process would effectively cap the upward extent of dust upwelling from the lower atmosphere, limiting the height to which dust would penetrate. Such a limitation is plausible and consistent with the occasional observation of water ice clouds visibly capping a dusty layer below. On Earth, in-cloud scavenging often results in aerosol particles being rapidly ‘washed out’ of the atmosphere, but on Mars, due to a lack of precipitation to the surface (at least outside polar regions), the impact is milder: dust particles are removed from the atmosphere at one altitude and either released at a lower level or remain within the cloud layer for an extended period.

Below-cloud scavenging is better understood theoretically, and several expressions have been formulated to describe the efficiency of the process, for specified sizes of precipitate and aerosol particles. It is known to vary strongly with dust particle size, as collection is dominated by Brownian diffusion for small particles and inertial impaction for large particles, leaving a characteristic minimum in scavenging efficiency for particle size, as collection is dominated by Brownian diffusion for small particles and inertial impaction for large particles, leaving a characteristic minimum in scavenging efficiency for small particles and inertial impaction for large particles, leaving a characteristic minimum in scavenging efficiency for small particles and inertial impaction for large particles, leaving a characteristic minimum in scavenging efficiency for small particles and inertial impaction for large particles, leaving a characteristic minimum in scavenging efficiency for small particles and inertial impaction for large particles, leaving a characteristic minimum in scavenging efficiency for small particles and inertial impaction for large particles, leaving a characteristic minimum in scavenging efficiency for small particles and inertial impaction for large particles, leaving a characteristic minimum in scavenging efficiency for small particles and inertial impaction for large particles.

The time taken to fall a distance of 1 km at three different heights in the atmosphere for particles ranging in size from 0.01–10 μm (calculated from equation (19)). The black lines represent particles of density $\rho_i = 2500$ kg m⁻³, while the shaded regions bound the possible values for particles with densities ranging from 1550–3300 kg m⁻³ (extreme values for likely martian dust density). Figure supplied by Dr Liam Steele, The Open University.

For water ice crystals on Mars, downward fluxes are so low (typically $\sim 10^{-10}$ kg m⁻² s⁻¹) that below-cloud scavenging should be negligible, even with 100% collection efficiency. CO₂ fluxes in the condensing winter polar atmosphere can, however, be significantly larger, making below-cloud scavenging potentially important, but strongly dependent on the sizes and densities of the particles involved. Typical CO₂ snow particle characteristics are not known with much certainty, but current best estimates are for radii of $r_e \sim 30–100$ μm (Colaprete and Toon 2002, Hayne et al 2014) and densities of 100–1000 kg m⁻³ (Smith et al 2001b, Matsuo and Heki 2009). Martian CO₂ snow particles are therefore smaller and more dense than the terrestrial particles of radius 500 μm and density ~50 kg m⁻³ used by Dick (1990). Both of these differences are such as to give larger collection efficiencies for the martian snow particles, at least for dust sizes of 0.1 μm. In fact, the larger $D$ and $u_e$ values for Mars mean that for a given ice particle size and density, both terms in $E$ are larger, so that the below-cloud scavenging process should be more efficient, by up to an order of magnitude, than it is on Earth. Scavenging of larger dust particles, of radius around 1 micron, requires CO₂ snow particles to be sufficiently large ($\geq 50$ μm), but recently observed cloud lifetimes, and inferred particle sedimentation rates, suggest that this is the case (Hayne et al 2014).

5. Volatile cycles and clouds

This section discusses the cycles of the two most important volatiles in the martian atmosphere, carbon dioxide (CO₂) and water (H₂O). The following subsections discuss their...
reservoirs, phase changes and the ability of both volatiles to form clouds. Finally, we consider how clouds can be modelled and how their role in the climate can be understood.

5.1. Reservoirs

The primary reservoir for carbon dioxide in the present climate cycle is in the form of ice at the polar caps, with roughly one third of the average mass of the atmosphere subliming from seasonal ice deposits and moving from the spring pole to the autumn pole, where it re-freezes, each year (e.g. Piqueux et al 2015). In addition to these seasonal polar caps, a small amount of CO₂, equivalent to a few per cent of the total atmospheric mass (Byrne and Ingersoll 2003), currently remains permanently frozen in a residual cap at the South Pole. The reasons for the persistence of a CO₂ ice cap at the South Pole but not at the North Pole are not fully understood, but are thought to involve the high southern cap albedo (Jakosky and Haberle 1990, Guo et al 2010). Nevertheless, the residual cap mass varies on relatively short timescales, and there is evidence both for net accumulation and for net ablation (Byrne 2009) occurring under present-day climatic conditions.

In addition, Phillips et al (2011) have identified buried CO₂ deposits up to 1000 m thick within the southern polar layered water ice cap. These might be released at higher obliquities, around 35°, conditions which have occurred as recently as 625 000 years ago, as seen in figure 37 (Laskar et al 2004). If all the CO₂ in this reservoir was to sublime, the atmospheric mass would increase by up to 80%, raising surface pressure from an average of 610 Pa to around 1000 Pa (Phillips et al 2011).

Diverse geomorphic evidence for substantial aqueous flow on the martian surface in the distant past requires a more massive atmosphere with a higher pressure in order that water may exist in a fluid form rather than the direct phase transition from ice to vapour, which occurs below the triple point at 610 Pa (see figure 31), perhaps coincidentally close to the average surface pressure on Mars today. This may also require a warmer atmosphere so that water does not remain primarily in the form of solid ice near to the poles. Section 6 will discuss the early Mars atmosphere problem in more detail, but it is worth noting that, with a less luminous early Sun, at the very least not only is much more atmospheric carbon dioxide required but other volatiles and perhaps volcanic gases and particulates are needed to maintain a sufficiently warm greenhouse environment, although the details remain very uncertain.

If a much more massive atmosphere existed on early Mars, the location of much of that carbon dioxide in the present day is still a mystery. The most likely scenario is that the long-term reservoir of carbon exists in the form of rocky deposits, as on Earth. Relatively few carbonates have been found to date, however, by either orbital surveys (Ehlmann et al 2008) or surface rovers (Clark et al 2005). Instead, the mineralogy possibly implies an acidic environment when bodies of water were present (Greenwood and Blake 2006), which did not form carbonate rocks. The lack of any substantial carbonates near the surface implies that, if there was a massive ancient atmosphere, it must now be stored deep below the surface.

Water vapour is, by comparison, a very minor constituent of the martian atmosphere, often amounting to less than a few precipitable microns (pr μm, the equivalent depth of the water in the atmospheric column if all were condensed at standard temperature and pressure), and very rarely approaching 100 pr μm (Smith 2002, Fedorova et al 2004, Smith et al 2009). The North Pole is covered by a permanent water ice cap, extending asymmetrically to 80–85° N, of thickness up to a few km (Zuber et al 1998), which provides a source of water for the atmosphere when this residual cap is exposed to insolation in spring, after the sublimation of the seasonal CO₂ cap. A smaller amount of residual water ice is present at the South Pole (Bibring et al 2004).

Additional, large amounts of water are stored in the subsurface in the form of ice. The total volume of water stored
at the polar regions, in the so-called polar layered deposits, is estimated to be equivalent to a global liquid layer $O(10\ \text{m})$ in depth (Smith et al. 2001a). Smaller quantities of subsurface water ice are present in mid-latitude regions (Holt et al. 2008). The distribution of these subsurface water reservoirs, which are largely inaccessible to the atmosphere under present conditions, holds information about past martian climates (see section 6).

5.2. Cloud formation and precipitation

Clouds in the martian atmosphere are composed of two condensibles: (1) carbon dioxide ice, mainly in the winter polar regions; and (2) water ice, both near the ground as fogs (see figure 32) and from about $10-30\ \text{km}$ in the atmosphere, particularly near the equator and over mountains (see figure 33) during northern summer, which coincides with aphelion, when Mars is furthest from the Sun. Water liquid droplet clouds, as are often found at lower altitudes on Earth, are not thought to occur on Mars as a result of the much lower pressure, since water can only exist as either vapour or ice below its triple point (about $610\ \text{Pa}$ under typical martian conditions; see figure 31).

