NOT TO APPEAR IN NONLEARNED J., 45.
Preprint typeset using LaTeX style emulateapj v. 12/14/05

A PHYSICAL MODEL FOR THE TOTAL SOLAR IRRADIANCE

A. D. CROUCH, P. CHARBONNEAU, G. BEAUBIEN, AND D. PAQUIN-RICARD
Département de Physique, Université de Montréal, Montréal, QC, H3C 3J7, Canada
Not to appear in Nonlearned J., 45.

ABSTRACT

We present a physical model for the total solar irradiance. The model takes the observed location, timing, and area of emerging active regions as input and produces a time-evolving size distribution of magnetic structures over the solar surface. We assume that the bright magnetic structures (faculae), that counteract the irradiance deficit caused by sunspots, consist of the products of active region decay. We simulate the decay process as a combination of fragmentation and boundary erosion of large-scale magnetic structures. The model has several adjustable parameters that control the decay processes and the irradiance contribution from the quiet-Sun and the small-scale magnetic elements that are produced during the decay process. We use a genetic algorithm to estimate these parameters by fitting to the observed irradiance and daily sunspot area time series over the 1978–2007 time interval. Given the simplifications associated with the model, the resultant parameter values are well constrained within the imposed ranges. In addition, the irradiance and daily sunspot area time series produced by the best fit models agree very well with the observations, although the sunspot area fits tend to perform better than the irradiance fits. However, it is evident that the model is neglecting a significant source of excess brightness, which manifests itself in two ways. First, the optimal values for the lifetime and intensity contrast of the bright, small-scale flux elements are both larger than expected. Second, the synthetic irradiance consistently underestimates the observations during the ascending phase of a cycle, despite the daily sunspot area fitting the observations quite well during these times. We also show that this genetic forward modelling approach can be used to detect a long term trend of decadal time scale in the quiet-Sun irradiance. Assuming a constant quiet-Sun irradiance, we reconstruct the total solar irradiance over the 1874–1978 time interval, for which observational data of emerging active regions is available.

Subject headings: Sun: activity — Sun: faculae, plages — Sun: magnetic fields — sunspots

1. INTRODUCTION

The magnitude of the role played by the Sun in the warming trend observed in the twentieth century remains a topic of controversy (see, e.g., Foukal et al. 2006, and references therein). Nearly continuous irradiance data now spanning almost three full solar activity cycles show a small ($\simeq 0.1\%$) increase in irradiance at activity maximum over the solar minimum value, contrary to what one would naively expect from the irradiance deficit associated with a greater surface coverage of cooler, darker sunspots. This is because active regions also contain magnetized structures that are brighter than the quiet photosphere, most notably faculae. Observations and modelling have demonstrated that faculae and the magnetic network away from active regions can jointly account for the observed irradiance excess (Foukal & Lean 1986, 1988; Foukal et al. 1991, and references therein). In fact, semi-empirical irradiance models including contributions from spots, faculae and the network reproduce quite well the observed irradiance variations over the appearance and decay of active regions, up to cycle-like timescales (e.g., Chapman, Cookson, & Dobias 1996; Lean et al. 1998; Fligge, Solanki, & Unruh 2000; Walton, Preminger, & Chapman 2003). From this point of view, irradiance variations are produced exclusively by the variations in relative surface coverage of various photospheric structures with distinct emissivities.

Another possibility is that irradiance variations are driven at least partly by magnetically-mediated alterations of convective energy transport. Without denying that irradiance does vary on shorter timescales in response to the varying coverage of sunspots and faculae, this class of explanations ascribes irradiance variations on solar cycle timescales (and perhaps longer) to a global, magnetically induced change in the structure of the solar convection zone (see Kuhn & Stein 1996; Li et al. 2003, for two different viewpoints on the matter). Indeed, Kuhn, Libbrecht, & Dicke (1988) have estimated that up to 50% of the observed irradiance variations on solar cycle timescales could be generated in this manner. Observationally, the idea is buttressed by spectroscopic observations suggesting cyclic variations in the photospheric temperature gradient (e.g., Gray & Livingston 1997) and observed cyclic changes in the frequency of solar acoustic p-mode frequencies (Woodard 1987; Bhatnagar, Jain, & Tripathy 1999). From the theoretical point of view, the magnitude of possible irradiance variations is strongly constrained by the high efficiency of convective energy transport, and high thermal capacity of the solar convective envelope as a whole (e.g., Spruit 2000). Yet this argument relies on a "diffusive" interpretation of convective energy transport, and so leaves open the possibility that changes in the global pattern of convection by the Lorentz force could produce a temperature signal at the surface, perhaps even with a latitude dependence (cf. Kuhn, Libbrecht, & Dicke 1988). Unfortunately, current MHD simulations of the solar convection zone (e.g., Brun, Miesch, & Toomre 2004) have not yet managed to produce solar cycle-like large-scale magnetic fields, and so cannot help in settling the issue at this juncture.

