Chapter 6

The solar cycle as a dynamo

> *If the sun did not have a magnetic field, it would be as boring a star as most astronomers believe it to be.*
>
> Attributed variously to E.N. Parker or R.B. Leighton

The aim of this chapter is to introduce observational aspects of the solar magnetic activity cycle that have most direct relevance as constraints to the dynamo mechanism and models that will occupy us in this third and final part of the course. Once again we turn to the sun as an exemplar of astrophysical magnetohydrodynamics, this time with regards to dynamo action. As with the wind models discussed in part II, this is not because the solar dynamo is more simple or complicated or interesting than other astrophysical dynamos, but simply because it is the dynamo for which we have the most observational information. Even more so than the geodynamo in the Earth's core, in fact, because as we shall see, the dynamo-powered solar magnetic cycle operates on a timescale commensurate with the human lifespan, rather than glacial or geological. By the time we're done, you will hopefully begin to appreciate the fact that the statement cited above is definitely not an understatement!

6.1 The sunspot cycle

6.1.1 Sunspots

Until the beginning of the twentieth century, the story of the solar activity cycle is coincident with the story of **sunspots**. As their name suggest, sunspots look like dark blemishes on the solar disk, but the vast majority are too small to be readily visible without a telescope. Only the largest sunspots can be visible to the naked-eye under suitable viewing conditions, for example when the sun is partially obscured by clouds or mist, particularly at sunrise or sunset. Numerous such sighting exist in the historical records, starting with Theophrastus (374-287 B.C.) in the fourth century B.C. However, by far the most extensive pre-telescopic records are found in the far east, especially in the official records of the Chinese imperial courts, starting in 165 B.C.

Figure 6.1 represents, to the best of our knowledge, the first surviving sunspot drawing, from a sighting on Saturday, 8 December 1128. The drawing is found in the Chronicles of John of Worcester, one of the many monks who contributed to the Worcester Chronicles. The accompanying text translates to something like:

> "...from morning to evening, appeared something like two black circles within the disk of the Sun, the one in the upper part being bigger, the other in the lower part smaller. As shown on the drawing." (trans. A. Van Helden)

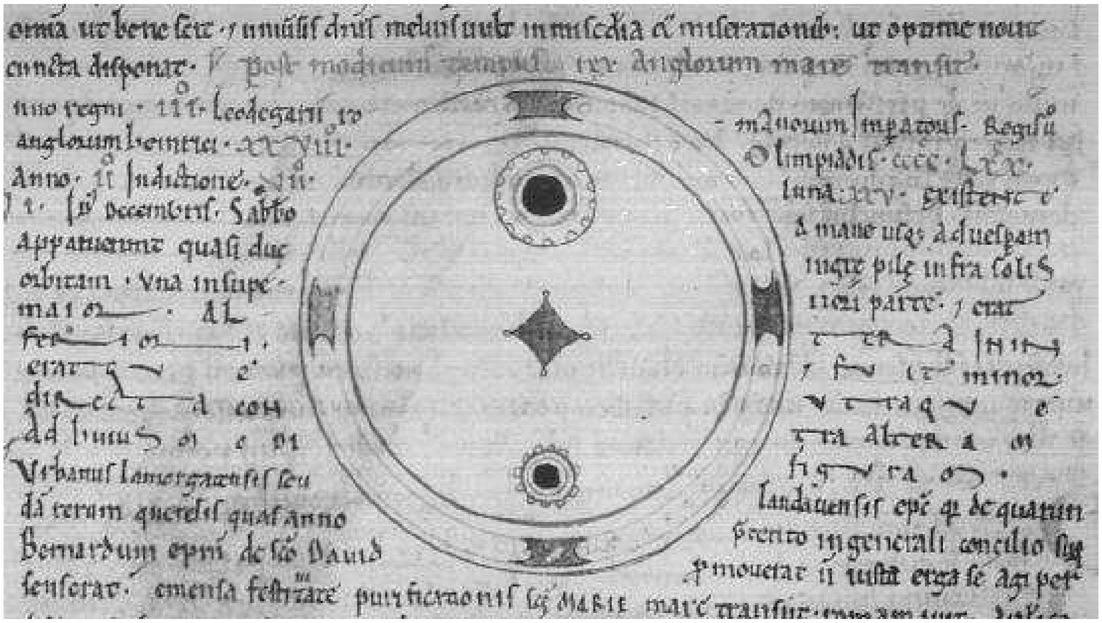

Figure 6.1: Sunspot drawing in the Chronicles of John of Worcester, twelfth century. Notice the depiction of the penumbra around each spot. Reproduced from R.W. Southern, *Medieval Humanism*, Harper & Row 1970, [Plate VII].

The fact that the Worcester monks could apparently distinguish the umbrae and penumbrae of the sunspots they observed suggests that these spots must have been exceptionally large.

A fascinating pre-telescopic sunspot sighting is certainly that of 28 May 1607 by none other than Johannes Kepler (1571-1630). Kepler had been observing the sun for over a month using his *camera obscura* projection technique, basically a pinhole camera. He was hoping to detect a transit of Mercury across the solar disk, as predicted by extant planetary ephemerides, and was well aware of the latter's deficiencies. But on May 28 he did noticed a small black spot on the solar disk, and concluded that he was indeed seeing Mercury in transit (see Fig. 6.2). It did not take long before he came to realize his mistake.

At the end of the first decade of the seventeenth century, four astronomers more or less simultaneously turned the newly invented telescope toward the Sun, and noted the existence of sunspots. They were Johann Goldsmid (1587-1616, a.k.a. Fabricius) in Holland, Thomas Harriot (1560-1621) in England, Galileo Galilei (1564-1642) in Italy, and the Jesuit Christoph Scheiner (1575-1650) in Germany. Fabricius was the first to publish his results in 1611, and to correctly interpret the apparent motion of sunspots in terms of axial rotation of the Sun. Like Harriot, Fabricius and his father (the then-well-known astronomer David Fabricius) first observed sunspots directly through their telescope shortly after sunrise or before sunset. The harrowing account of their observations is worth quoting: (excerpt from the translation in the paper by W.M. Mitchell cited below):

> "... Having adjusted the telescope, we allowed the sun's rays to enter it, at first from the edge only, gradually approaching the center, until our eyes were accustomed to the force of the rays and we could observe the whole body of the sun. We then saw more distinctly and surely the things I have described [sunspots]. Meanwhile clouds interfered, and also the sun hastening to the meridian destroyed our hopes of longer observations; for indeed it was to be feared that an indiscreet examination of a lower sun would cause great injury to the eyes, for even the weaker rays of the setting or rising sun often inflame the eye with a strange redness, which may last for two days, not without affecting the appearance of objects."

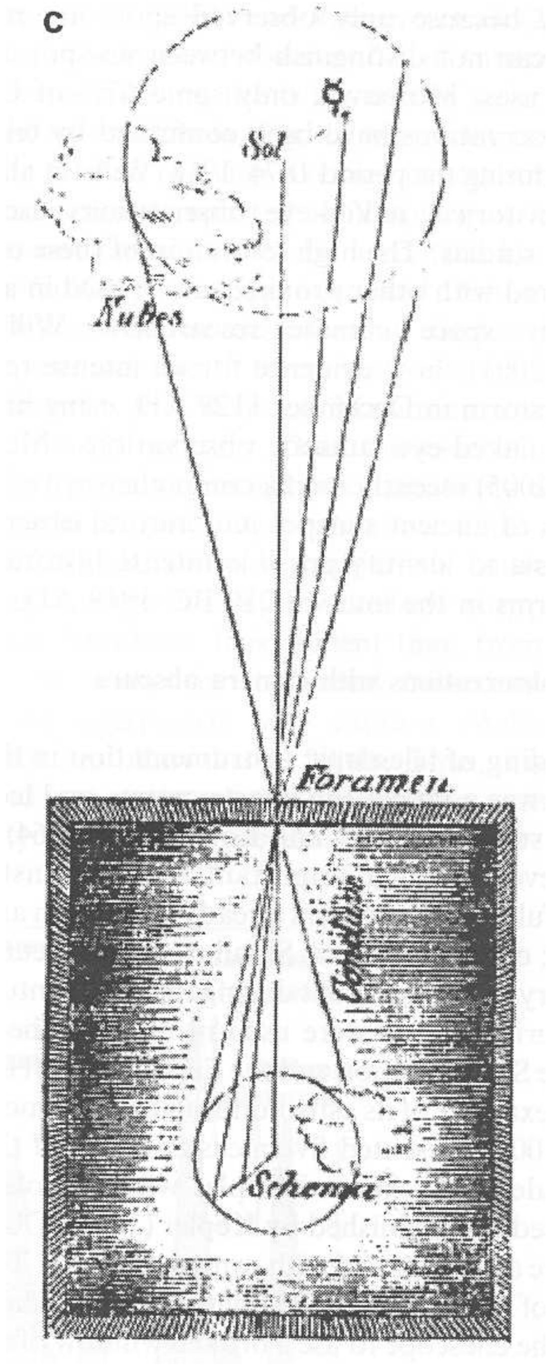

Figure 6.2: Naked-eye observation of a sunspot on 18 May 1607 by Johannes Kepler. Observing the sun intermittently on a cloudy day, Kepler could only make a few observations, and concluded he had had the good fortune of catching the planet Mercury in transit across the solar disk. Diagram reproduced from Vaquero, J.M. 2007, *Adv. Sp. Res.*, **40**, 929 [Fig. 2].