Thick carbon dioxide clouds form in the polar night, at altitudes ranging from the surface to almost $40\ \text{km}$. They release substantial latent heat that holds the atmospheric temperature close to $145\ \text{K}$, the same as the near-surface atmosphere which is in contact with exposed carbon dioxide ice on the surface. Since there is no incoming solar radiation in the polar night, the clouds have little radiative impact at visible and shorter wavelengths. If CO$_2$ snow falls, it will either accumulate on the surface or re-evaporate on the way down should it encounter warmer conditions, taking up latent heat and so cooling the local atmosphere. These processes will tend to bring the atmosphere around and below the polar hood clouds to an isothermal state, at the same temperature as the ice on the surface. This means that the clouds also have limited infrared impact, with emission to space from the tops of the clouds being similar to that from the ice-covered surface in cloud-free regions.

Upon reaching the surface, CO$_2$ snow undergoes metamorphosis (Elusztkiewicz 1993) and transforms, over the course of several days, from granular ice into non-porous slab ice, decreasing its reflectivity in the process. Nevertheless, the presence of more reflective, freshly fallen snow grains may be important to reducing sublimation rates and allowing the preservation of the

---

**Figure 32.** Image of the western part of Valles Marineris, taken by the High Resolution Stereo Camera (HRSC) aboard Mars Express on 25 May 2004, showing dense ground fog (Möhlmann et al. 2009). Image credit: ESA/DLR/FU.
perennial south polar residual cap (Hayne et al. 2012). CO₂ ice can also form directly on the surface, even at low latitudes if the altitude is high enough, such as near the peak of Arsia Mons on the Tharsis ridge at night (Cushing and Titus 2008). Over high topography, the pressure and thermal inertia of the atmosphere is very low, so temperature can fall low enough at night such that condensation might occur, even allowing for the slightly reduced CO₂ frost point at low pressure. In the daytime the temperature at such locations would, similarly, rise very quickly and the CO₂ frost would sublime. This diurnal behaviour is very different to the lower altitude, polar ice caps that remain at 145 K throughout the diurnal cycle.

The principal source of atmospheric water vapour is the residual ice cap at the North Pole: vapour sublimes as sunlight warms the polar surface in the spring and early summer, after any covering layer of CO₂ ice has first sublimated. The relatively moist air is then transported southwards (as seen in figure 34), initially by waves around the polar cap near the northern spring equinox \((L_s = 0°)\) and then primarily by the large-scale, zonally symmetric Hadley Circulation (see figures 11 and 12), especially moving into northern summer \((L_s = 90°\) onwards). This air rises at northern mid-latitudes and is carried equatorward at higher levels. Condensation at high, cold levels forms an equatorial band of ice clouds, known as the aphelion cloud belt, in northern summer (figure 34), and thereby prevents some of the water vapour from reaching the southern hemisphere (Clancy et al. 1996). Later in the year, water vapour returns to the north polar cap, or condenses onto the cold South Pole. In southern hemisphere spring and summer, water vapour is emitted from the polar cap in smaller amounts, and no equatorial cloud belt forms (figure 34), as Mars is then closer to the Sun and slightly warmer on average.

On the Earth, the latent heat of water plays a crucial role in determining the stability of the atmosphere. Gill (1982) defines an equivalent potential temperature \(\theta_r\) to take account of this,

\[
\theta_r = \theta \exp \left(\frac{Lq_w c_p}{e} T\right)
\]

with \(L\) the latent heat of sublimation of water ice (or of evaporation of liquid water, as appropriate), \(q_w\) the mass mixing ratio of water, and \(\theta\) the potential temperature for a dry air parcel at temperature \(T\) and pressure \(p\) (see equation (5)). As discussed above, in section 2.2.1, an unstable column of air is one with an equivalent potential temperature profile such that \(d\theta_r/dz < 0\) which, on Earth, occurs more readily than does the equivalent condition for a dry air column, \(d\theta/dz < 0\). In such a situation, convection occurs, which brings the profile back towards neutral stability.

On Mars, Hinson et al. (2014) have calculated that even for a (highly unlikely) atmosphere that is fully saturated throughout the vertical column, \(d\theta_r/dz\) differs from \(d\theta/dz\) by only about 0.1 K km\(^{-1}\) as a result of the low vapour pressure of water. This is insignificant compared to typical static stabilities which are close to two orders of magnitude larger in most cases. Hence the latent heat of water on Mars is of little dynamical importance.

This contrasts strongly with the radiative impact of water ice clouds on Mars, which have been shown to produce local perturbations of 20–30 K around 25 km above the martian surface, especially near the equator (Wilson et al. 2008; see figure 35). These effects can have strong influences in producing layers of differing stability just above layers of cloud (Hinson et al. 2014) despite their apparently moderate optical depth at visible wavelengths, roughly equivalent to thin cirrus layers on Earth.

5.3. Cloud microphysics and modelling

Clouds affect the atmosphere of Mars on both global and local scales. On a global scale, the formation and subsequent
sedimentation of ice particles concentrates vapour near the surface, possibly enhancing the exchange of water with the regolith (Daerden et al 2010) and influencing large-scale transport (Clancy et al 1996). On a local scale, cloud radiative effects (which are dependent on ice particle size) can lead to local heating or cooling of the atmosphere (Wilson et al 2008), and scavenging by sedimenting ice particles could remove dust from the atmosphere, as described in section 4. In order to further understand the role played by clouds in the martian climate, it is therefore necessary for atmospheric models to be able to accurately predict their formation and evolution. To do this, we need to model the relevant microphysical processes: nucleation, growth and sedimentation.

As the martian atmosphere is colder and more tenuous than the Earth’s, martian microphysics schemes can be simplified in comparison to their terrestrial counterparts. The lack of liquid water in Mars’ atmosphere means that clouds are formed solely of ice, and have been likened to cirrus clouds (Whiteway et al 2009) or polar stratospheric clouds (Burlakov and Rodin 2012) on Earth. Therefore, liquid-only and liquid–ice processes (such as melting and riming) need not be considered. On the other hand, ice nucleation processes on Earth are relatively poorly understood, and the lower temperatures and pressures of the martian atmosphere introduce a region of phase space that existing modelling methods may be ill-equipped to represent, in their current form.

As the processes of nucleation, growth and sedimentation are dependent upon particle size and particle interaction, a detailed study of the formation and evolution of clouds requires the simultaneous consideration of many different aerosol types and sizes. Bin-resolved schemes account for particles of ice, dust, and ice-coated dust, with each type having a range of discrete size bins. As theory predicts that nucleation and condensation occur on short timescales (Lamb and Verlinde 2011), such schemes also require small model integration timesteps.

When calculating nucleation rates, the presence of numerous dust nuclei at cloud-forming heights means that the probability of heterogeneous nucleation is much greater than that of homogeneous nucleation (Määttänen et al 2005). As such, generally only heterogeneous nucleation is considered, though homogeneous nucleation becomes important at high supersaturations (Burlakov and Rodin 2012). Ice particle growth is then usually assumed to proceed via vapour diffusion only (Colaprete and Toon 2000), since aggregation of particles through Brownian motion is important only for very small (<0.1 μm) ice particles (Michelangeli et al 1993). Sedimentation rates in each size bin are calculated as for small dust particles (see section 4.4.1) with the Stokes relationship, modified by a correction factor to account for the large mean free paths in Mars’ atmosphere.