Numerous reconstructions of the solar irradiance on century timescales can be found in the literature, based on either of the above approaches (e.g., Hoyt & Schatten 1993; Lean, Beer, & Bradley 1995; Solanki & Fligge 1998; Foukal 2002; Preminger & Walton 2005; Wang, Lean, & Sheeley 2005; Tapping et al. 2007). The vast majority of these reconstructions are based on semi-empirical or statistical relationships established using twentieth century solar observations. Generally speaking, great caution must be exerted in extrapolating statistical models outside of the range in which they are calibrated. This difficulty becomes quite critical in the solar context, since cycles 21, 22 and 23 did not differ very much in their overall activity levels (but do see de Toma et al. 2004). Yet one would certainly like to estimate how solar irradiance behaves during extended periods of strongly suppressed magnetic activity, such as the 1645–1715 Maunder Minimum, or during a secular rise in activity levels, such as in the time period 1900—1950. The latter is particularly germane to current debates regarding the respective roles of solar and anthropogenic influences on global warming. In both cases extrapolation uncertainties present a major obstacle to the reliability of irradiance reconstruction in the (relatively) distant past.

One way out of this difficulty is to establish a *physical*, rather than statistical, relationship between irradiance excess contributors such as the network and faculae, and observables such as sunspots for which a relatively homogeneous time series exists all the way back to the mid-seventeenth century. What we mean by "physical" is that facular coverage (for example) is calculated via a physical model producing it from a single time series of well-defined observables, such as the location and time of emerging active regions. As with statistical models based on empirical correlations, such a model, once calibrated on the irradiance time series for cycles 21 through 23, is to be applied outside the parameter regime within which is was calibrated. Yet, one would like to think that the physical mechanisms through which a sunspot decays (for example) are universal, in the sense that they do not depend explicitly on the overall activity level. Extrapolating such a model outside of its defining parameter regime is then, one would hope, a safer procedure than direct extrapolation of purely statistical correlations between, say, facular and sunspot coverages.

In this paper, we develop and use a stochastic fragmentation/erosion model whereby individual sunspots are continuously "injected" into a model "solar surface", and subsequently fragment and/or erode into smaller magnetic structures, leading to the buildup of a population of small, bright magnetic elements which then produce an irradiance excess over the quiet photosphere. More specifically, what this fragmentation model produces is a time-evolving size distribution of magnetic structures over the solar disk. This is then combined with a semi-empirical "contrast function" measuring the observed irradiance excess or deficit of a magnetic structure as a function of its size and position on the disk, thus yielding an irradiance time series. While the model does not yet include explicitly coherent spatial structures such as faculae and network, it reproduces their effects on irradiance, in an average sense if not in all details. It has the great advantage that it can be extrapolated to earlier centuries with a minimum of uncertainty.

The model itself is described in § 2. In § 3 we use a genetic algorithm to estimate the model parameters by fitting to the 1978–2007 irradiance and daily sunspot area time series, with observed areas of emerging active regions as input to the model. This fitting procedure turns out to be far from trivial, because of the inherently stochastic nature of the fragmentation/erosion model along with the uncertainties related to active regions that emerge on the far side of the Sun or near the limb. The best-fit models are then used in § 4 to reconstruct an irradiance time series over the 1874–1978 time interval, for which observational data of emerging active regions is available. In § 5 we use this technique to search for the existence of a long-term trend in the quiet-Sun irradiance over the interval 1978–2007.

2. THE MODEL

2.1. *The injection of active regions*

Where possible the injection of individual sunspots onto the model solar surface is based on observations of emerging active regions. We derive the position, time, and area from the daily sunspot observations produced by the Royal Greenwich Observatory (RGO) from May, 1874 to the end of 1976 and from the US Air Force (USAF) and the US National Oceanic and Atmospheric Administration (NOAA) for later times (up to March, 2007). This information has been made publicly available by David Hathaway[1]. It should be noted that the formats are slightly different for the two data sets; for example, the position angles are rounded differently in each data set. The change occurs almost two years before the irradiance record begins; consequently, the calibration of the fragmentation/erosion model is done entirely with USAF/NOAA data. When we perform the extrapolation to times before the irradiance observations commence, we assume that the sunspot observations are homogeneous; although, the USAF/NOAA active region areas are multiplied by a correction factor of 1.4 as required.