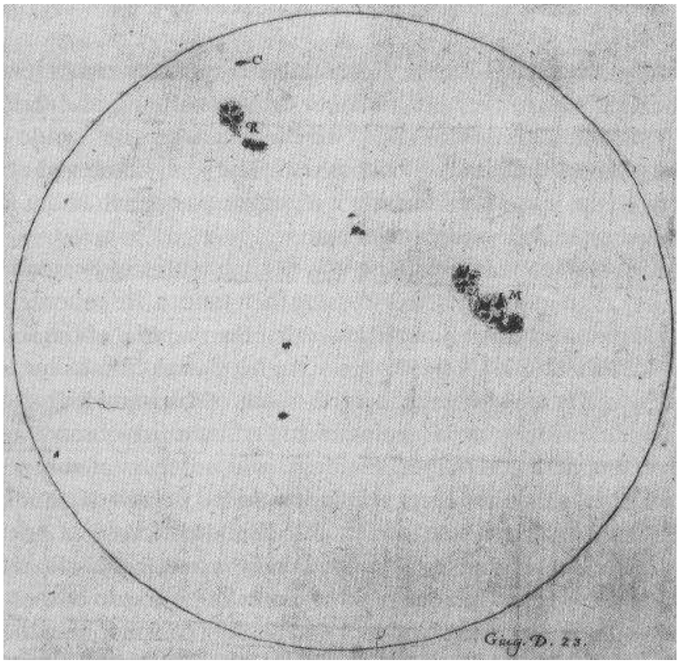

Figure 6.3: Reproduction of one of Galileo's sunspot drawings for 23 June, 1612. The umbrae/penumbrae structure is clearly visible here.

Galileo and Scheiner, however, were the most active in using sunspots to attempt to infer physical properties of the Sun (Figs. 6.3, 6.4). To Galileo belongs the credit of making a convincing case that sunspots are indeed features of the solar surface, as opposed to intra-Mercurial planets (Scheiner's original position).

6.1.2 The eleven-year cycle

Early sunspots observers noted the curious fact that sunspots rarely appear outside of a latitudinal band of about ±30° centered about the solar equator, but otherwise failed to discover any clear pattern in the appearance and disappearance of sunspots. In 1826, the German amateur astronomer Samuel Heinrich Schwabe (1789-1875) set himself about the task of discovering intra-mercurial planets, whose existence had been conjectured for centuries. Like many before him, Schwabe realized that his best chances of detecting such planets lay with the observation of the apparent shadows that they would cast upon crossing the visible solar disk during conjunction; the primary difficulty with this research program was the ever-present danger of confusing such planets with small sunspots. Accordingly, Schwabe began recording very meticulously the position of any sunspot visible on the solar disk on any day that weather would permit solar observation. In 1843, after 17 years of observations, Schwabe had not found a single intra-mercurial planet, but had discovered something else of great importance: the cyclic increase and decrease with time of the *average* number of sunspot visible on the Sun, with a period that Schwabe originally estimated to be 10 years.

As Schwabe's discovery of the sunspot cycle gained recognition, the question immediately arose as to whether the cycle could be traced farther in the past on the basis of extant sunspot observations. In this endeavour the most active researcher was without doubt the Swiss astronomer Rudolf Wolf (1816-1893). Faced with the daunting task of comparing sunspot observations carried out by many different astronomers using various instruments and observing techniques, Wolf defined a **relative sunspot number** (r) as follows:

$$r = k(f + 10g) \ , \tag{6.1}$$

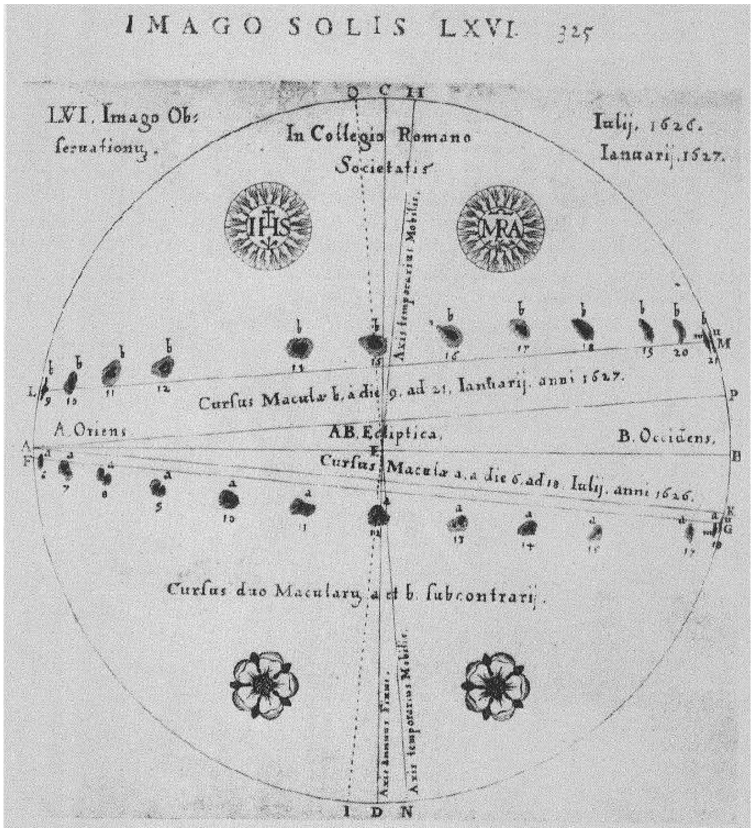
IMAGO SOLIS LXVI. 325
LXVI. Imago Obſervationũ.
In Collegio Romano Societatis
Iulij, 1626. Ianuarij, 1627.
O C H
Axis temporarius Mobilis
Curſus Maculæ b. a die 9. ad 21. Ianuarij. anni 1627.
A Oriens
AB. Ecliptica.
B. Occidens.
Curſus Maculæ a. a die 6. ad 18. Iulij. anni 1626.
Curſus duo Macularũ a et b. ſubcontrarij.
Axis Annuus Fixus
Axis temporarius Mobilis
I D N

Figure 6.4: Solar rotation as inferred from the drift of sunspots. Diagram from Scheiner's *Rosa Ursina*, actually meant to illustrate the variations of the Sun's apparent rotation axis (as seen from Earth) in the course of the year.

where g is the number of sunspots groups visible on the solar disk, f is the number of individual sunspots (including those distinguishable within groups), and k is a correction factor that varies from one observer to another (with $k = 1$ for Wolf's own observations, by definition). This definition is still in used today, but r is now usually called the Wolf (or Zürich) sunspot number. Wolf succeeded in reliably reconstructing the variations in sunspot number as far as the the 1755–1766 cycle, which has has since been known conventionally as "Cycle 1", with all subsequent cycles numbered consecutively thereafter; at this writing (August 2008), we are in the minimal activity phase delineating cycle 23 from the upcoming cycle 24.

Figure 6.5 shows two time series of the relative sunspot number. The first (thin black line) is the monthly-averaged value of r as a function of time, and the thick red line is a 13-month running mean of the same. Note how the amplitude, duration and even shape of sunspots cycles can vary substantially from one cycle to the next. The period, in particular, ranges from 9 (cycle 2) to 14 years (cycle 4), although it remains costumary to speak of the "11-year cycle".

6.1.3 The Waldmaier and Gnevyshev-Ohl Rules

Starting with Wolf himself, the sunspot number time series (monthly, monthly smoothed, yearly, etc) has been analyzed to death in every possible manner known to statistics, nonlinear dynamics, and numerology[1]. Many otherwise serious and respectable people engaged in this type

[1] Two colleagues of mine, both world-renowed experts in the analysis of time series, have independently remarked to me that the sunspot number time series are quite possibly the "natural" time series having produced

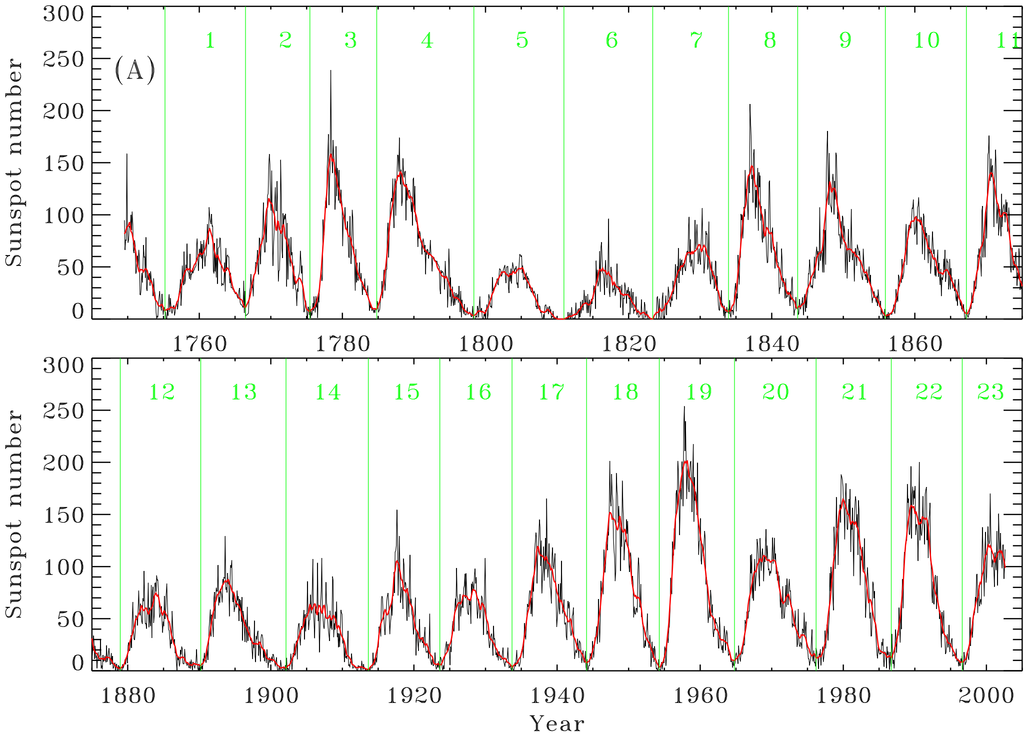

Figure 6.5: Two time series of the celebrated Wolf Sunspot Number. The thin black line is the monthly-averaged sunspot number, and the thick red line a 13-month running mean thereof. These and other related data are publicly available at the Solar Influences Data Analysis Center in Brussels, Belgium (`http://sidc.oma.be`).

of work seem to forget that the definition of the sunspot number is largely arbitrary, and its link to the real dynamical quantity, the sun's magnetic internal field, uncertain at best.