Bin-resolved schemes are typically used to interpret the results of observations. For example, Colaprete et al (1999) compared model output to data returned from Mars Pathfinder
in order to study cloud formation and radiative effects; Montmessin et al (2002) compared simulated limb profiles with observations from the Viking Orbiter; and Daerden et al (2010) compared model cloud predictions to Phoenix LIDAR observations of boundary layer clouds. While bin-resolved schemes allow for detailed microphysics calculations, their results depend crucially upon parameters that are difficult to determine, particularly the contact parameter, $m$, which exerts a strong control on the rate of nucleation (Michelangeli et al 1993, Montmessin et al 2002, Määttänen et al 2005, Burlakov and Rodin 2012). The value of $m$ varies depending on the type of aerosol, the surface roughness and whether the aerosol is preactivated (has previously had ice formed on it). To date, most bin-resolved schemes have used values of $m$ ranging from 0.925–0.975, but recent laboratory work has shown that $m$ can range from 0.84–0.98 (Iraci et al 2010), which will have implications for simulated nucleation and hence subsequent growth and sedimentation.
To understand the large-scale influences of clouds, by simulating cloud microphysics in a GCM or mesoscale model, computational constraints necessitate the use of ‘bulk’ schemes, in which model macrophysical output at each grid point (such as temperature, pressure and water content) is used as input to a somewhat simplified microphysical model. Early attempts at using a GCM bulk scheme showed that the specification of a uniform ice particle size was an oversimplification (Richardson et al 2002). A variable particle size can instead be used, by assuming that ice particles nucleate heterogeneously around all the available dust nuclei, and that their sizes follow a statistical distribution, such as a log-normal or gamma distribution. With this method, Montmessin et al (2004) successfully reproduced the important observed features of ice particle size seasonal and spatial variability; namely, the formation of larger particles (3–4 μm) in the aphelion cloud belt, and smaller particles (1–2 μm) in the southern winter polar region (Clancy et al 2003). Similar approaches have captured many of the large-scale cloud formations observed by spacecraft, such as polar hoods, mountain clouds and the aphelion cloud belt (Machtoub 2012). However, such schemes are limited by the fact that microphysics calculations are only performed every time the model’s physics schemes are called (which is usually the order of tens of minutes), whereas microphysical properties can vary on the order of seconds.

More sophisticated ‘combined’ schemes attempt to retain more of the information available from bin-resolved schemes by using bulk methods for transport of ice particle mass and number densities, but calculating microphysical processes for discrete size distributions, obtained by conversion according to an assumed size distribution. This combination allows improvements to be made in the calculations of nucleation, growth and sedimentation, while keeping the computing time low when calculating global transport. A combined scheme such as this has been used in a mesoscale model to study mountain-induced vertical transport of water (Michaels et al 2006), and in a GCM to study equatorial water ice deposits (Nelli et al 2009), clouds and fogs observed by Phoenix (Nelli et al 2010) and the complete hydrological cycle with radiatively active clouds (Navarro et al 2014). The results of such simulations match well with available spacecraft data. As yet, however, there have been no comparisons between the predictions of bulk and combined schemes, so it is difficult to say to what extent incorporating bin-resolved microphysical processes into a GCM improves cloud predictions. Future plans to incorporate a 1D bin-resolved scheme into a GCM have also been put forward by Burlakov and Rodin (2012).

6. Outstanding problems

In this review we have attempted to outline many of the key physical processes that govern the climate of Mars, focusing on the basic principles, albeit to some extent at the expense of many of the observational details. The martian atmospheric structure and near-surface environment are determined largely by many processes that have direct counterparts on Earth, but under conditions that are experienced only rarely on Earth, if at all. In many respects, much of Mars is similar to the polar desert terrains found in places such as the Canadian Arctic or in the dry valleys of Antarctica, where precipitation is rare, temperatures are seldom above the freezing point of water and water-related erosional processes are very slow (see http://quest.nasa.gov/challenges/marsanalog/marsanalog.html for more details). Such terrains are comparatively hostile to terrestrial life, though some species are evidently able to cope with such conditions, from which lessons may be learnt about the potential habitability of Mars itself.

But the analogy between such harsh and extreme terrestrial environments and the surface of Mars can only ever be partial, not least because of the different atmospheric composition of the martian atmosphere and the much lower surface pressure. So it is important to evaluate insights gained from studying any potential analogues against a first principles approach which focuses attention on the most important physical processes that may be relevant. In this section, we briefly summarize some of the key points that emerge from our overview of these physical processes, with particular reference to our understanding of the present martian climate and how this might have evolved over the history of the planet on both geologically short and long timescales.

6.1. Deep-time past climate: warm and wet Mars?

Some of the most intriguing and poorly understood questions relating to the past climate of Mars concern the origin and history of morphological features at the martian surface that appear to require the action of large and sustained volumes of a flowing liquid, such as water. These features include widespread networks of sinuous valleys in the southern highlands of Mars, widespread plains in the tropics and northern sub-tropics that appear to be populated by erosion features consistent with massive flooding events, streamlined ‘teardrop-shaped’ islands, broad channels and chaotic terrains that appear to have been formed by catastrophic melting of ground ice, and even ‘fossil coastlines’, along with the suggested presence of a large-scale ocean covering much of the northern plains to a depth of up to 500 m (Baker et al 1991, Parker et al 1993, Head et al 1999). Although suggestions of other materials (such as liquid CO\textsubscript{2} (Hoffman 2000), wind erosion (Leovy 1999) or lava flows (Leverington 2011)) have been made in the past, the ubiquity of such features and their detailed morphology seem to demand an interpretation in terms of a fluid with a very low viscosity (ruling out most lavas) that must almost certainly involve large amounts of water in liquid form. Even in regions devoid of such massive flood features, mineral deposits have been observed, both from orbit and (recently) from surface rover investigation (Squyres et al 2008), consistent with sustained lakes of liquid water or brines.

Cratering statistics suggest that such landforms are typically very old, with many dating from earlier than 3 Gyr BP (before present). Indeed many of the sinuous valley networks in the southern highlands appear to date from a fairly narrow range of ages from around 3.4–4 Gyr BP (Hartmann 2005, Fassett and Head 2008). Their formation appears to require
CO2 to values up to 5 bar, and suggested this might be sufficient to produce the temperatures required, simply by increasing the surface pressure of the Sun was significantly lower than today, by a factor of around 20–30% (Gough 1981).

Such a massive change in atmospheric mass and composition suggests a major evolution since around 2.5–3 Gyr BP. Geological evidence indicates that large-scale flooding events have been much rarer since around 2.7 Gyr BP (in the so-called Amazonian era), suggesting that such massive changes took place apparently irreversibly around that time, possibly leading to the permanent loss of large amounts of atmospheric mass either to space or to the deep subsurface. Current understanding suggests that accelerated loss to space may have played a major role, and so NASA’s MAVEN mission (Jakosky et al 2015) has been launched recently to attempt to obtain detailed measurements of loss mechanisms at the atmospheric exobase under present conditions, with the hope of gaining clues as to what might have changed since the Noachian era, more than 3 Gyr BP.

The current debate over the nature of the early martian climate and how to achieve a much more active hydrological cycle than in the present climate has tended to focus on three main areas of uncertainty:

1. How to sustain temperatures near the surface sufficient to maintain water (or brines) in liquid form, despite a faint young Sun (FYS)? Could a different composition have produced a more substantial greenhouse warming to enable this? Could brines (with a much reduced melting point) account for widespread liquid reservoirs?
2. How to sustain large areas of open water in the presence of polar ice caps and the hemispheric asymmetry in surface topography?
3. What was the role of large impact events? Could they have produced sufficiently large-scale melting to sustain water in liquid form for long enough to produce the observed erosion?