Typically, after the initial appearance of an active region its area will grow over a period of a few days and reach some peak value before decreasing. For simplicity, we inject an active region into the simulation at the time it reaches its maximum size. The initial growth phase of the active region and its subsequent contribution to the irradiance and daily sunspot area are neglected. Because active regions generally consist of two sunspots, we split the active region into two equal parts and allow each part to decay according to the fragmentation/erosion model described in § 2.2. More complex initial distributions could be employed to account for active regions that consist of more than two spots; however, such distributions can result naturally from the fragmentation process and are therefore included, though indirectly and with some time-delay.

We impose three constraints on the observations that are consistent with constraints imposed by others that have examined the RGO data (e.g., Moreno-Insertis & Vazquez 1988; Martinez Pillet et al. 1993; Hathaway & Choudhary

[1] `http://solarscience.msfc.nasa.gov/greenwch.shtml`

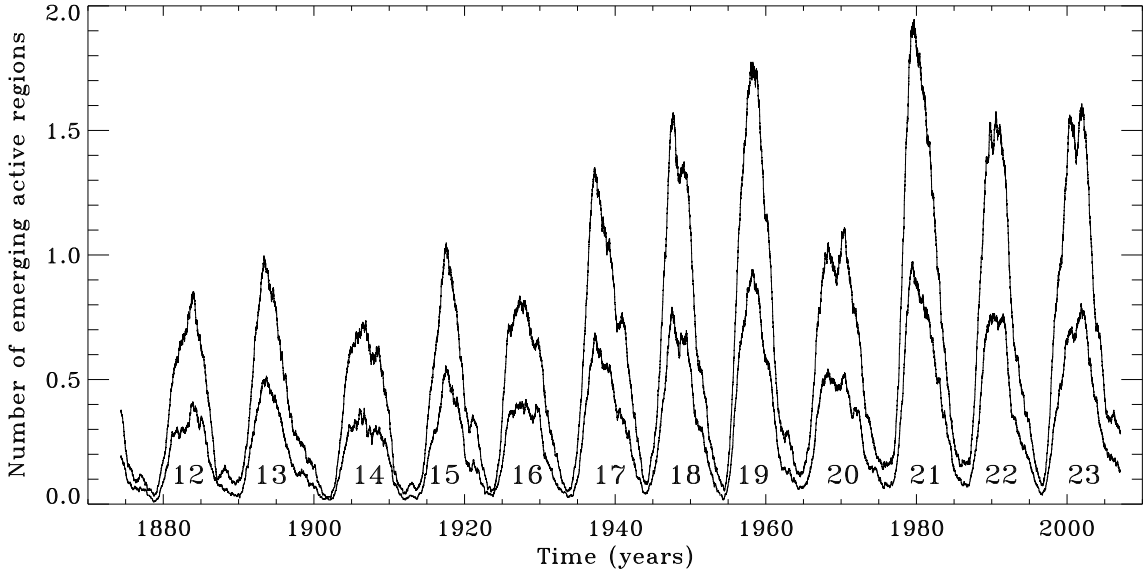

FIG. 1.— Average number of emerging active regions per day as a function of time. The lower curve is the average number of active regions per day that emerge on the observed section of the solar surface, whereas the upper curve is for the unobserved section. In each case, the data is boxcar averaged with a box size of 365 days. Cycles are numbered according to the convention introduced in the nineteenth century by Rudolf Wolf.

2004; Baumann & Solanki 2005). First, because sunspot area corrections (for projection effects) are large for observations near the limb, we only consider active regions that attain a peak area within 60° in longitude of the central meridian. In line with this constraint, no active regions are observed to emerge at latitudes greater than 50° from the equator. Second, we only include active regions with a corrected peak area greater than or equal to 35 millionths of a solar hemisphere (μHem), which is sufficiently above the spatial resolution limit of the RGO telescope. Third, the active region must be observed for at least three days during a single disk passage; although the observations need not be made on consecutive days (data gaps are allowed). It is also worth noting that we have accounted for recurrent active regions (those that cross the disk more than once). According to these constraints we will include active regions that did not have an observed growth phase (i.e., the peak area occurs on the first day it was observed). Consequently, in a small fraction of cases we may be underestimating the area of the active regions when they emerge.

For the sections of the solar surface that cannot be observed reliably or at all – in the vicinity of the limb and the far side – the properties of the injected active regions must be modelled. We do this in a manner that produces properties that are statistically consistent with the properties of observed active regions. In the near future, it may be possible for helioseismic far side imaging to reduce the uncertainties associated with this procedure (Lindsey & Braun 2000; González-Hernández et al. 2007).