Of the various patterns uncovered in the sunspot number time series, some actually appear to be robust, in that they do not depend too much on the manner the analysis is being carried out, and are also found in other indicators of solar activity; to the point in fact that they have been upgraded to the status of empirical "Rule". We'll consider here only the two most convincing ones.

The Waldmaier rule refers to the fact that an anticorrelation seems to exist between cycle amplitude and rise time (or duration). Starting for example from the time series of smoothed monthly sunspot number (red line on Figure 6.5, it is straightforward to assign to each cycle n a peak amplitude A_n and a duration T_n, the latter being simply the time interval between the two minima bracketing a given cycle. Similarly, the rise time is the time interval between a minimum and the subsequent maximum. Figure 6.6A shows a correlation plot of cycle rise tiem and amplitude, which is characterized by a linear correlation coefficient of $r = -0.7$, which is definitely large enough to merit attention. A similar, through weaker anticorrelation exists between cycle amplitude and duration. These anticorrelations are intriguing, because one might have (naively) expected that high amplitude cycles should take longer to build up and also last longer, but in fact the opposite seems to hold.

Another intriguing sunspot cycle amplitude pattern is known as the Gnevyshev-Ohl rule, and is illustrated on Fig. 6.6B. Cycle peak amplitude A_n are plotted as solid dots, versus cycle number n. For reasons that will become clear shortly, odd-numbered cycles (according to

the largest number of research journal pages per byte of actual data!

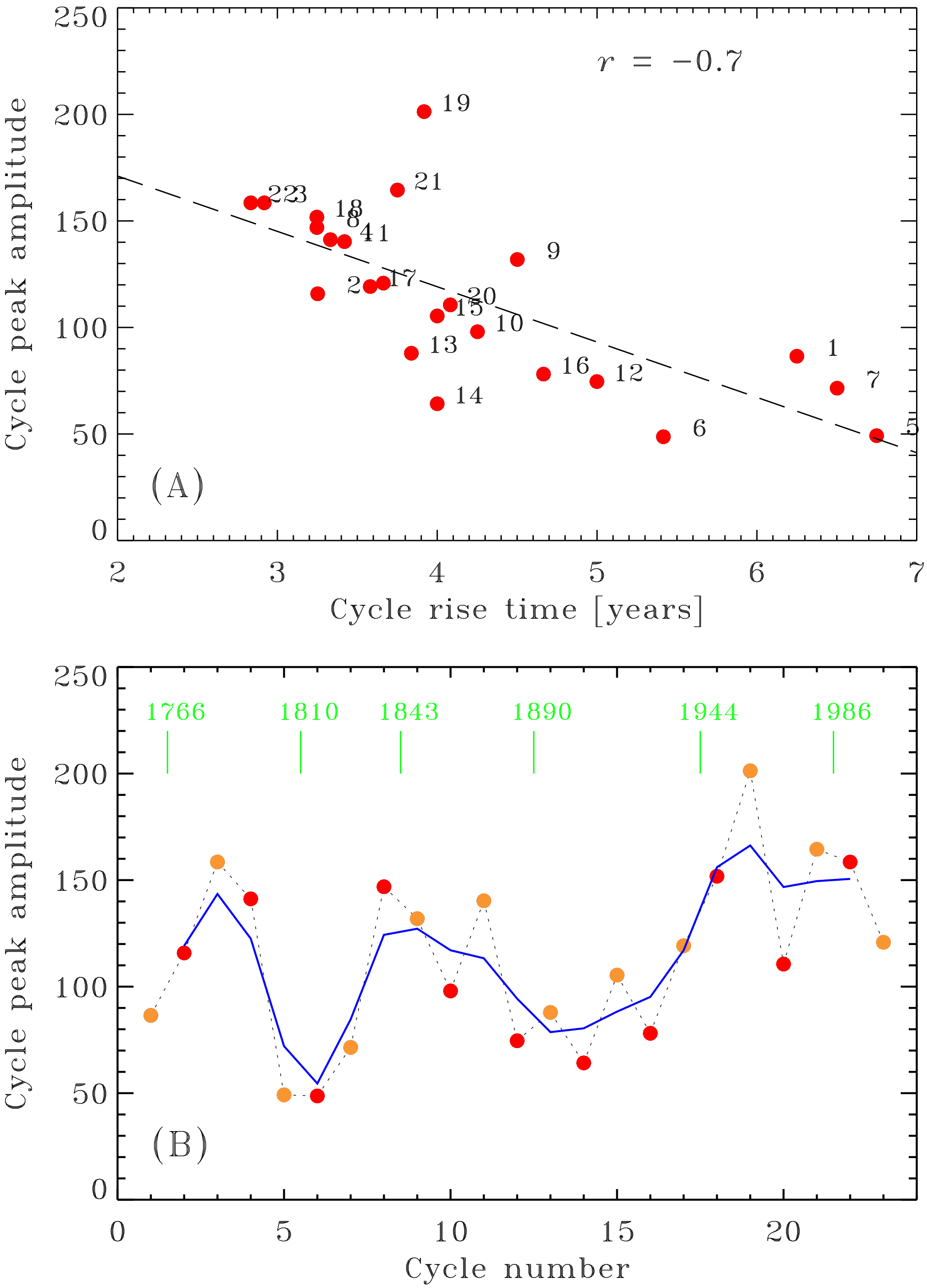

Figure 6.6: (A) The anticorrelation between cycle rise time and amplitude, known as the Waldmaier Rule. A similar correlation, although weaker, characterizes cycle amplitudes and durations; (B) The Gnevyshev-Ohl Rule. Under Wolf's numbering convention, the odd-numbered cycles (orange dots) are more often found above the running mean (blue line) than even-numbered cycles (red dots), a pattern that held true uninterrupted from cycle 9 to 21 inclusively.

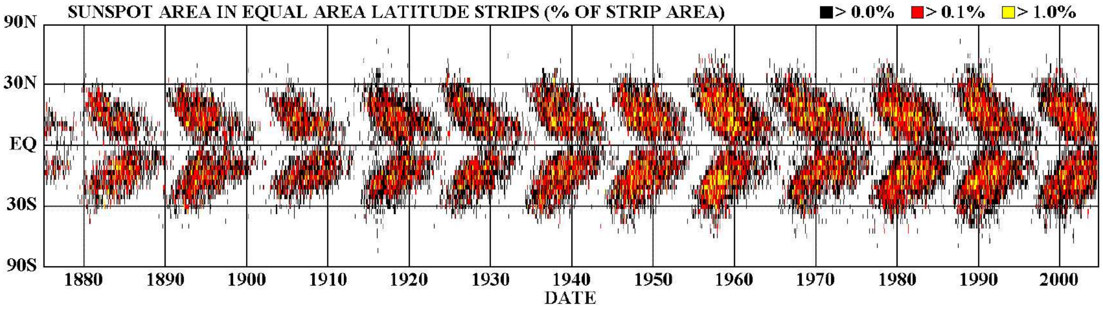

Figure 6.7: A sunspot butterfly diagram, showing the equatorward migration of sunspot latitudes in the course of each cycle. The sunspot number peaks about midway though the equatorward migration Data and graphics courtesy of David Hathaway, NASA/MSFC.

Wolf's numbering convention) have been plotted in orange, and even-numbered cycles in red. Compute now a 1-2-1 running mean of cycle amplitude, i.e.,

$$\langle A_n \rangle = \frac{1}{4}(A_{n-1} + 2A_n + A_{n+1}) , \qquad n = 2, 3, ... \tag{6.2}$$

The resulting time series for $< A_n >$ is plotted as a thick blue line on Fig. 6.6B; notice now how most odd-numbered cycles (orange) lie *above* the running mean curve, while even-numbered cycles (red) usually lie below. In fact, from cycle 9 to 21 inclusive, the pattern has held true without interruption. As we will see in due time, both the Waldmaier and Gnevyshev-Ohl Rules pose quite a challenge to most current solar dynamo models.

6.1.4 The butterfly diagram

To the striking cyclic pattern uncovered by Schwabe was soon added an equally striking *spatial* regularity. In 1858, Gustav Spörer (1822-1895) and Richard Carrington (1826-1875) independently pointed out that sunspots are observed at relatively high ($\sim 40°$) heliocentric latitudes at the beginning of a sunspot cycle, but are seen at lower and lower latitudes as the cycle proceeds, until at the end of the cycle they are seen mostly near the equator, at which time spots announcing the onset of the next cycle begin to appear again at $\sim 40°$ latitude. This is illustrated on Figure 6.7, in the form of a **butterfly diagram** for the time period 1875—2003. The construction of sunspot butterfly diagrams was first carried out by the husband-and-wife team of Annie and E. Walter Maunder in 1904, and proceeds as follows: one begins by laying a coordinate grid on, for example, a solar white light or calcium image, with, as in the case of geographic coordinates on Earth, the rotation axis defining the North—South direction. The visible solar disk is then divided in latitudinal strips of constant projected area, and for each such strip the percentage of the area covered by sunspots and/or active regions is calculated and color coded. This defines a one-dimensional (vertical) array describing the average sunspot coverage at one time. By repeating this procedure at constant time intervals and stacking the arrays one besides the other, one obtains a two-dimensional image of average sunspot coverage as a function of heliospheric latitude (vertical axis) and time (horizontal axis).

The absence of sunspots at high latitudes ($\gtrsim 40°$) at any time during the cycle, and the equatorward drift of the sunspot distribution as the cycle proceeds from maximum to minimum, are both particularly striking on such a diagram. Note how the latitudinal distribution of sunspots is never exactly the same, and how for certain cycles (for example the 1965—1976 cycle) there exists a pronounced North–South asymmetry in the hemispheric distributions. Note also how, at solar minima, spots from each new cycle begin to appear at mid-latitudes while spots from the preceding cycle can still be seen near the equator, and how sunspots are almost never observed within a few degrees in latitude of the equator. Sunspot maximum (1991, 1980, 1969,...) occurs about midway along each butterfly, when sunspot coverage is maximal at about 15 degrees latitude.