6.1.1. Early martian greenhouse warming? Allowing for the possibility of accelerated loss of atmospheric mass since the late Noachian era, initial suggestions for sustaining temperatures sufficient to keep water in liquid form focused on the ability of a much more massive CO2 atmosphere alone to achieve this. The early study by Pollack et al (1987) computed the warming produced by increasing the surface pressure of CO2 to values up to 5 bar, and suggested this might be sufficient to produce the temperatures required, simply by increasing the overall infrared opacity of the clear atmosphere. It was realized somewhat later, however, by Kasting (1991) that such a massive CO2 atmosphere would become supersaturated at higher altitudes in the middle martian atmosphere, because of the tendency for the atmosphere to cool with height. Thus, much of the CO2 with a 5 bar surface pressure would condense into ice clouds in the middle atmosphere and begin to precipitate. Such dense clouds would also substantially increase the albedo, reflecting sunlight back out to space, causing the surface to cool, eventually allowing solid CO2 to accumulate onto the surface until the surface pressure reduced to a maximum of around 350 hPa. With such a pressure the atmosphere would not be sufficiently opaque in the infrared to provide sufficient greenhouse warming to sustain pure water in liquid form unless the Sun were no less luminous than around 85% of its current luminosity. This would rule out this mechanism alone as a means of accounting for the observed hydrological erosion features.

The ability of dissolved salts to reduce the freezing point of water has also been suggested as a means of mitigating the effects of a FYS and insufficient greenhouse warming to reach 273 K (Haberle et al 2001). However, most commonly occurring mineral salts will only depress the freezing point by around 5 K or so, which is not sufficient to mitigate the effects of a FYS and weak greenhouse. Recent observations, e.g. by the Phoenix Lander (Renno et al 2009) and Curiosity Rover (Glavin et al 2013, Martin-Torres et al 2015), however, have revealed the presence of perchlorate salts in some subsurface ices and soils, even close to the equator. Perchlorates can depress the freezing point of eutectic mixtures with water to around 200 K, but the viscosity of such mixtures is much higher than that of pure water. Recent Curiosity observations indicating the presence of perchlorate brines in Gale Crater at low latitudes may indicate the widespread existence of subsurface brines across the planet, though it is perhaps unlikely that such perchlorates would be present in sufficient quantities to fill a substantial ocean on a planetary scale. But it is possible that they could lead to the transient activity of liquid-induced erosion, with impacts on geomorphology and geochemistry.

More recent studies of greenhouse warming on Mars have looked elsewhere for means to increase the greenhouse potential of the martian atmosphere. The likelihood of increased vulcanicity during martian early history suggests that other chemical trace species might have been present in large quantities during the Noachian era. Suggested species have included NH3, SO2, CH4, H2S, H2O, organic hazes and even molecular H2 itself (e.g. see Johnson et al 2008, Mischna et al 2013, Ramirez et al 2013, Sagan and Mullen 1972). Sulphur compounds have received quite a lot of attention (see Johnson et al 2008, Mischna et al 2013, for more discussion), since they can provide opacity in regions of the infrared spectrum that complements that due to CO2 and H2O. A likely snag, however, is that sulphur compounds can readily decompose photochemically to form solid aerosol particles that are efficient scatterers, thus contributing to a cooling effect to offset greenhouse warming. However, combinations of CO2, H2O and SO2, together with appropriate conditions in terms of surface albedo etc, might be able to conspire to create local conditions conducive to maintaining liquid water for at least part of the martian year (Mischna et al 2013), even under FYS conditions. H2 is another recently suggested possible constituent of the early martian atmosphere (Ramirez et al 2013)
which could also efficiently increase infrared opacity through collision-induced absorption effects. However, it is not clear how long-lived molecular hydrogen would be in the martian atmosphere, which would determine how effective it might be on the timescales necessary to sustain a hydrological cycle sufficient to produce such massive erosion features.

Not all aerosols necessarily act to cool the atmosphere, however. Volcanic ash and black carbon (if emitted), for example, are more likely to warm than cool, depending upon their albedo in the visible. In an early study, Forget and Pierrehumbert (1997) showed that thick CO2 ice clouds would act as efficient scatterers of infrared radiation, which could act, at least in principle, to enhance greenhouse warming of early Mars. More recent work using 3D circulation models (Forget et al 2013), however, indicates that CO2 cloud cover needs to be fairly widespread to be effective as a greenhouse agent, and even then will only apparently provide up to \( \sim 15 \) K of warming (at around 2 bar surface pressure). As discussed earlier in this paper, water ice clouds can also have a strong influence on the radiative budget of the atmosphere, which can also lead to surface warming under some circumstances. Urata and Toon (2013) and Wordsworth et al (2013) have recently explored this in 3D circulation models, but find the amount of warming may be highly localized and is quite sensitive to a number of relatively uncertain parameters.

6.12. Warming by impacts. An alternative approach to the question of obtaining relatively warm surface temperatures on Mars has been extensively discussed in the context of the much higher rate of meteoritic impacts during the early history of planets such as Mars. This “early heavy bombardment” era represents the end-point of the formation of the Solar System in the form that is currently familiar, during which migration of the orbits of the gas giant planets is thought to have perturbed clouds of debris left over from the initial burst of planet formation, causing this debris to impact onto the surfaces of the main planets and forming much of the heavily cratered terrain that still survives on planets like Mars, the Moon and Mercury to the present day.

Such a large and energetic impact can deposit a huge amount of energy, and matter including water ices and other volatiles, into the atmosphere. To have an effect on a planetary scale, impactors need to be larger than a few tens of km in diameter, as a result of which the atmosphere may be made significantly wetter and warmer, at least for a while. However, calculations suggest (e.g. see Segura et al 2002) that, even with large impacts (\( \sim 250 \) km in diameter; large enough to create the 1500 km diameter Hellas Basin on Mars), the effective warming associated with the energy of the impact is unlikely to persist for more than a few tens of years. Whether or not this would be sufficient to produce a vigorous enough hydrological cycle to account for the features observed is unclear, although it might perhaps be able to explain the formation of massive flood landforms. Moreover, the interval of massive impacts was relatively short-lived and was largely concluded by around 3.2 Gyr BP; so more recent hydrological activity may well require an alternative explanation.

Thus, we see that the character of the martian climate has changed enormously since the early history of the planet, with likely intervals when the planet was able to sustain liquid water at the surface (moderated either by changes in atmospheric composition, enhancing greenhouse warming, or via major meteoritic impacts), interspersed with extended periods when the climate was much drier and colder (e.g. see Wordsworth et al (2015) for some recent scenarios using a 3D climate model). As time progressed, much of the early atmosphere was lost until we arrived at the present situation, when the climate is slowly modulated by cyclic changes in orbit and rotation. But this corner of the subject remains in a somewhat uncertain state with a number of questions still unanswered—at least to the level of confidence one might wish for—and there remains a strong need for more observational information and measurements from the surface and subsurface to constrain evolutionary models.

6.2. Stability of the current climate and cyclic variability

Although recent research has unveiled many of the key factors that influence the form and structure of the current martian climate, as reviewed above, it is of significant interest to examine to what extent the present or recent climate represents a stable equilibrium state, or whether it is subject to transient changes with time. The previous subsection of this review has discussed some of the issues related to the state of martian climate in the distant past, more than 1 Gyr ago. But there is a significant body of evidence to suggest that martian climate has undergone variations on much shorter timescales recently (in geological terms), measured in millennia or Myr. These include the occurrence of geological features that are apparently relatively recent, such as those associated with glacial action and erosion (e.g. Head et al 2005) or the formation of “gully” flow features (Malin and Edgett 2000), which may be only a few Myr old at most (according to age determinations based on cratering statistics). Such features would seem to indicate prevalent conditions (in terms of humidity and action of water at the surface, albeit sometimes in frozen form) that differ markedly from what we observe today, although recent evidence (e.g. Cushing and Titus 2008, Raack et al 2015) may suggest other explanations for present-day gully formation involving CO2 ice processes.