The section of the Sun that is not observed, outside 60° in longitude of the central meridian, represents two thirds of the solar surface in longitude. For each third, we set the number of emerging active regions on a given day to the number of emerging active regions in the observed third but on a randomly selected day in the recent past or near future. The selection probability is normally distributed about the current day with a standard deviation of 100 days. Larger values of the standard deviation result in a loss of agreement in the temporal variation; whereas, shorter values cause the unobserved regions to behave almost identically to the observed regions. Figure 1 shows a comparison of the average number of active regions per day that emerge in the observed and unobserved parts of the Sun. As expected roughly twice as many active regions emerge in the unobserved part in comparison to the observed part (a total of 28670 and 14375, respectively, for this particular realisation). Figure 1 shows that the temporal behaviour of the modelled emergence rate is very similar to the observations, including the shape and relative magnitude of the peaks. It can also be verified that the probability distribution of the number of emerging active regions and its variation with different phases of the solar cycle agree well with observations.

The same technique is used to assign a latitude to each modelled active region; that is, the modelled latitude is copied from the latitude of an observed active region from the recent past or near future. On days with multiple observed emerging active regions the selected latitude is chosen at random. The longitude of each modelled active region is randomly distributed in a uniform fashion within each unobserved third; thus, active longitudes are ignored.

The area of each modelled emerging active region is drawn from a distribution which is based on a lognormal fit to the observed distribution. The lognormal distribution was chosen because it has been shown that it is a good approximation for both the distribution of individual sunspot umbral areas (Bogdan, Gilman, Lerche, & Howard 1988) and the distribution of maximum sunspot group areas (Baumann & Solanki 2005). Figure 2 shows the probability density function (PDF) for the areas of the observed and modelled emerging active regions. For each bin, the probability density is defined as

$$\mathrm{PDF}\,(A) = \frac{1}{N}\frac{\Delta N}{\Delta A}\,, \tag{1}$$

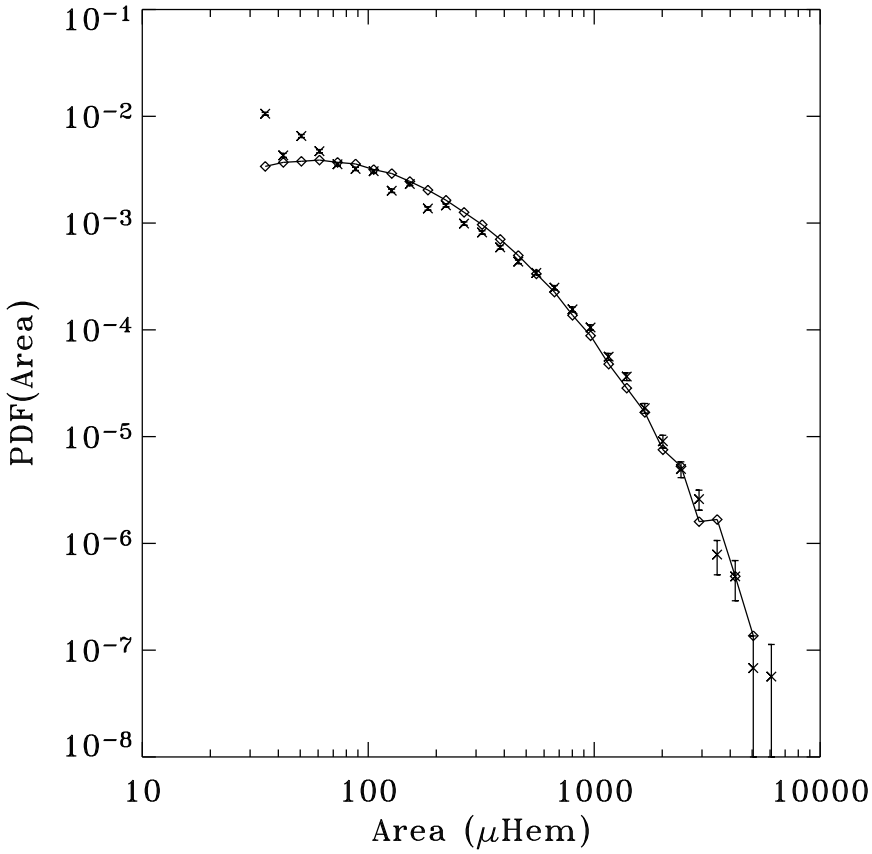

FIG. 2.— Probability density function (PDF) of the area of the emerging active regions. The crosses with error bars represent the PDF for the observed active regions. The error bars are based on the square root of the raw count in each bin (assuming Poisson counting statistics), they are of similar size to the crosses at small area. The diamonds joined by full curves are the PDF for the modelled active regions in the unobserved section of the Sun (the curve is only included to guide the eye).