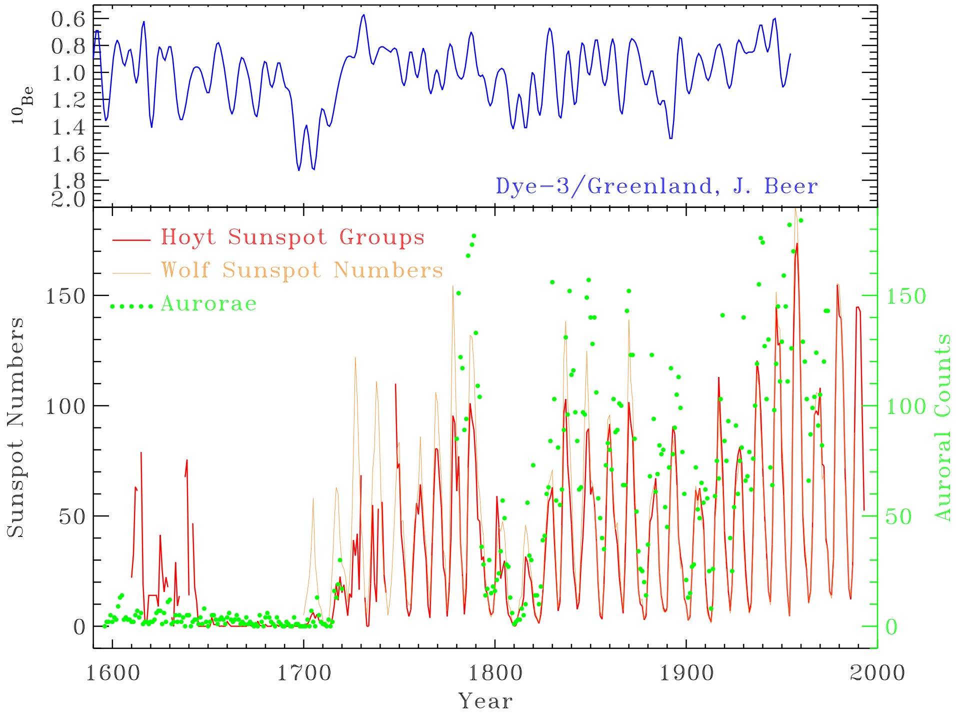

Figure 6.8: The Maunder minimum, as seen through cosmogenic radioisotopes (top panel) and sunspot and auroral counts (bottom panel). The thick red line is the so-called Group Sunspot Number, a reconstruction similar to Wolf's (thin orange line) but deemed more reliable in the eighteenth century because it relies exclusively on the more easily observable sunspot groups. Beryllium 10 data courtesy of J. Beer, EAWAG/Zürich.

6.1.5 The Maunder Minimum

One final, peculiar feature associated with the sunspot cycle needs to be discussed, because of its implications for dynamo modelling. The historical reconstructions began by Wolf have been pushed as far back as the invention of the telescope in the opening decade of the seventeenth century, which marks the beginning of regular sunspot monitoring by astronomers. One such full reconstruction, starting in 1610, is shown on Figure 6.8 (bottom panel). While observations are a tad patchy from 1610 to 1640, coverage is actually quite good beyond this date. The lack of sunspots in the period 1645-1715 is therefore not due to lack of data, but represents a phase of strongly suppressed solar activity now known as the **Maunder Minimum**, after the solar astronomer E.W. Maunder, who, following the pioneering historical investigations of Gustav Spörer, was most active and steadfast in investigating the dearth of sunspot sightings by astronomers active in the second half of the seventeenth century. The documented occurrence of exceptionally cold winters throughout Europe during those years may be causally related to reduced solar activity, although this remains a topic of controversy.

That this is not just a matter of failing to form sunspots is confirmed by historical reconstructions of auroral counts, which are also strongly reduced during the Maunder Minimum (cf. Fig. 6.8). On the other hand, cosmogenic radioisotopes such as ^{10}Be, whose production frequency is known to be modulated by the frequency of solar eruptive phenomena, continue to show a cyclic pattern throughout the Maunder minimum (Fig. 6.8, top panel), indicating that the cycle had actually not come to a complete standstill.

The cosmogenic isotope record also indicates that episodes of markedly reduced solar activity occurred in 1282-1342 (Wolf minimum) and 1416-1534 (Spörer minimum), and that solar activity was significantly above its modern average in the time period 1100-1250 (dubbed Me-

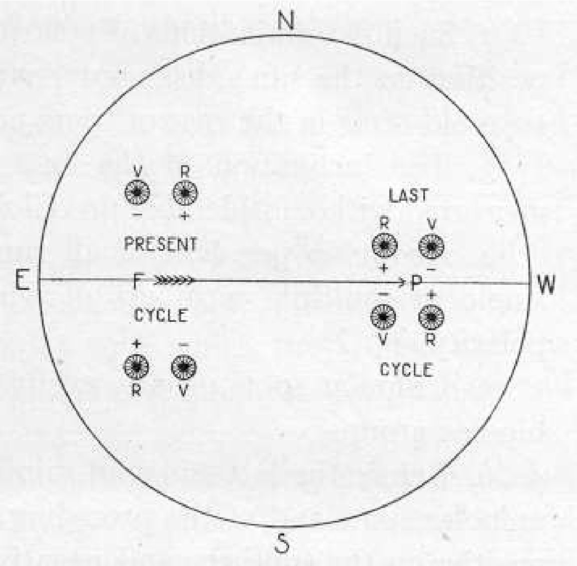

Figure 6.9: A diagram taken from the 1919 paper by G.E. Hale, F. Ellerman, S.B. Nicholson, and A.H. Joy (in *The Astrophysical Journal*, vol. **49**, pps. 153-178), illustrating Hale's polarity laws "V" and "R" refer to the circular polarization components, which allow to determine magnetic polarity. This presented solid evidence for the existence of a well-organized large-scale magnetic field in the solar interior, oriented predominantly in the East-West direction, which cyclically changes polarity approximately every 11 years.

dieval Maximum by Min/Max aficionados). Some recent such reconstructions (see bibliography) have in fact identified some XXX grand minima in the past 11,000 years.

6.2 The (magnetic) solar cycle

6.2.1 Hale's polarity laws

The study of sunspots and their 11-year cycle was finally put on a firm physical footing by George Ellery Hale (1868-1938). In the decade following their groundbreaking discovery of sunspot magnetic fields in 1907-1908, Hale and collaborators went on to establish what are now known as **Hale's polarity laws**:

1. At any given time, the polarities of the leading spots of sunspot pairs are the same in a given solar hemisphere;

2. At any given time, the polarities of the leading spots of sunspot pairs are opposite in the N and S hemispheres;

3. Sunspot polarities reverse in each hemisphere from one 11-yr sunspot cycle to the next;

(see Fig. 6.9). The most straightforward interpretation of this common opposite polarity grouping is that we are seeing the surface manifestation of a large-scale **toroidal field** residing somewhere below the photosphere, having risen upwards and pierced the photosphere in the form of a so-called "Ω–loop" (see Figure 6.10).

Because the flux rope can be expected to expand as it rises buoyantly through the convective envelope, neither the size or magnetic field strength of sunspots can be assumed to be identical

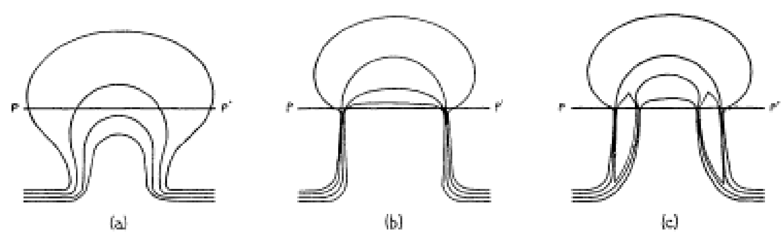

Figure 6.10: Schematic representation of a sunspot pair as the manifestation of an underlying toroidal flux rope having risen through the photosphere. The magnetic fields impedes convective energy transport, so that cooling leads to a collapse of the magnetic field into two sunspots of opposite polarities. Reproduced from the 1955 paper by E.N. Parker's in *The Astrophysical Journal*, vol. **121**, pps. 491-507 [Figure 2, p. 496].

to that of the underlying toroidal flux ropes. However, if the rope maintains its cohesion throughout the rise and emergence processes then its **magnetic flux**

$$\Phi_B = \int (B_\phi \hat{\mathbf{e}}_\phi) \cdot \hat{\mathbf{n}} \mathrm{d}S \tag{6.3}$$

is a conserved quantity, as per flux-freezing in the ideal MHD limit. Observations indicate that for sunspots $\Phi_B \sim 10^{13}$—10^{15} Wb (Wb≡T m^2), with 10^{14} Wb a representative value for a "typical" sunspot.

Hale's polarity rules, interpreted in terms of this "model" of sunspots, imply that the toroidal component of the solar internal magnetic field is antisymmetric about the equator[2], and evolves cyclically on a $\simeq 22$ year timescale. So, from a physical —rather than botanical— standpoint, the true length of the solar cycle is not 11 years, but rather 22 years. Yet astronomers are creatures of tradition, and solar astronomers are no exception; a century after Hale's discovery of the sunspot polarity law, it remains customary to speak of the "11 year solar cycle".

6.2.2 Joy's law

Hale and collaborators also showed that the line segment joining two members of a sunspot pair shows a systematic tilt angle ($\tilde{\theta}$) with respect to the East-West direction, the sunspot farther ahead (in the direction of solar rotation) being closer to the equator. Although there exists considerable variations in observed tilt angles, statistically the magnitude of the tilt increases with increasing heliocentric latitude (see Fig. 6.11). This is known as Joy's Law, after the poor bastard who was told by his boss G.E. Hale to go back and measure the tilts of all sunspot pairs to be found on the sunspot drawings of Carrington and Spörer.