Close to both poles, though outside the permanent ice caps, further evidence of transient climate variability is inferred from the presence of extensive layers in the soil of varying mixtures of water ice, dust and sand, the patterns of which appear to be coherent across large distances, based on both visual observations (Milkovitch and Head 2005) and ground-penetrating radar (Phillips et al 2008, Putzig et al 2009). Such layers appear to have been laid down gradually over many millennia by atmospheric deposition of varying concentrations of dust and ice, and have been widely interpreted as indicating cyclic climate change on timescales of millennia to Myr. Figure 36 shows a set of MOC images of the north polar cap, where frames A–C gradually zoom in onto the margins of a patch of ice to reveal a sequence of these layers, each containing varying amounts of ice and dust. The layers range
in thickness from $\sim$5–30 m and are coherent horizontally over distances of several hundred km.

Such observations have been further compounded by the discovery of landforms in low and middle latitudes (e.g. Holt et al. 2008, Murray et al. 2005) indicating the presence of significant amounts of subsurface ice that models suggest should not be in long-term equilibrium, but should have sublimed away long ago.

6.2.1. Cyclic climate variability. The idea that martian climate might vary cyclically on these timescales has recently gained much credibility with the discovery (Laskar et al. 2004) that, thanks to gravitational perturbations from other planets within the Solar System, the rotational and orbital elements of Mars undergo slow (but chaotic) oscillations with periods ranging from a few tens of millennia to around 2.5 Myr (see figure 37). These oscillations are analogous in many respects to the so-called Milankovitch cycles, affecting planetary obliquity, precession and orbital eccentricity, which are thought to act as ‘pacemakers’ of the cycles of glaciation on Earth on timescales of $10^3$–$10^5$ yr (Hays et al. 1976). Precisely how such ‘pacemaking’ works for the Earth is still somewhat controversial (e.g. Wunsch 2004), since the amplitude of insolation variations near both poles is much weaker than apparently needed to induce major glacial episodes. The mechanism may be more akin to a weak but coherent nonlinear synchronization between certain components of the externally driven Milankovitch oscillations (mainly in obliquity and eccentricity) and a naturally self-induced internal oscillation of the climate system that favours timescales between glaciations of around 100 kyr (Tziperman et al. 2006). In the martian case, however, the situation seems more clear-cut, since the amplitude of several of the inferred variations in external forcing is much larger than for the Earth, with variations e.g. in obliquity over the range $\sim$15°–45° (with a current value of 25°; cf an obliquity of 23° ± 1° for the Earth). This would be expected to lead to changes in annual mean insolation at high latitudes over a range $\sim$150–450 W m$^{-2}$ (see figure 37(c) (Laskar et al. 2004)). Even relatively simple, energy-balance climate models (e.g. Armstrong et al. 2004) suggest that this will lead to large changes in the equilibrium climate over the whole planet, given time for the atmosphere and near surface environment to adjust.

Although changes in orbital ellipticity and the phase of precession will have some effect on the distribution of incoming sunlight in space and time during the martian year, probably the most substantial climatic impact comes from slow changes in the planetary obliquity (Laskar et al. 2004). This is
the dominant parameter in determining the nature and intensity of the seasonal cycle on a planet, with a small obliquity favouring only weak seasonal variations with perpetually weak insolation onto both poles. At high obliquity angles, however, the polar regions are subject to large changes in insolation between winter and summer, with high latitudes remaining unilluminated for much of the winter but then relatively strongly heated during summer time. This can have the effect of melting and evaporating large volumes of polar ice during summer at high latitudes, which may be transported as vapour to other regions of the planet. Averaged over the year, high obliquity angles tend to reduce the overall difference in insolation between poles and the equator, with the poles actually receiving more integrated energy than the equator for obliquity angles $\geq 50^\circ$ (Ward 1974, Laskar et al 2004). The timescales for deposition and evaporation of water and CO$_2$ surface and subsurface ice may differ markedly, so such changes in seasonal insolation levels can have the overall effect of causing water ice to migrate systematically from/to high latitudes to/from lower latitudes as obliquity increases/decreases over intervals of decades to millennia or even longer. High obliquities will also tend to favour the release of CO$_2$ that may have been stored in the subsurface in polar regions during lower obliquity intervals, given time for the increased heating from polar insolation to penetrate deep into the soil (which can take $\geq O(10^4$ yr) for depths of 1 km or greater). Thus, the strong variations in martian obliquity with periods of up to 2.5 Myr may be expected to result in major changes in atmospheric pressure, humidity and distribution of surface ice.

Mars’s current obliquity is close to that of the Earth, and so the equilibrium distribution of surface water ice favours accumulations close to both poles, much as observed. At lower obliquities, this trend becomes even more pronounced,
with lower mean temperatures at both poles. This may result in more of the CO$_2$ atmosphere condensing onto the poles quasi-permanently, since the lower amplitude of the seasonal cycle leads to cooler summers at high latitudes, which may be insufficient to evaporate all of the CO$_2$ that condensed during winter. So the mean surface pressure during the year across the whole planet will tend to reduce as more CO$_2$ ice accumulates near the poles. At higher obliquities than at present, however, the opposite will tend to occur, releasing any CO$_2$ ice currently resident beneath the polar ice caps and increasing the mean surface pressure. A key uncertainty, however, is the factor by which the surface pressure can increase, since the extent of any residual subsurface CO$_2$ ice reservoir is still largely unknown, mainly because much of it is probably inaccessible to current remote sensing techniques. However, subsurface water ice close to the surface can be inferred from gamma ray spectrometer data (Boynton et al. 2002), which clearly show the existence of large ice reservoirs locked up in the polar layered terrains surrounding the main ice caps in both hemispheres. Some subsurface ice is also found at low and middle latitudes, in regions where one would not expect ice to be in equilibrium (e.g. Head et al. 2005, Holt et al. 2008). This would seem to be a left-over from a previous high obliquity interval, when low latitude ice was in equilibrium.

6.2.2. Uncertainties in recent palaeoclimate models. Various approaches have been taken in attempts to quantify the response of the martian climate system to such strong changes in insolation associated with cyclic variations in obliquity and other rotational and orbital elements. The earliest work (e.g. Jakosky et al. 1995, Kieffer and Zent 1992, Leighton and Murray 1966, Paige 1992, Ward 1974) made use of simple, one-dimensional energy balance models of the surface and mean atmospheric column, solving for the radiative balance at the surface and throughout the atmospheric column, essentially averaged across the planet. This enabled a first-order evaluation of the typical response of the climate system, including impacts on surface and atmospheric temperatures, ice deposits and clouds, though did not provide much information on how different latitudes would respond.

These models were subsequently extended to provide some resolution in latitude (e.g. Armstrong et al. 2004, Greve et al. 2010, Paige 1992), allowing for differing levels of insolation at different latitudes and seasonal variations. The relative simplicity of these models enables them to simulate evolution of their climate over long timescales, up to ~1 Gyr, including the cyclic variations of orbital and rotational elements of the planet. However, despite their complexity in relation to radiative and thermodynamic balances, they need to make a number of simplifications, in particular neglecting the effects of atmospheric transport and transport of ice within ice sheets. The latter can be taken into account by coupling atmospheric models with explicit ice sheet models (e.g. Stenzel et al. 2007), although at the expense of representing much of the 3D spatio-temporal details.

More recently, a number of groups have attempted to adapt full-scale three-dimensional time-dependent GCMs and similar kinds of model to study this problem with much more geographical detail. For epochs up to a few Myr ago this problem is reasonably well-posed, since the topography of the surface is unlikely to have changed significantly over those timescales, at least on large spatial scales. It is then relatively straightforward to adapt even fairly complex models of the current climate to different orbital and rotational elements to obtain seasonally and diurnally varying insolation patterns consistent with past epochs (e.g. Haberle et al. 2003, Mischna 2003, Newman et al. 2005, Richardson and Wilson 2002a). The most recent models (e.g. Forget et al. 2006, Madeleine et al. 2009) include fairly full representations of cloud formation and the hydrological cycle, and attempt to reconstruct in some detail the accumulation and precipitation patterns of surface ice during epochs of high obliquity. The results are often highly suggestive of explanations for some of the geomorphological features that indicate glacial action in the geologically recent past.