where A is the area of the active region, ΔA is the bin width (which is constant in log-space in Figure 2), ΔN is the number of active regions in the size range $[A\,, A + \Delta A)$, and N is the total number of active regions examined. The PDF for the observed active regions does not show any significant variation with time, either from cycle to cycle or at different phases of each cycle. This is consistent with the findings of Bogdan, Gilman, Lerche, & Howard (1988). Likewise, the area PDFs for emerging active regions before 1976 (RGO format) and after 1976 (USAF/NOAA format) are not substantially different. The USAF/NOAA area PDF tends to have a slightly higher fraction of the smallest and largest areas; however, it is difficult to determine if the discrepancy is due to the difference in data source or if it is solar in origin. Consequently, to estimate the parameters associated with the lognormal distribution we combine observations from all times (both data sets). Each modelled active region is required to have a minimum area of 35 millionths of a solar hemisphere (consistent with the constraint on the observations) and a maximum size no greater than the observed maximum. Figure 2 shows that the PDFs do not match perfectly over the entire area range – the model appears to underestimate the frequency of the smallest areas. However, for the intermediate range of areas the two distributions appear to agree quite well.

2.2. *A stochastic fragmentation/erosion model for active region decay*

The processes governing active region decay are not well understood. For sunspots, the observational evidence tends to favor a decay law in which the spot area is a quadratic function of time (Petrovay & van Driel-Gesztelyi 1997). However, it is difficult to distinguish between linear and quadratic decay laws for sunspots and sunspot groups (Moreno-Insertis & Vazquez 1988; Martinez Pillet et al. 1993). Linear and quadratic decay laws have very different consequences for the theory governing sunspot decay. Gokhale & Zwaan (1972), Meyer et al. (1974), and Krause & Ruediger (1975) have proposed models which result in a linear decay law, which implies that the decay rate is independent of the size of the spot. In the Gokhale & Zwaan model sunspot decay is due to Ohmic dissipation across a current sheet, whereas in the models of Meyer et al., and Krause & Ruediger the mechanism is turbulent diffusion. Another promising suggestion is that sunspot decay is due to the erosion of the sunspot boundary (e.g., Simon & Leighton 1964; Petrovay & Moreno-Insertis 1997; Martínez Pillet 2002; Thomas et al. 2002). This corresponds to a quadratic decay law and a decay rate that is proportional to the sunspot circumference. This suggestion is also supported by the presence of moving magnetic features (MMFs) in the moat flow surrounding sunspots (e.g., Sheeley 1969; Harvey & Harvey 1973; Vrabec 1974; Zhang et al. 2003). However, it is not clear if the net flux transported by MMFs is sufficient to completely explain sunspot decay (e.g., Hagenaar & Shine 2005). Another potentially important

process in active region evolution is the fragmentation of the constituent flux elements (Wallenhorst & Topka 1982; Martin et al. 1985; Schrijver 2001).

After it emerges each active region is broken into two equally sized pieces. Thereafter, each piece undergoes a decay process that we approximate as a combination of classical fragmentation and boundary erosion. We assume that this decay process is solely responsible for the production of the small-scale magnetic flux elements that constitute faculae (plage) and the magnetic network. For faculae, this idea is partly supported by the results of Preminger & Walton (2005) and Chapman & Hoffer (2006). However, it should be noted that this assumption is not entirely correct because active region decay is not the only source of faculae (e.g., Foukal 1998). Additionally, the flux that sustains the magnetic network mostly emerges outside active regions as small-scale elements that come from ephemeral regions and possibly local dynamo action (e.g., Harvey & Zwaan 1993; Wang et al. 1995; Schrijver et al. 1997; Zhang et al. 1998b; Cattaneo 1999; Hagenaar 2001; De Pontieu 2002; Hagenaar et al. 2003; Abramenko et al. 2006; Zhang et al. 2006). The details of how faculae and the network are produced are not modelled here. To properly account for the emergence, dispersal, and interaction of flux elements on a range of size scales would require a sophisticated magnetic flux dispersal scheme that is too computationally expensive for the task we have upon us (e.g., Schrijver et al. 1997; Parnell 2001; Simon et al. 2001; Schrijver 2001; Krijger & Roudier 2003; Crouch et al. 2007). Consequently, the interaction and motion of the decay products is ignored.