Least-squares fits to observations yield a parametric representation of the form:

$$\sin\tilde{\theta} = 0.48\cos(\theta) + 0.03 \tag{6.4}$$

where θ is the usual polar angle (sometimes called "colatitude"). This pattern plays an important role in some of the solar cycle models to be considered in later chapters. This is because the existence of a finite, systematic tilt implies a net dipole moment, which can contribute to the net solar **poloidal field**.

All these regularities carry a very important message; no matter how vigorous convective fluid motions may be in the convective envelope, they are not vigorous enough to completely disrupt the solar internal toroidal magnetic field.

[2] which, as we shall see in chapter 7 below, is what one would expect from the kinematic shearing of a dipolar magnetic field by axisymmetric differential rotation.

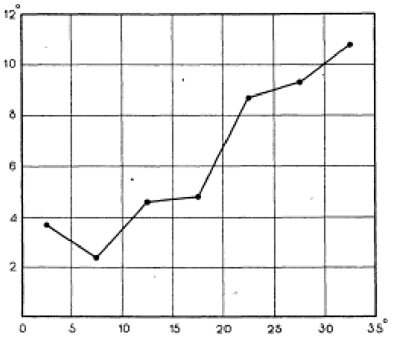

Figure 6.11: Variation of the average tilt angle $\tilde{\theta}$ (ordinate) with respect to heliocentric latitude (abcissa, both in degrees). Reproduced from Hale, Ellerman, Nicholson & Joy 1919, *Astrophys. J.*, **49**, 153 [Figure 5, p. 168].

6.2.3 Modeling the buoyant rise of magnetic flux ropes

In translating the cartoon of Figure 6.10 into a quantitative physical model, we have a number of issues that need to be clarified. Perhaps the most pressing are: (1) to identify the region(s) of the solar interior from which the flux ropes originate, and (2) to estimate the time required for a magnetic flux tube to rise through the convective envelope[3]. It turns out that both questions are very much related.

Consider a toroidal flux tube of diameter a and mean field strength B, embedded in a convective envelope with pressure and density profiles $p(r), \rho(r)$ and scale height $h = kT/\mu m_p g$, where g is the gravitational acceleration and $a \ll h$. Lateral pressure equilibrium demands that

$$p(r) = p_i(r) + \frac{B^2}{2\mu_0} \ , \tag{6.5}$$

where p_i is the gas pressure within the tube, and the second term on the RHS is the **magnetic pressure**. Clearly eq. (6.5) can only be satisfied provided that $p_i < p(r)$, i.e., the flux tube is evacuated. Given the high thermal diffusivity provided by radiation under solar interior conditions, it is reasonable to assume that the temperature is the same inside and outside the tube; this implies $\rho/\rho_i = p_i/p$, so that

$$\rho(r) - \rho_i(r) = \frac{\rho(r)B^2(r)}{2\mu_0 p(r)} \ . \tag{6.6}$$

The (radial) buoyancy force per unit length along the tube is then

$$F = \pi a^2 g(\rho - \rho_i) \ . \tag{6.7}$$

As a consequence of this buoyancy, the tube is accelerated upwards and begins to rise towards the surface. If thermal equilibrium is maintained between the tube and its surroundings, the only force left to equilibrate the buoyancy force is the aerodynamic drag:

$$F_D = \frac{C_D}{2}\rho u^2 a \ , \tag{6.8}$$

where u is the rise velocity of the tube, and the coefficient of aerodynamic drag C_D is a number of order unity for low viscosity subsonic flows. Equating eqs. (6.7) and (6.8) yields, after making use of the definition for the scale height h, the following expression for the terminal rise velocity:

$$u^2 = \frac{B^2}{\mu_0 \rho}\left(\frac{\pi a}{C_D h}\right) \ . \tag{6.9}$$

[3] What follows is largely inspired from the 1975 paper by E.N. Parker cited in the bibliography at the end of this chapter

The rise time τ for a magnetic flux tube starting at a depth r_0 within the envelope is then approximately

$$\tau \simeq \frac{R_\odot - r_0}{u} \ . \tag{6.10}$$

If you start plugging in numbers in the above expressions (you get to do just that in problem 6.3 below!), you soon come to the conclusion that for flux ropes of strengths $\gtrsim 0.1\,$T and magnetic fluxes $\gtrsim 10^{13}\,$Wb released at $r/R_\odot = 0.8$, the rise time is well under one year. More elaborate calculations, taking into account heat exchange between the tube and its surroundings as well as the slightly superadiabatic stratification of the envelope yield even shorter rise times. As we shall see in later chapter, amplification of the solar magnetic field by the dynamo process requires that the field remains in its generating region for a few years before sufficiently high field strengths are produced. This has led to the conclusion that the solar magnetic field is stored —maybe even produced— not in the convective envelope proper, but rather immediately below it.

The issue of *storage* of the magnetic flux ropes below the core-envelope interface is far from trivial. Basically, the flux ropes are subject to non-axisymmetric instabilities that take the form of growing waves of low azimuthal wavenumbers. The growth time for the instability turns out to depend rather sensitively on the thermodynamic structure (namely, the degree of subadiabaticity) of the storage layers. The bibliography at the end of this chapter lists a few good papers concerned with such stability analyses. Such calculations indicate that magnetic flux tubes of strength 6—16 T can be stored immediately beneath the core-envelope interface for time periods of a few years, with the growth time for the instability decreasing rapidly with increasing field strength.

Considerable efforts have also gone into making more realistic models of the rise of thin magnetic flux tubes through the solar convective envelope. Such models include all kinds of reasonable things like rotation, non-axisymmetric perturbations, storage below the core-envelope interface, etc. While this represents a considerable improvement, mathematically flux tubes are still treated as structureless, flux-carrying material lines and so these kinds of calculations cannot properly take into account the interaction of the tube with the surrounding turbulent fluid motions. With this caveat in mind, thin flux tube modeling has produced the following two important results:

1. The flux ropes rise essentially radially if they have a field strength $B \gtrsim 6$—$10\,$T; otherwise the Coriolis force deflects the rising flux tubes to high latitudes.

2. The flux ropes emerge without any tilt for $B \gtrsim 10^2\,$T, and with tilts compatible with Joy's Law for fields strengths in the range 6—16 T.

Now, this is *great* stuff: the observed emergence of sunspots at low heliocentric latitudes puts a *lower* limit on the strength of the participating flux ropes; Joy's Law, on the other hand, translate into an *upper* limit on the field strength. One concludes that the sunspot-forming toroidal flux ropes must have magnetic field strengths in the rather narrow range

$$6 \lesssim B \lesssim 16\,\mathrm{T} \ . \tag{6.11}$$

The basic physical mechanism underlying these two remarkable results is the same: if the rise time of the flux ropes is much smaller than the solar rotation period, the Coriolis force has a strong influence. It is the Coriolis force that deflects the rising flux ropes to high latitudes, and gives rise to the twist that, upon emergence, manifests itself as Joy's Law. If the field is strong enough for the rise time to be much shorter than the rotation period, then the rising flux rope does not "feel" the rotation, rises radially, and emerges without a tilt.

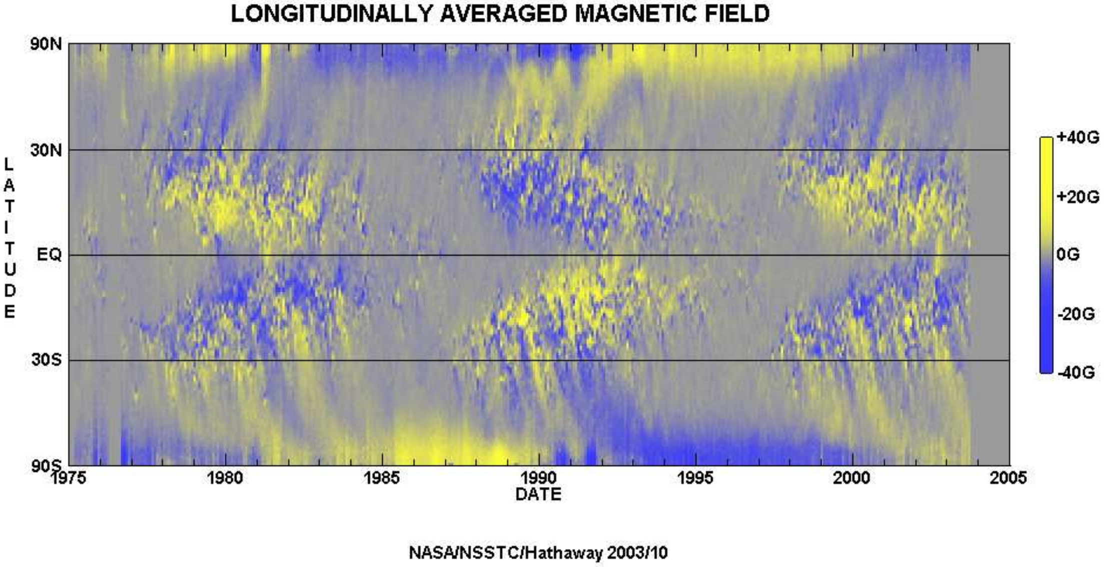

Figure 6.12: A synoptic magnetogram covering the last three sunspot cycles. The radial component is azimuthally averaged over a solar rotation, and the resulting latitudinal strips stacked one against the other for successive rotations. Data and graphics courtesy of David Hathaway, NASA/MSFC.