However, these models rely heavily on extrapolating the behaviour of microphysical and radiative parameterizations to conditions that are well removed from those of present Mars, with rather higher atmospheric pressures and humidities than experienced currently. The role of dust particles as condensation nuclei is fairly crucial to these predictions, but our understanding of dust lifting and deposition processes is limited even under current conditions, so the application of dust schemes to past climates remains speculative. This is particularly problematic to studies of recent palaeoclimate since dust is also likely to play important roles in modulating atmospheric temperatures and large-scale circulation strength controlling the transport of water between equatorial and polar regions (Haberle et al. 1982, Montmessin et al. 2007), and even in affecting the stability of subsurface ice deposits by altering their albedo through dry and wet deposition processes (Toon et al. 1980, Hofstadter and Murray 1990).

The lack of knowledge of the size of any residual subsurface CO$_2$ reservoirs in the martian system leads to significant uncertainty in the equilibrium surface pressure at other epochs. Thus, many of the basic uncertainties in the dust and hydrological cycles for modeling the present climate on Mars become substantially amplified when dealing with past epochs, though the sense of likely trends found in such models may well be robust. Nevertheless, a further source of uncertainty is the inability to run such climate models for more than a few martian years, because of their computational expense. So it is necessary to extrapolate the equilibrated response of those parts of the climate system that only respond on very long timescales.

6.3. Present climate

From the discussion presented elsewhere in this review, it is clear that the overall energy balance of the martian surface is determined much more by direct radiative exchanges between the surface, involving solar heating and direct thermal radiation to space, than on Earth, because of the relative transparency of the almost pure CO$_2$ martian atmosphere in the infrared. Although atmospheric convection and transport may
be quite active, their overall effect is relatively weak compared with the Earth. This means that surface conditions are determined primarily by latitude and time of day. Variations in the altitude of the surface therefore have much less effect, e.g. on mean surface temperature, than on Earth, where atmospheric heat transport is much greater, due to its greater atmospheric density, and surface temperatures are determined much more by the tendency for convecting air to cool adiabatically with height. Combined with the absence of open oceans on Mars, so the surface thermal inertia is relatively small compared with Earth, this implies that the large amplitude topography of Mars exerts both a mechanical and a thermal perturbation on the martian atmospheric circulation, with high elevation acting as a heat source as well as a mechanical obstacle to the flow.

The relatively low thermal inertia of both surface and atmosphere means that the martian atmosphere is relatively strongly damped by radiative cooling effects. Combined with the effect of the much smaller planetary radius compared with the Earth, models suggest that the martian atmosphere ought to be in a different dynamical regime to the Earth, for which baroclinic instability might be expected to lead to disturbances of longer wavelength compared to the planetary radius. In such a regime, there are many fewer unstable modes available than on Earth, which simple models suggest should lead to a less chaotic and turbulent atmosphere, at least on large scales (though the strong local daytime heating at the surface and rugged topography would be expected to lead to strongly turbulent flows on a small scale). This seems to be at least partially borne out in observations of martian weather systems, and in more complex and realistic models. However, Mars is also subject to some strongly nonlinear processes that may promote chaotic variability, associated e.g. with the lifting and transport of radiatively active mineral dust, that play a much stronger and more widespread role on Mars than on Earth.

6.3.1 Uncertainties in the dust cycle. As discussed in section 4 above, one of the strongest sources of nonlinearity in the martian climate system at present relates to the lifting, large-scale transport and deposition of dust. Because the atmosphere has such a low thermal inertia and strong response to radiative forcing, suspended dust particles can rapidly affect the thermal state of the air, resulting in a strong positive feedback on the circulation which, in turn, can stimulate increased dust lifting activity, both locally and potentially at distant locations (Newman et al 2002b, Strausberg et al 2005, Martínez-Alvarado et al 2009b). The threshold dependence of the onset of dust lifting with wind strength adds to the nonlinearity of the process, resulting in chaotic and potentially explosive growth of storm events. Moreover, because much of the martian surface is a dust-laden desert, dust storms can grow to huge scales across large distances before dissipating—in stark contrast to Earth where dust-rich deserts are geographically confined. Thus, although large dust storms on Earth are not unknown, they are relatively rare on a continental scale and pretty soon run into either moist ground or the oceans, which cuts off the surface supply of dust.

Models of the martian dust cycle have made rapid progress in many respects, utilizing techniques that directly parallel those that have been used successfully in climate models for the Earth. But as discussed in section 4, there are still many respects in which these models fall short of capturing the observed behaviour of dust events on Mars. Reproducing the observed climatology and statistics of regional and large-scale dust storms clearly poses major challenges to such models. In their present form, global models typically require some careful tuning of (semi-empirical) lifting parameters to achieve just the right sensitivity of lifting rates to reproduce the intermittency and intensity of growth rates observed. Moreover, such tuning turns out to be highly sensitive to the spatial resolution used in the model, with little evidence so far of convergent behaviour as resolution is increased. Attempts to refine and improve some of the physical processes being parameterised (e.g. taking into account sub-gridscale storage and release of dust (Mulholland et al 2013)) have helped to alleviate this problem somewhat, but have by no means solved this issue.

Even once dust lifting is initiated within a model dust storm, the form of the lifted dust cloud may not resemble the form observed, especially in relation to the evident formation of reasonably long-lived elevated layers and concentrations of dust over large regions of the planet. This remains one of the key unsolved issues in our understanding of how dust storms develop and evolve on Mars. In particular, it is still not entirely clear whether these layers form as a result of intense and horizontally confined vertical injections of dust to high altitude (the so-called ‘rocket dust storm’ events (Spiga et al 2013)) or due to some selective removal process (e.g. forms of wet deposition) that clears the atmosphere of dust at lower levels. Preliminary model investigations seem to suggest that scavenging of dust particles by water ice condensation may not be efficient or effective enough to achieve this, mainly because the associated precipitation is insufficient to reach the ground before subliming back to the vapour phase. But the situation is not yet conclusive. Some further observational evidence to support either approach would be beneficial.

The precise ways in which dust may be lifted from the surface layers of the atmosphere into the freely convecting regions are still not well understood. This manifests itself in models in the continuing need to rely on ad hoc empirical parameterizations of the threshold lifting rates and their dependence on near-surface wind speeds. The classical approaches based on simple saltation and parameterized sand fluxes may be too simplistic to reflect realistic scaling laws, and further research is clearly needed to quantify these dependences (ideally in experiments and field studies (e.g. Merrison et al 2007)). A closer link between Mars- and Earth-oriented research in small-scale dust processes would clearly be of benefit here, since there is much in common on both planets.

6.3.2 Volatiles and condensibles. Recent studies have clearly shown the importance of condensible species, such as CO₂ and H₂O, in affecting the surface climate in various ways. Clouds may form which strongly affect the local radiative energy budget, either by scattering sunlight or upwelling
infrared radiation. Surface ice may also have a strong impact on the surface albedo, again influencing the surface and atmospheric energy budget. But until recently the approach towards representing such clouds in models has been at best rather crude, if included at all. A major issue is that cloud formation and dissipation relies on processes taking place on a very small (micron) scale, under conditions on Mars that are difficult to relate to those more familiar on Earth and amenable to direct observation and measurement. Even on Earth, there are many uncertainties concerning the details of cloud formation and its impact on large-scale climate (Boucher et al 2013). On Mars, this is further compounded by the much lower atmospheric pressures and almost pure CO₂ composition of the atmosphere, which necessitates specific investigations of the microphysical processes in both CO₂ and water ice under the relevant conditions.