The model consists of two components: large- and small-scale magnetic flux elements. Large-scale elements are defined as flux concentrations with areas larger than $A_f^* = \pi(r_f^*)^2$, assuming each element is circular. Thus, r_f^* is the scale that divides large-scale magnetic elements from small-scale ones. As we discuss below, r_f^* could be much smaller than the radius of a typical sunspot. Therefore, the class of objects that we refer to here as large-scale elements may include sunspots and pores. Large-scale elements are subject to fragmentation and boundary erosion, the products of which are small-scale elements.

We work with time steps of one day. On each day, each existing large-scale element has two chances of fragmenting. If fragmentation occurs at the first attempt then the products may also undergo further fragmentation on the second attempt. For each attempt a trial is performed where a random number x_1, uniformly distributed in the range 0–1, is calculated. The probability of fragmentation is denoted by p_{frag}; thus if $x_1 < p_{frag}$ then fragmentation occurs and the large-scale element is broken into two separate pieces with areas $x_2 A$ and $(1 - x_2)A$, where x_2 is a random number uniformly distributed in the range 0.1–0.9 and A is the area of the large-scale element at the beginning of the trial.

Whether or not fragmentation occurs, each resultant large-scale element can suffer boundary erosion. The area removed by erosion on each day is perimeter dependent and is equal to

$$\Delta A_{erode} = \pi r_{erode} \left(2r - r_{erode}\right) , \tag{2}$$

where r is the original radius of the large-scale element and $r_{erode} = x_3\, r_{erode}^*$, with r_{erode}^* being the model parameter that controls the maximum radial thickness of the eroded annular area and x_3 a random number uniformly distributed in the range 0–1. The area removed ΔA_{erode} is converted into small-scale elements with an efficiency, η; meaning that the collective area of small-scale elements produced is $\eta \Delta A_{erode}$. The inclusion of η is designed to allow for the possibility that some area (or magnetic flux) is lost during the decay process as opposed to being converted to small-scale flux elements. Possible causes of this could be submergence, cancellation or reconnection (e.g., Wallenhorst & Howard 1982; Rabin et al. 1984; Martin et al. 1985; Zirin 1985; Harvey et al. 1999; Kálmán 2001; Chae et al. 2004). If the large-scale element has a radius smaller than r_{erode} prior to undergoing boundary erosion then boundary erosion does not occur; because r_{erode}^* is an upper bound this large-scale element may suffer boundary erosion (and/or fragmentation) at a subsequent time step. It should be noted that in reality only structures with penumbra (i.e., sunspots) may suffer significant boundary erosion as MMFs are advected away by the moat flow (Martínez Pillet 2002). Thus, the slowing of the erosion as the element decays is reasonable. Fragmentation and erosion continue to diminish the size of each large-scale element until its radius falls below the threshold radius r_f^*. At this point we assume the large-scale element disintegrates totally into small-scale elements, also with an efficiency η.

Small-scale elements are not subject to fragmentation or boundary erosion as described above. We assume that the collective area of the small-scale elements associated with a particular active region decays exponentially, so the area removed from the collection of small-scale elements on a given day is equal to

$$\Delta A = \lambda A , \tag{3}$$

where λ is the model parameter that controls the rate of decay and A is the area of the collection of small-scale elements at the beginning of the day. When the collective area of the small-scale elements associated with a particular active region gets smaller than $A_t^* = \pi(r_t^*)^2$ we assume that the collection of small-scale elements vanishes altogether (of course, more can be produced at a later time). The inclusion of a threshold size for the small-scale elements is primarily computational – to avoid tracking infinitesimally small areas that make a negligible contribution to the TSI.

2.3. *Transport of magnetic flux*

We assume that each active region is advected poleward by the meridional flow. We approximate this with the function (Komm et al. 1993)

$$M\left(\phi\right) = 12.9 \sin\left(2\phi\right) + 1.4 \sin\left(4\phi\right) \quad \left(\mathrm{m\ s^{-1}}\right) , \tag{4}$$

where ϕ is the latitude. For realistic parameter settings the meridional flow is most important for the small-scale flux elements because they tend to survive long enough for the impact to be significant. On the other hand, large-scale flux elements decay relatively quickly so the meridional flow does not have sufficient time to transport them a significant distance.

Our model Sun is assumed to be spherical with a synodic rotation period of 27.2753 days at the equator. We also include the influence of solar differential rotation. Thus, on each day the longitude of each active region is shifted relative to the equator by an amount

$$\Delta\Omega\left(\phi\right) = -1.95\sin^2\phi - 2.09\sin^4\phi \quad \left(\text{deg day}^{-1}\right) , \tag{5}$$

where the numerical coefficients are derived from the results of Charbonneau et al. (1999) for the sidereal rotation rate. Differential rotation turns out to be a very minor factor in our calculations but is included nevertheless.