6.2.4 Poloidal field reversals

While the surface magnetic field on Fig. 2.4 may look like a total mess, on long timescales a well-defined spatiotemporal pattern once again emerges. Figure 6.12 is once again a synoptic (time-latitude) diagram of the radial magnetic field component (averaged in longitude on the visible disk) covering three sunspot cycles. New poloidal field first shows up at mid-latitudes (e.g., 1977), a year or two after the new cycle sunspots have begun to appear at high latitudes, and then migrates to higher and possibly lower latitudes in the course of the cycle. The situation is greatly complicated by the active region fields, which make a very strong contribution to the line-of-sight magnetograms at low heliocentric latitudes. Furthermore, the tilt of active regions amounts to a net dipole moment, which is carried to higher latitudes by the poleward surface meridional flow (more on this later) following the decay of the active regions. This poleward transport is clearly visible on Fig. 6.12, in the form of elongated, inclined stripes extending from mid to high latitudes. Whether this transport of poloidal field contributes to —or even dominates— the evolution of the high latitude poloidal field remains an open question.

At high heliocentric latitude ($\gtrsim 50°$) there exist a cleaner pattern of polarity changes occurring on the solar cycle period. For example, polarity reversal occurs in 1980, at solar maximum. During the 1976—1986 cycle the toroidal field was negative in the N-hemisphere; taken at face value, Figure 6.12 then indicates that the high latitude poloidal field lags the toroidal field by a phase interval $\Delta\varphi \simeq \pi/2$.

A different observational tracer that yields similar results is the count of **polar faculae**, concentrated regions of relatively strong magnetic field often seen at high latitudes. Under the assumption that the structure of the faculae themselves is independent of the phase in the solar cycle, their number at any given time gives a measure of the overall poloidal field strength, once calibrated against magnetograms. Reliable polar faculae records exists for nearly one hundred years, allowing to reconstruct the solar cycle evolution of the large-scale solar poloidal field back to the beginning of the twentieth century.

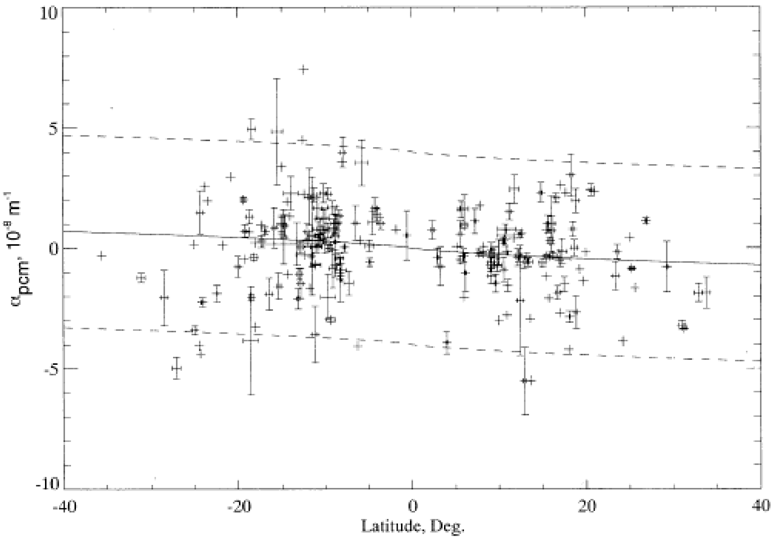

Figure 6.13: Current helicity of active regions versus heliospheric latitude, as determined from 3D vector magnetogram. There is a weak but statistically significant trend whereby current helicity is negative in the Northern solar hemisphere, and positive in the Southern. The solid line is the mean twist expected from the interaction of an initially untwisted flux rope interacting with helical convection, with the dashed dashed lines corresponding to ± one standard deviation. Diagram taken from Canfield & Pevtsov 2000, *J. Astrophys. Astr.*. **21**, 213 [Fig. 2, p. 216].

6.2.5 Current helicity in active regions

Vector magnetographic observations of sunspots and active regions allow to reconstruct the vertical component of the electrical current density, which then allows a calculation of the current helicity ($\mathcal{H}_J$; see §1.11) of the magnetic flux rope having formed the sunspots. Since this is expected to be closely related to the true magnetic helicity, and that the latter is a conserved quantity in ideal MHD, one is allowed to suppose that current helicity patterns characterizing sunspots and active regions are representative of the current helicity of the underlying, large-scale sunspot-forming toroidal magnetic field. An important caveat is the possibility that the helicity of the buoyantly rising flux ropes is altered during its crossing of the convection zone, due to interaction with turbulent fluid motions, themselves imbued with kineric helicity by the Coriolis force. Simple models of this interaction suggest that the bulk of the inferred helicity may be due to this interaction, as shown on Figure 6.13. What is plotted there is in fact the parameter α for a force-free magnetic field with non-zero currents (cf. eq. (1.116), defined via the relation

$$[\nabla \times \mathbf{B}]_z = \alpha B_z = \mu_0 J_z \ , \tag{6.12}$$

the second equality arising from Ampère's Law in its MHD-usual pre-Maxwellian form.

It turns out that inferred current helicities for active regions show a fairly well-defined hemispheric pattern: $\mathcal{H}_J$ is negative in the Northern solar hemisphere, and positive in the Southern (see again Figure 6.13). And since the cycle-to-cycle magnetic polarity reversals flip both $\mathbf{B}$ and $\mathbf{J}$, this sign pattern remains fixed from one cycle to the next. A similar result is obtained by inferring the sign of active region twists from X-Ray observations of magnetic loops linking the two components of classical bipolar sunspot pairs. So the helicity trend is likely real, the question remaining is whether it truly reflects the state of the magnetic field in the deep-seated flux-rope forming region.

6.2.6 Cyclic modulation of solar activity

The solar magnetic cycle also modulates the solar luminosity, the sun being about 0.1% brighter at sunspot maximum than at minimum (see Fig. 6.14, top panel). The emission of short-wavelength, non-thermal emission also varies in phase with the solar cycle; in the far-ultraviolet regions of the solar spectrum ($\lambda \lesssim 120\,\mathrm{nm}$), variations by $\sim 100\%$ are observed between solar minimum and maximum, with corresponding variations by over a factor of ten in the X-Ray domain. The sun's radio emission, indicative non-thermal acceleration of electrons in the lower corona, also follows the sunspot cycle quite closely (see Fig. 6.14, third panel). Finally, the frequency of all solar eruptive events (flares, coronal mass ejections, eruptive filaments, etc.) are all strongly modulated by the solar cycle, although the distribution of event sizes appears to be invariant with respect to the phase of the cycle; put differently, it's not that large flares cannot happen near activity miminum, they are just a lot less frequent. Indeed, the Fall of 2007, far into the descending phase of cycle 23, witnessed some of the largest X-Ray flares on record since the beginning of continuous solar X-Ray monitoring by the GOES patrol satellites.

6.3 Summary of solar cycle characteristics

For convenience, let's now collect a short list of fundamental observational features that a physical model of the solar magnetic cycle should reproduce (omitting for the time being anything related to amplitude fluctuation):

1. A large-scale magnetic field, axisymmetric to a good approximation and antisymmetric about the solar equator;

2. A cyclic variation of this large-scale magnetic field, characterized by polarity reversals with a $\sim 20\,\mathrm{yr}$ oscillation period;

3. An internal toroidal field of strength ~ 1—$10\,\mathrm{T}$, concentrated at low solar latitudes ($\lesssim 45°$, say), and migrating equatorward in the course of the cycle with minimal spatiotemporal overlap between successive cycles;

4. A large-scale surface poloidal field of a few $10^{-3}\,\mathrm{T}$, migrating poleward in the course of the cycle, and reversing polarity at sunspot maximum.

5. Negative current helicity in the N-hemisphere, and positive helicity in the S-hemisphere, across all cycles.

This last constraint remains subject to the caveat described in §6.2.5, and for that reason, although legitimate, it is not considered a particularly strong constraint on dynamo models of the solar cycle.

6.4 A simple dynamo

Before moving on with astrophysical dynamos, we will first consider the following simple example, which illustrates nicely how the idea of amplyfying magnetic field by bodily moving electrical charges across a magnetic field is not so mysterious as one may initially think.

One of the many practical invention of Michael Faraday was a DC electric current generator based on the rotation of a conducting metallic disk threaded by an external magnetic field. Figure 6.15(A) illustrates the basic design: a circular disk of radius a mounted on an axle, rotating at angular velocity ω through the agency of some external mechanical force (like Faraday turning a crank). A vertical magnetic field is imposed across the disk. Electrical charges in the disk will feel the usual Lorentz force $\mathbf{F} = q\mathbf{u} \times \mathbf{B}$ where, (initially) $\mathbf{u}$ is just the