Several recent studies have highlighted the fact that cloud nucleation on Mars may be rather different to the equivalent terrestrial process. These primarily relate to the ease with which nucleation occurs when the atmosphere becomes saturated with water vapour. In an analysis of SPICAM data Maltagliati et al (2011) found that supersaturations do occur in the martian atmosphere: in fact, 60% of the data samples showed supersaturation, and supersaturation ratios as large as 5 were observed. At the same time, laboratory nucleation studies (Iraci et al 2010, Ladino and Abbatt 2013) have found that the critical supersaturation increases to values substantially greater than 1 at the low temperatures (∼200 K) experienced on Mars. These results point to a heterogeneous nucleation process that is much more inhibited in martian conditions than has previously been assumed, and which leads to the formation of supersaturations where dust nuclei are in relatively short supply. Therefore, more laboratory and modelling studies of the nucleation efficiencies in various conditions and on various aerosols would be beneficial, to understand this temperature effect further (for example, Määttänen et al 2007) have recently modelled two-component (H₂O and CO₂) heterogeneous nucleation.

Also, particle sizes are assumed to follow log-normal or gamma distributions, but as found by Montmessin et al (2002), sometimes a bi-modal distribution produces a better match to observations. More research is required in order to determine the typical size distributions of various aerosols, as this plays a role in bulk schemes (for calculating sedimentation), in bin-resolved schemes (for setting up an initial dust distribution) and in combined schemes (for converting mass and number densities to size bins).

6.3.3. Liquid water on Mars? As mentioned already, present martian conditions are such that while the combination of atmospheric pressure and near-surface temperature on average does not support the existence of water in the liquid phase, the values are close enough to the triple point (figure 31) that it is possible that liquid water could form sporadically on Mars. This could occur either in situations in which the temperature reaches 273 K, the most favourable combination of factors for which would be a Sun-facing slope in low latitudes, during southern summer, when the planet makes its closest approach to the Sun, or if the salt content of the liquid solution is large enough that the freezing point is lowered beyond the triple point of pure water.

The second of these cases has been confirmed to occur, in a quite extreme form, by observations made by the Phoenix lander (Renno et al 2009), situated at a relatively high, cold latitude (68° N). Phoenix detected the temporary appearance of a liquid brine at atmospheric temperatures of 200–220 K, believed to originate from ice stored very near the surface and then exposed in the presence of the lander. At the least, this confirmed previous satellite measurements (Boynton et al 2002) indicating that significant quantities of water ice are present in the upper metre of the martian soil (Renno et al 2009), available as a potential source of the liquid form. Liquid brines, forming in the near-subsurface at night and evaporating in daytime, have more recently been detected at equatorial latitudes by the Curiosity rover (Martín-Torres et al 2015).

Indirect evidence for the relatively recent presence of flowing water also exists. Geologically recent gully features on the surface have been determined to have been formed by flowing pure liquid water, and not a briny fluid with lower freezing point (Malin and Edgett 2000, Heldmann et al 2005). However, there is uncertainty as to exactly when this flow occurred, considering that atmospheric pressure variations on seasonal and particularly on orbital timescales (as described in section 6.2) could shift the local atmosphere above or below the triple point at the locations of the gullies (Heldmann et al 2005). Possibly more useful indirect evidence for the presence of liquid water (again, possibly in the form of a saline solution) has come in the form of visual observations of ‘recurring slope lineae’. These are flow-like features seen on sloping terrains, particularly those facing the equator, the appearance of which has been seen to vary with changing season, suggesting the action of a flowing liquid (McEwen et al 2014).

So we see that, as mentioned in our Introduction, one of the most pressing uncertainties in both the present and past Mars climate systems continues to be the extent to which water exists, or has existed in the past, in persistent liquid form across the planet. At present this most likely seems to reside in the form of concentrated brines within sub-surface layers, perhaps only a few cm or metres below the surface itself. Conditions there are tantalisingly similar to those found on Earth to nurture extemophile species of micro-organisms, despite the harsh chemical and radiation environment prevalent above the surface, and suggest the possibility of environmental niches on Mars where life might continue to thrive. Such a possibility keeps alive the prospect eventually of finding evidence for either extant life or fossil remains of past life, which will continue to motivate detailed exploration of the Red Planet for the foreseeable future.

Acknowledgments

The authors are grateful to L Montabone and an anonymous referee for their constructive comments that helped to improve this manuscript. Thanks are also due to P Irwin for providing
References


Anderson D L T 1976 The low-level jet as a western boundary current Mon. Weather Rev. 104 907–21

Andrews D G 2001 An Introduction to Atmospheric Physics (Cambridge: Cambridge University Press)


Bagnold R A 1954 The Physics of Blown Sand and Desert Dunes (London: Methuen)


Balme M and Greeley R 2006 Dust devils on Earth and Mars Rev. Geophys. 44 RG3003


Barnes J R 1984 Linear baroclinic instability in the Martian atmosphere J. Atmos. Sci. 41 1536–50


Basu S, Richardson M I and Wilson R J 2004 Simulation of the Martian dust cycle with the GFDL Mars GCM J. Geophys. Res. 109 E11006

Benson J L, Kass D M and Kleinböhl A 2011 Mars north polar hood as observed by the Mars Climate Sounder J. Geophys. Res. 116 E03008

Benson J L, Kass D M, Kleinböhl A, Mc Cleese D J, Schofield J T and Taylor F W 2010 Mars south polar hood as observed by the Mars Climate Sounder J. Geophys. Res. 115 E12015

Bibring J-P et al 2004 Perennial water ice identified in the south polar cap of Mars Nature 428 627–30


Brügler A F C and Murphy J R 1998 Mars’ surface pressure tides and their behavior during global dust storms J. Geophys. Res. 103 8587–601


Byrne S 2009 The polar deposits of Mars Annu. Rev. Earth Planet. Sci. 37 535–60

Byrne S and Ingersoll A P 2003 A sublimation model for Martian south pole ice features Science 299 1051–3

Cantor B A, James P B and Calvin W M 2010 MARCI and MOC observations of the atmosphere and surface cap in the north polar region of Mars Icarus 208 61–81

Castellanos A 2005 The relationship between alternating driver particle forces and bulk behaviour in dry and uncharged fine powders Adv. Phys. 54 263–376

Chapman S and Lindzen R S 1970 Atmospheric Tides (Dordrecht: Reidel)

Charney J G 1947 The dynamics of long waves in a baroclinic westerly current J. Meteor. 4 135–52


Clancy R T, Wolff M J and Christensen P R 2003 Mars aerosol studies with the MGS TES emission phase function observations: Optical depths, particle sizes and ice cloud types versus latitude and solar longitude J. Geophys. Res. 108 5098

Clark B et al 2005 Chemistry and mineralogy of outcrops at Meridiani Planum Earth Planet. Sci. Lett. 240 73–94

Claudin P and Andreotti B 2006 A scaling law for aeolian dunes on Earth and Mars J. Geophys. Res. 111 E04015


Colaprete A and Toon O B 2000 The radiative effects of martian water ice clouds on the local atmospheric temperature profile Icarus 145 524–32

Colaprete A and Toon O B 2002 Carbon dioxide snow storms during the polar night on Mars J. Geophys. Res. 107 10501


Collins M, Lewis S R, Read P L and Houdrín F 1996 Baroclinic wave transitions in the Martian atmosphere Icarus 120 344–57

Conrath B J 1976 Influence of planetary-scale topography on the diurnal thermal tide during the 1971 martian dust storm J. Atmos. Sci. 33 2430–9


Cushing G E and Titus T N 2008 MGS-TES thermal inertia study of the Arria Mons Caldera J. Geophys. Res. 113 E06006


Dick A L 1990 A simple model for air/snow fractionation of aerosol components over the Antarctic peninsula J. Atmos. Chem. 11 179–96

Dickinson R E 1973 Method of parameterization for infrared cooling between altitudes of 30 and 70 km J. Geophys. Res. 78 4451–7


Dundas C M, Diniega S and McEwen A S 2015 Long-term monitoring of martian gully formation and evolution with MRO/HiRISE Icarus 251 244–63

Eady E T 1949 Long waves and cyclone waves Tellus 1 33–52

Eckermann S D, Ma J and Zhu X 2000 Scale-dependent infrared radiative damping rates on Mars and their role in the deposition of gravity-wave momentum flux Icarus 211 429–42