2.4. *Calculating the total solar irradiance*

On each day, the total solar irradiance, S, is calculated by summing the contributions from each Earth-facing active region, which consists of time-varying amounts of large- and small-scale flux elements. Thus, the resultant irradiance is calculated according to the three component model,

$$S = S_Q + \Sigma_i \Delta S_s + \Sigma_j \Delta S_{fac} , \tag{6}$$

where S_Q is the quiet-Sun irradiance, and ΔS_s and ΔS_{fac} are the contributions to the irradiance from individual sunspots i and faculae j, respectively (the sums represent the total contribution in each case). It should be noted that in equation (6) we are now characterising magnetic structures of various sizes as sunspots and faculae.

We assume that each large-scale element with a radius larger than r_s^* is a sunspot (from the point of view of its contribution to the irradiance). Large-scale elements with radii smaller than r_s^* make no contribution to the irradiance. For each sunspot, the irradiance deficit is calculated with the standard formula (e.g., Hudson et al. 1982; Fröhlich et al. 1994; Lean et al. 1998),

$$\frac{\Delta S_s}{S_Q} = \frac{1}{2}\mu A_s \left(3\mu + 2\right) \alpha_s , \tag{7}$$

where $\mu = \cos\theta\cos\phi$ (θ is the central meridian angle), A_s is the spot area in units a solar hemisphere ($2\pi R_\odot^2$), and α_s is the sunspot intensity contrast. Equation (7) assumes that sunspots and quiet-Sun show the same limb-darkening. We employ the sunspot intensity contrast defined by Brandt et al. (1994), which is well constrained by observations,

$$\alpha_s = -\left[0.2231 + 0.0244\log_{10}\left(A_s \times 10^6\right)\right] , \tag{8}$$

where the argument of the logarithm must be provided in terms of millionths of a solar hemisphere. The area dependence of α_s helps account for some of the properties which vary from spot to spot, such as the effective temperature and area ratios of umbra to penumbra.

There are several uncertainties regarding the facular contribution to equation (6). The most obvious is our lack of a detailed model for the dispersal and interaction of small-scale flux elements. Consequently, we can not distinguish those flux elements that form faculae from those that form network. In addition, there remain significant uncertainties concerning the precise centre-to-limb variation and the intensity contrast for both facular and network. In light of these uncertainties we opt for a simple approach. We ignore the irradiance contribution from the network entirely and assume that the irradiance contribution from the small-scale flux elements can be approximated by the photometric facular index formula (e.g., Chapman 1980; Chapman & Meyer 1986),

$$\frac{\Delta S_{fac}}{S_Q} = \frac{1}{2}\mu A_{fac} \left(3\mu + 2\right) \left(1/\mu - 1\right) \alpha_{fac} , \tag{9}$$

where A_{fac} is the area of the collection of small-scale flux elements in units of a solar hemisphere and α_{fac} is facular intensity contrast, which is assumed to be constant and has a typically quoted value of 0.036 (Chapman & Meyer 1986). Equation (9) assumes that the irradiance contribution from the faculae is zero at disk centre and peaks at the limb. This is consistent with the findings of Lawrence (1988), Foukal et al. (2004), and Ermolli et al. (2007), but not with those of Libbrecht & Kuhn (1985) and Ortiz et al. (2002, 2006) whose results indicate that the faculae contrast peaks somewhat prior to the limb. More complicated formulae could be employed with more free parameters (e.g., Chapman 1987; Steinegger et al. 1996; Lean et al. 1998; Ortiz et al. 2002, 2006). However, we would argue that it is more important to first construct a model to properly account for the emergence, dispersal, and interaction of small-scale flux elements (e.g., Schrijver et al. 1997; Parnell 2001; Simon et al. 2001; Schrijver 2001; Krijger & Roudier 2003; Crouch et al. 2007). This would allow a sensible four component version of equation (6) to be constructed that included a term for the network contribution (e.g., Foukal et al. 1991; Steinegger et al. 1996). This is beyond the scope of the present investigation so, for now, we retain equation (9).