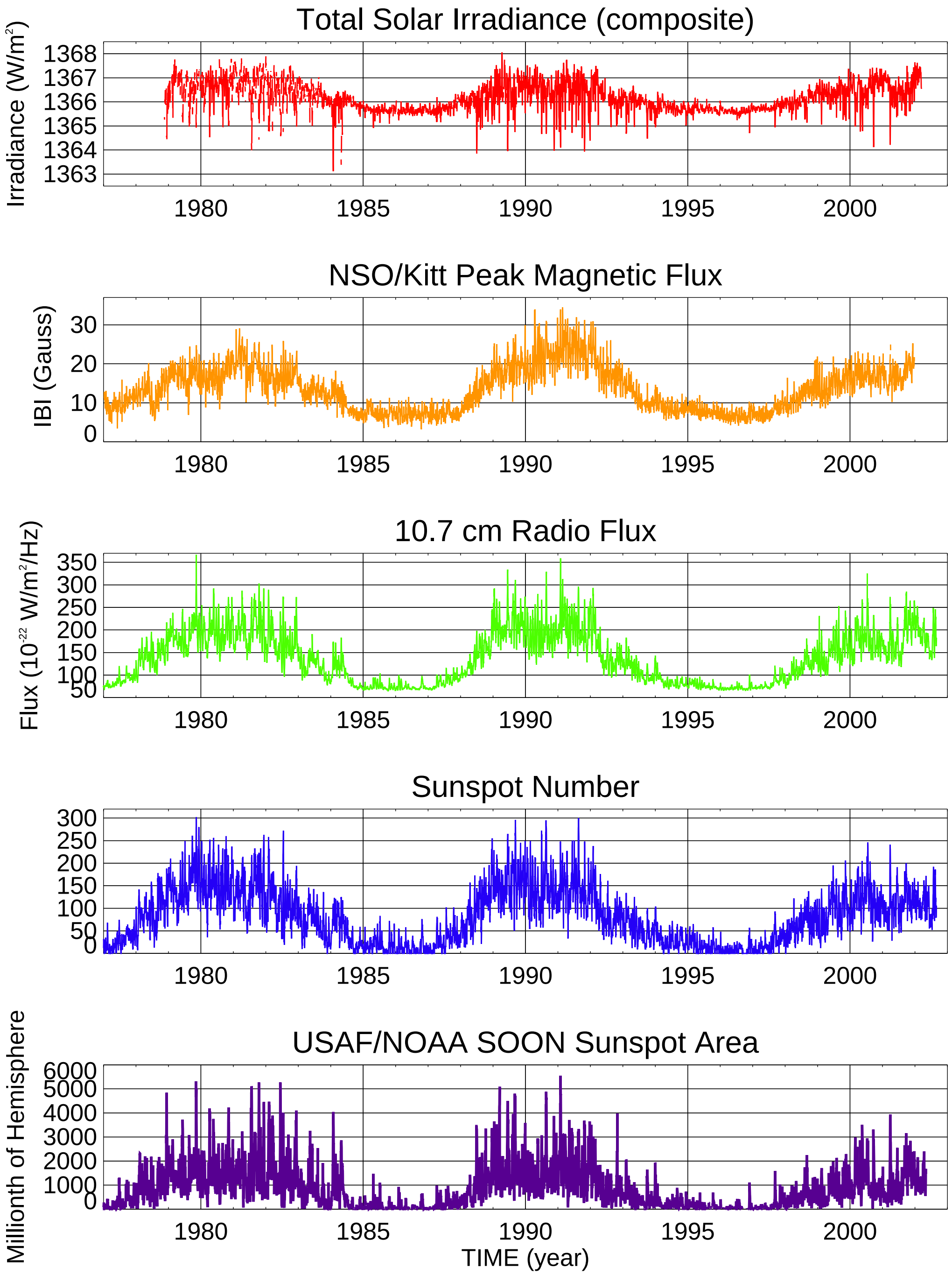

Figure 6.14: Variation of various solar activity indicators with the solar cycle. The NSO/Kitt Peak magnetic flux includes the contribution of magnetic fields outside of sunspots. The 10.7cm radio flux is a measure of non-thermal processes in the lower corona. Data and graphics courtesy of Giuliana DeToma, HAO/NCAR.

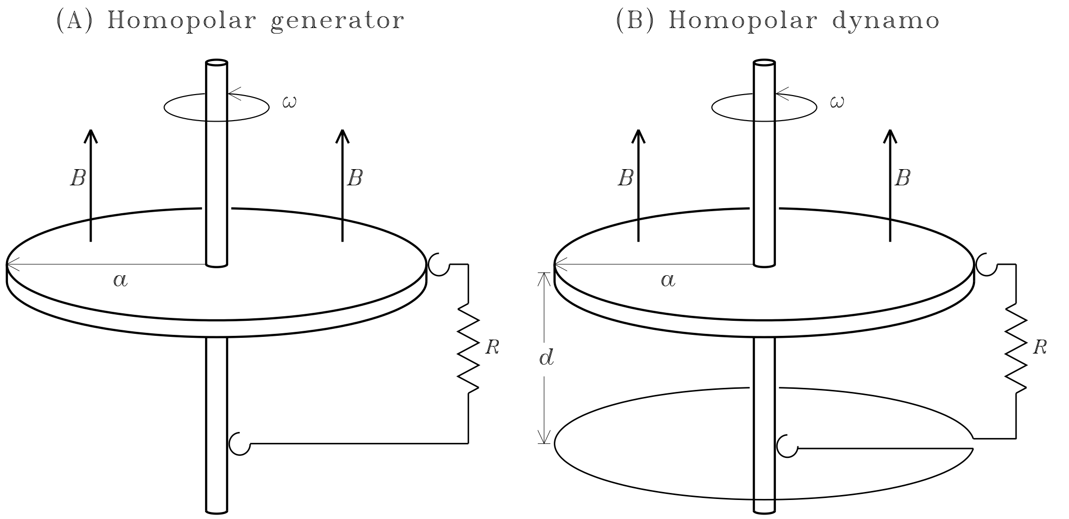

Figure 6.15: A homopolar generator (A) versus a homopolar dynamo (B). An external magnetic field B is applied across a rotating conducting disk, producing an electromotive force that drives a radial current, a wire connecting the edge of the disk to the axle, forming a circuit of resistance R. The only difference between the two electro-mechanical devices illustrated here is that in the latter case, the wire completing the circuit by connecting on the axle is wrapped into a loop in a plane parallel to the disk, so that a secondary vertical magnetic field is produced (see text).

motion imposed by the rotation of the disk. Working in cylindrical coordinates (s, ϕ, z) one can write

$$\mathbf{u} = (\omega s)\hat{\mathbf{e}}_\phi \ , \tag{6.13}$$

$$\mathbf{B} = B_0\hat{\mathbf{e}}_z \ . \tag{6.14}$$

so that

$$\mathbf{F} = (q\omega s B_0)\hat{\mathbf{e}}_s \ . \tag{6.15}$$

Now consider the circuit formed by connecting the edge of the disk to the base of the axle via frictionless sliding contacts. With the lower part of the circuit away from the imposed magnetic field, the only portion of the circuit where the magnetic force acts on the charges is within the disk, amounting to an electromotive force

$$\mathcal{E} = \oint_{\text{circuit}} \left(\frac{\mathbf{F}}{q}\right) \cdot \mathrm{d}\boldsymbol{\ell} = \int_0^a \omega B_0 s \mathrm{d}s = \frac{\omega B_0 a^2}{2} \ . \tag{6.16}$$

Neglecting for the time being the self-inductance of the circuit, the current flowing through the resistor is simply given by $I = \mathcal{E}/R$. This device is called a **homopolar generator**.

There is a subtle modification to this setup that can turn this generator into a **homopolar dynamo**, namely a device that converts mechanical energy into self-amplifying electrical currents and magnetic fields. Instead of simply connecting the resistor straight to the axle as on 6.15(A), the wire is wrapped around the axle in a loop lying in a plane parallel to the disk, and then connected to the axle, as shown on 6.15(B). Use your right-hand rule to convince yourself that this current loop will now produce a secondary magnetic field B_* that will superpose itself on the external field B_0. The magnetic flux through the disk associated with this secondary field will be proportional to the current flowing in the wire loop, the proportionality constant being defined as the inductance (M):

$$MI = \Phi = \pi a^2 B_* \ , \tag{6.17}$$

where the second equality comes from assuming that the secondary field is vertical and constant across the disk; but what really matters here is that $B^* \propto I$ since the geometry is fixed. We now write an equation for the electrical current, this time taking into consideration the counter-electromotive force associated with self-inductance of the circuit:

$$\mathcal{E} - L\frac{\mathrm{d}I}{\mathrm{d}t} = RI \tag{6.18}$$

where L is the coefficient of self-inductance, and the current I is now a function of time. Substituting eqs. (6.16) and (6.17) into this expression, leads to

$$L\frac{\mathrm{d}I}{\mathrm{d}t} = \frac{\omega a^2}{2}\left(B_0 + \frac{MI}{\pi a^2}\right) - RI \tag{6.19}$$

indicating that the current –and thus the magnetic field– will grow provided that initially,

$$\frac{\omega a^2 B_0}{2} > RI \ . \tag{6.20}$$

which it certainly will at first since $I = 0$ at $t = 0$. There will eventually come a time (t_*) when the secondary magnetic field will be comparable in strength to the externally applied field B_0, at which point we may as well "disconnect" B_0; eq. (6.19) then becomes

$$L\frac{\mathrm{d}I}{\mathrm{d}t} = \left(\frac{\omega M}{2\pi} - R\right) I \ , \tag{6.21}$$

which integrates to

$$I(t) = I(t_*) \exp\left[\frac{1}{L}\left(\frac{\omega M}{2\pi} - R\right) t\right] \ . \tag{6.22}$$

indicating that the current —and magnetic field— will grow provided the externally-imposed angular velocity exceeds a critical value:

$$\omega > \omega_c = \frac{2\pi R}{M} \ . \tag{6.23}$$

This is not a (dreaded) case of perpetual motion, or creating energy out of nothing, or anything like that. The energy content of the growing magnetic field ultimately comes from the biceps of the poor bastard working ever harder and harder to turn the crank and keep the angular velocity ω at a constant value, as you'll get to verify in one of the problem at the end of this chapter in the simpler context of the homopolar generator.

There are many features of this dynamo system worth noting, and which all find their equivalent in the MHD dynamos to be studied in chapters to follow:

1. There exist a critical angular velocity that must be reached for the self-inductance to beat Ohmic dissipation in the resistor, leading to an exponential growth of the magnetic field; below this critical value, the field decays away exponentially once the initial field B_0 is removed.

2. Not all circuits connecting the edge of the disk to the axle will operate in this way; if we suddenly reverse the rotation of the disk, or wrap the wire the other way around the axle, the magnetic field produced by the loop will *oppose* the applied field;

3. The externally applied magnetic field B_0 is only needed as a *seed field* to initiate the amplification process.

4. The homopolar dynamo is really nothing more than a device turning mechanical energy into electromagnetic energy, more specifically magnetic energy.