Eluszkiewicz J 1993 On the microphysical state of the Martian seasonal caps Icarus 103 43–8

Fassert C I and Head J W 2008 Valley network-fed, open-basin lakes on Mars: distribution and implications for Noachian surface and subsurface hydrology Icarus 198 37–56


Fenton L K and Michaels T I 2010 Characterizing the sensitivity of daytime turbulent activity on Mars with the MARMS LES: early results Int. J. Mars Sci. Explor. 5 159–71

Findlater J 1969 A major low-level air current near the Indian Ocean during the northern summer Q. J. R. Meteorol. Soc. 95 362–80

Fjørtoft R 1953 On the changes in the spectral distribution of kinetic energy for two-dimensional nondissipative flow Tellus 5 225–30


Forget F, Haberle R M, Montmessin F, Levraud B and Head J W 2006 Formation of glaciers on Mars by atmospheric precipitation at high obliquity Science 311 368–71


Forget F and Pierrehumbert R T 1997 Warming early Mars with carbon dioxide clouds that scatter infrared radiation Science 278 1273–6


Gillette D 1978 A wind tunnel simulation of the erosion of soil: Effect of soil texture, sandblasting, wind speed and soil consolidation on dust production Atmos. Environ. 12 1735–43


Greenwood J P and Blake R E 2006 Evidence for an acidic ocean on Mars from phosphorus geochemistry of Martian soils and rocks Geology 34 953


Güillot T 2010 On the radiative equilibrium of irradiated planetary atmospheres Astron. Astrophys. 520 A27

Guo X, Richardson M I, Soto A and Toigo A 2010 On the mystery of the perennial carbon dioxide cap at the south pole of Mars J. Geophys. Res. 115 E04005

Guzewich S D, Toigo A D, Kulowski L and Wang H 2015 Mars Orbiter Camera climatology of textured dust storms Icarus 258 1–13
Hinson D P et al 2014 Initial results from radio occultation measurements with the Mars Reconnaissance Orbiter: a nocturnal mixed layer in the tropics and comparisons with polar profiles from the Mars Climate Sounder Icarus 243 91–103

Hoffman N 2000 White Mars: a new model for Mars’ surface and atmosphere based on CO2 Icarus 146 326–42

Hofstadter M D and Murray B C 1990 Ice sublimation and rheology: implications for the Martian polar layered deposits Icarus 84 352–61

Hollingsworth J L and Barnes J R 1996 Forced stationary planetary waves in Mars’s winter atmosphere J. Atmos. Sci. 53 428–48


Holt J W et al 2008 Radar sounding evidence for buried glaciers in the southern mid-latitudes of Mars Science 322 1235–8


Head J W et al 2005 Tropical to mid-latitude snow and ice accumulation, flow and glaciation on Mars Nature 434 346–51


Hess S L 1950 Some aspects of the meteorology of Mars J. Meteor. 7 1–13


Hide R 1969 Dynamics of the atmospheres of the major planets with an appendix on the viscous boundary layer at the rigid bounding surface of an electrically-conducting rotating fluid in the presence of a magnetic field J. Atmos. Sci. 26 841–53


Hinson D P 2006 Radio occultation measurements of transient eddies in the northern hemisphere of Mars J. Geophys. Res. 111 E05002

Hinson D P, Pláźold M, Tellmann S, Häusler B and Tyler G L 2008 The depth of the convective boundary layer on Mars Icarus 198 57–66
Matsuo K and Heki K 2009 Seasonal and inter-annual changes of volume density of martian CO₂ snow from time-variable elevation and gravity Icarus 202 90–4
McCleese D J et al 2008 Intense polar temperature inversion in the middle atmosphere on Mars Nat. Geosci. 1 745–9
McCleese D J et al 2010 Structure and dynamics of the Martian lower and middle atmosphere as observed by the Mars climate sounder: seasonal variations in zonal mean temperature, dust and water ice aerosols J. Geophys. Res. 115 E12016
McEwen A S, Dundas C M, Mattson S S, Toigo A D, Ojha L, Wray J J, Chojnicki M, Byrne S, Murchie S L and Thomas N 2014 Recurring slope lineae in equatorial regions of Mars Nat. Geosci. 7 53–8
Miljkovitch M and Head J W 2005 North polar cap of Mars: polar layered deposit characterization and identification of a fundamental climate signal J. Geophys. Res. 110 E101005
Montabone L, Lewis S R and Read P L 2005 Interannual variability of Martian dust storms in assimilation of several years of Mars Global Surveyor observations Adv. Space Res. 36 2146–55
Montabone L et al 2015 Eight-year climatology of dust optical depth on Mars Icarus 251 65–95
Mulholland D P, Read P L and Lewis S R 2013 Simulating the interannual variability of major dust storms on Mars using variable lifting thresholds Icarus 223 344–58
Murray J B et al 2005 Evidence from the Mars Express high resolution stereo camera for a frozen sea close to Mars equator Nature 434 352–6
Nelli S M, Renno N O, Murphy J R, Feldman W C and Bouguer S W 2010 Simulations of atmospheric phenomena at the Phoenix landing site with the Ames general circulation model J. Geophys. Res. 115 E00E21
Newman C E, Lewis S R and Read P L 2005 The atmospheric circulation and dust activity in different orbital epochs on Mars Icarus 174 135–60
Newman C E, Lewis S R, Read P L and Forget F 2002b Modeling the Martian dust cycle, 2, Multianuual radiatively active dust transport simulations J. Geophys. Res. 107 5124
Owen T 1982 The composition of the Martian atmosphere Adv. Space Res. 2 75–80
Pankine A A and Ingersoll A P 2002 Interannual variability of Martian global dust storms: Simulations with a low-order model of the general circulation Icarus 155 299–323
Pedlosky J 1971 Finite-amplitude baroclinic waves with small dissipation J. Atmos. Sci. 28 587–97
Petrosyan A et al 2011 The Martian atmospheric boundary layer Rev. Geophys. 49 RG3005
Phillips R J et al 2011 Massive CO₂ ice deposits sequestered in the south polar layered deposits of Mars Science 332 838–41
Pollack J B, Kasting J F, Richardson S M and Pollakoff K 1987 The case for a wet, warm climate on early Mars Icarus 71 203–24
Raack J, Reiss D, Appéré T, Vincendon M, Ruesch O and Hiesinger H 2015 Present-day seasonal gully activity in a south polar pit (Sisyphi Cavi) on Mars Icarus 251 226–43
Rasmussen K R, Kok J F and Merrison J R 2009 Enhancement in CO₂ and H2O419 Nature 43
Richardson M I and Wilson G 2002b A topographically forced enhancement in CO₂ and dust on present-day solar radiation and climate on Mars Q. J. R. Meteorol. Soc. 125 483–93
Sagan C and Mullen G 1972 Earth and Mars: evolution of atmospheres and surface temperatures Science 177 52–6
Savijärvi H and Siilis T 1993 The Martian slope winds and the nocturnal PBL jet J. Atmos. Sci. 50 77–88
Seiff A 1982 Post-Viking models for the structure of the summer atmosphere of Mars Adv. Space Res. 2 3–17
Smith D E, Zuber M T and Neumann G A 2001a Seasonal variations of snow depth on Mars Science 294 2141–6
Smith D E et al 2001a Mars Orbiter Laser Altimeter: experiment summary after the first year of global mapping of Mars J. Geophys. Res. 106 23689–722
Smith M D 2002 The annual cycle of water vapor on Mars as observed by the thermal emission spectrometer J. Geophys. Res. 107 5115
Spiga A 2011 Elements of comparison between martian and terrestrial mesoscale meteorological phenomena: katabatic winds and boundary layer convection Planet. Space Sci. 59 915–22
Spiga A and Forget F 2009 A new model to simulate the Martian mesoscale and microscale atmospheric circulation: validation and first results J. Geophys. Res. 114 E02009