6.5 The astrophysical dynamo problem(s)

Copper wires and sliding contacts being a rather sparse commodity in the universe, we must now figure out to apply the general idea of a dynamo to astrophysical fluids. In the MHD limit, our hope lies evidently with the induction term $\nabla \times (\mathbf{u} \times \mathbf{B})$ in the induction equation

$$\frac{\partial \mathbf{B}}{\partial t} = \nabla \times (\mathbf{u} \times \mathbf{B}) - \nabla \times (\eta \nabla \times \mathbf{B}) \ . \tag{6.24}$$

Remember that there are no true source term in eq. (6.24); if $\mathbf{B} = 0$ at some t_0, then $\mathbf{B} = 0$ for all subsequent $t > t_0$. We must therefore assume that some seed field exists to start up the dynamo process, just as in the homopolar dynamo we just looked into. As we saw in §2.10, there exist viable candidates to produce this seed field, most notably battery mechanism associated with mechanical separation of electric charges.

In its simplest form, the **dynamo problem** consists in finding a flow field $\mathbf{u}$ that can sustain a magnetic field against Ohmic dissipation. We must distinguish **kinematic dynamo**, where the flow field $\mathbf{u}$ is considered given *a priori* and constructed without any regards for its underlying dynamics, from what can only be called (for lack of a generally agreed-upon terminology) the **full dynamo problem**, in which the flow $\mathbf{u}$ results from a solution of the full set of MHD equations (§1.7), including the backreaction of the magnetic field on the flow via the Lorentz force term $\mathbf{J} \times \mathbf{B}$ on the RHS of the Navier-Stokes equation.

The kinematic regime carries the immense practical advantage that the induction equation then becomes truly linear in $\mathbf{B}$, and the dynamo problem reduces to finding a (smooth) flow field $\mathbf{u}$ that has the requisite topological properties to lead to field amplification. In the following chapters we will concentrate mostly on this kinematic regime, but will occasionally touch upon the much more difficult dynamical problem, mostly via direct numerical simulation of the full set of MHD equations.

As we'll see in the following chapter, there are flows that can amplify a magnetic field during a transient time interval, after which $\mathbf{B}$ decays again. So we tighten our definition of the dynamo problem by demanding that a flow be a dynamo if it can lead to $\mathcal{E}_{\rm B} > 0$ for times much larger than all relevant advective and diffusive timescales of the problem. To make things even harder, we'll add the additional condition that no electromagnetic energy be supplied across the domain boundaries S, i.e., $\mathbf{S} \cdot \mathbf{n} = 0$ in eq. (1.87). It is readily shown that this latter condition is satisfied if either (1) $\mathbf{B} = 0$ on the boundary, or (2) the components normal to the boundaries of $\mathbf{U}$, $\mathbf{B}$, and $\mathbf{L}$ all vanish on S (do problem XXX!).

The **solar dynamo problem** can be tackled either in kinematic or fully dynamical form. The aim there is to reproduce observed spatiotemporal patterns of magnetic field evolution, a minimum list of features having already been listed at the end of §6.1. As will become obvious in the following chapter, even this basic short list is a pretty tall order. Yet, from solar irradiance variations and their possible influence on Earth's climate to space weather prediction, it all begins with the solar cycle. Keep this in mind as we now start to dig into the mathematical and physical intricacies of magnetic field generation in electrically conducting fluids. We'll seem to venture pretty far away from the sun and stars at times, but stick to it and you'll see it all fitting together at the end. And now, into the abyss...

Problems:

1. Estimate the solar rotation period from the apparent motion on the sunspots drawing reproduced on Fig. 6.4. What are the primary difficulties in carrying out this kind of analysis?

2. The sunspot number time series reproduced on Fig. 6.5 are almost certainly the most intensively studied time series in All Of Astrophysics, as measured by the number of

published research papers per data point. So you need to try your hand at crunching it a little bit. First, go to the SIDC's Web Page:

`http://sidc.oma.be`

click on "Sunspot archive & graphics", and grab the dataset for the 13-month running mean of the monthly sunspot number (red line on Fig. 6.5 herein). Then,

(a) Measure the duration of each cycle, and compute the mean sunspot cycle period;

(b) Measure the peak and integrated (i.e., area-under-the-curve) cycle amplitude; do these two measures of cycle amplitude correlate well?

(c) Measure the rise time, i.e., the time elapsed from start of a cycle to its peak. Does this correlate to anything you have extracted so far (cycle duration, amplitude, etc.)?

(d) Do a lag analysis by looking for correlation between the amplitude of one cycle, and that of the preceeding cycle; that of two cycles ago; three cycles ago, etc. Do you find any significant correlation for some lag?

(e) Finally, calculate a power spectrum of the time series. Do you find significant peaks at periods other than $\sim 11\,$yr?

3. This problem has you quantify and reflect upon some of the statements made in §6.2.3.

(a) Fill in all missing mathematical steps leading to eq. (6.10).

(b) Compute and plot curves showing the variations of the rise time with assumed magnetic field strength, for four fixed values of the magnetic flux: $\log\Phi_B = 12$, 13, 14 and 15 (in weber). In all cases you may assume that the participating flux tube are released at a depth $r_0/R_\odot = 0.80$ within the convective envelope, where the density and temperature assume values $\rho = 4000\,\mathrm{kg\ m}^{-3}$ and $T = 5\times10^6\,$K. Remember that fixing the magnetic flux implies a relationship between a and B.

(c) Make a list of all the assumptions having entered this little derivation; which are the most/least reasonable ones?

4. homopolar generator: compute work done against magnetic force by externally applied torque; verify that it is equal to the energy dissipated in the resistor.

Bibliography:

On the telescopic re-discovery of sunspots in the seventeenth century, and ensuing debates over priority and physical interpretation, see

Mitchell, W.M. 1916, "The history of the discovery of the solar spots", in *Popular Astronomy*, **24**, 22-ff,

Shea, W.R. 1970, "Galileo, Scheiner, and the interpretation of Sunspots", *Isis*, **61**, 498-519,

Van Helden, A. 1996, "Galileo and Scheiner on sunspots", in *Proc. Am. Phil. Soc.*, **140**, 358-396,

and especially Galileo's original writings on the topic:

Galileo, G. 1613, *Letters on Sunspots* [in S. Drake (trans.) 1957, *Ideas and Opinions of Galileo*, Doubleday].

If such historical issues are of interest to you, you can also consult the ever-being-enlarged Web site "Great Moments in the History of Solar Physics":

`http : //www.astro.umontreal.ca/ ~ paulchar/history.html`.

Hale's original papers on sunspots are still well worth reading. The two key papers are:

Hale, G.E. 1908, "On the probable existence of a magnetic field in sunspots", *The Astrophysical Journal*, **28**, 315-343,
Hale, G.E., Ellerman, F., Nicholson, S.B., and Joy, A.H. 1919, *The Astrophysical Journal*, **49**, 153-178.

On the Maunder minimum, see

Eddy, J. A., 1976, *Science*, **192**, 1189-1202,
Eddy, J. A., 1983, *Solar Phys.*, **89**, 195-207,
Ribes, J. C., and Nesme-Ribes, E. 1993, *Astron. Ap.*, **276**, 549-563.

and on cosmogenic radioisotopes:

Beer, J. 2000, *Sp. Sci. Rev.*, **94**, 53-66.
Usoskin, I.G., Solanki, S.K., & Kovaltsov, G.A. 2007, *Astron. Ap.*, **471**, 301.

The study of rising toroidal flux ropes a proxy for the emergence of the solar internal toroidal field in the form of sunspot pairs is a topic that has exploded in the past 15 years or so. Among the many noteworthy contributions in this field , we recommend the following as starting points:

Moreno-Insertis, F. 1986, A&A 166, 291,
Choudhuri, A.R., & Gilman, P.A. 1987, *Astrophys. J.*, **316**, 788,
Fan, Y., Fisher, G.W., & DeLuca, E.E. 1993, *Astrophys. J.*, **405**, 390,
D'Silva, S., & Choudhuri, A.R. 1993, A&A 272, 621,
Caligari, P., Moreno-Insertis, F., & Schüssler, M. 1995, *Astrophys. J.*, **441**, 886.

Important earlier papers on the topic are:

Parker, E.N. 1955, *Astrophys. J.*, **122**, 293,
Parker, E.N. 1975, *Astrophys. J.*, **198**, 205,
Schüssler, M. 1977, A&A 56, 439,
Moreno-Insertis, F. 1983, A&A 122, 241.

On the interaction of a rising magnetic flux rope with helical convection, see

Fisher, G.H., Fan, Y., Longcope, D.W., Linton, M.G., & Pevtsov, A.A. 2000, *Solar Phys.*, **192**, 119-139,

and references therein. The thin flux tube approximation used in most of these calculations is usually credited to

Spruit, H.C. 1981, A&A 98, 155.

Considerable effort is currently being put into doing away with the thin flux tube approximation, in order to see which of the above results remains robust once the flux tube is no longer treated as a one-dimensional object. This is a rapidly moving field, so for the latest see the following recent on-line review by Yuhong Fan:

$\texttt{http}: //\texttt{www.livingreviews.org/lrsp} - 2004 - 1.$

as well as these now-classics:

Longcope, D.W., Fisher, G.W., & Arendt, S. 1996, *Astrophys. J.*, **464**, 999,
Emonet, T., & Moreno-Insertis, F. 1998, *Astrophys. J.*, **492**, 804,
Fan, Y., Zweibel, E., and Lantz, S.R. 1998, *Astrophys. J.*, **502**, 968.

On the storage and stability of toroidal flux ropes below the solar convective envelope, see

Ferriz-Mas, A., & Schüssler, M. 1994, *Astrophys. J.*, **433**, 852,
Ferriz-Mas, A. 1996, *Astrophys. J.*, **458**, 802.

The amount and variety of solar data (numbers, images, movies) available online is quite simply staggering. Use your Web-surfing skills to locate the Web sites of NASA's Marshall Space Flight Center, of the Solar Stanford Center, of the High Altitude Observatory, of NASA's Goddard Space Flight Center, of SOHO and TRACE, to name but a few.