In § 5 we explore the possibility that the solar irradiance may be varying on a longer timescale. To this end we allow the quiet-Sun irradiance, S_Q, to vary in a linear fashion over the period spanning the observed irradiance time series 1978–2007, i.e.,

TABLE 1
Model parameters

Symbol	Description	Free/Fixed	Value/Allowed range
	Fragmentation chances per day per element	fixed	2
	Fragmentation range	fixed	0.1–0.9
p_{frag}	Fragmentation probability	free	0–1
r_f^*	Threshold radius for large-scale elements	fixed	250 km
r_{erode}^*	Maximum radial thickness of the eroded annulus	free	100 km – 2 Mm
η	Area conversion efficiency	free	0–1
λ	Decay rate for small-scale elements	free	0–0.5
α_{fac}	Faculae intensity contrast	free	0.018–0.072
r_s^*	Threshold radius for sunspots	fixed	1 Mm
S_Q, $S_{Q,c}$	Quiet-Sun irradiance (and intercept of linear trend)	free	1365–1366 W m^{-2}
$S_{Q,m}$	Gradient of Quiet-Sun irradiance linear trend	free and fixed	[-1, 1] W m^{-2} over 1978–2007
r_t^*	Collective threshold radius for small-scale elements	fixed	25 km

$$S_Q = S_{Q,c} + S_{Q,m}\,(t - t_0)\ , \tag{10}$$

where t is the time step, t_0 is the time step of first irradiance observation in 1978, and $S_{Q,m}$ and $S_{Q,c}$ represent the gradient and vertical intercept of the straight line, respectively. If such a variation exists it will likely be more complicated than a simple linear trend (e.g., Tapping et al. 2007). Therefore, we do not intend to extrapolate the above trend outside the time interval over which it is validated. We will simply show that it is possible for a technique such as ours to detect an underlying trend of this type.

2.5. *Model Parameters*

Table 1 provides a summary of the various parameters associated with the model described in §§ 2.2 – 2.4. In some cases we have chosen to fix a parameter, in others we allow the parameter to vary over some range (referred to as free in the third column of Table 1). In either case, we justify the chosen value or range of values with observational evidence or theoretical arguments where possible. This section is devoted to discussing the various choices made for each parameter.

In our model the fragmentation probability, p_{frag}, is the same for all large-scale elements. One would assume that the probability of a large flux concentration fragmenting on a given day is quite high. However, the same may not be true for moderately sized flux concentrations (see Schrijver 2001, for a scheme where the fragmentation probability is flux-dependent). Given the lack of observational evidence or theoretical arguments to constrain p_{frag} we do not restrict its allowed range; we explore values in the range 0–1.

The fragmentation range controls the size of the individual pieces that are produced when fragmentation occurs. This is also difficult to constrain with observational evidence or theoretical models. For this reason we allow the fragmentation products to have areas between 10% and 90% of the original element area; thus, fragments with areas less than 10% of the original element are prevented.

Like the other parameters controlling the fragmentation of large-scale elements, the number of chances each element has to fragment on a given day is very difficult to constrain observationally or theoretically. We give each element 2 chances, because we found that the agreement between the modelled irradiance time series and the observations was better than the case with a single chance. Larger settings prove costly from a computational perspective and are avoided here.

The threshold radius for large-scale elements, r_f^*, is the size below which elements no longer undergo fragmentation or boundary erosion. We assume that fragmentation products smaller than this are converted completely into small-scale flux elements. There is observational evidence that indicates that fragmentation and shredding (i.e., boundary erosion) occur down to the smallest observable scales (e.g., Berger & Title 1996; Zhang et al. 1998a,b; De Pontieu 2002). However, we do not model these processes for each individual small-scale flux element. Instead, we assume that these effects are captured (on average) by λ, the parameter controlling the decay of the small-scale flux elements associated with each active region. The inclusion of r_f^* is primarily for computational efficiency (i.e., to avoid tracking the fragmentation and erosion of numerous individual elements over a broad range of spatial scales). Nevertheless, some care must be taken in choosing the value for r_f^*. It is reasonable to assume that r_f^* must be smaller than the radius of the smallest observed sunspots and pores, because there is no observational evidence to suggest these structures spontaneously disintegrate. By definition, the value for r_f^* also needs to be larger than the scale of the individual small-scale elements that constitute faculae and network. Ideally, r_f^* would lie in the range where magnetic flux elements tend to have a brightness comparable to the quiet-Sun, although this depends on the heliocentric angle (Topka et al. 1997; Solanki 2002). Considering all of these issues we set $r_f^* = 250$ km.

The model parameter that controls the maximum radial thickness of the annular area removed by boundary erosion from each large-scale element on each day is r_{erode}^*. According to Martinez Pillet et al. (1993) the average and median area decay rates for isolated sunspots are 19 and 15 μHem/day, respectively. Whereas, for sunspot groups the average and median are 41 and 31 μHem/day, respectively. Martinez Pillet et al. also find that the decay rates appear to be distributed in a lognormal fashion. Consequently, in addition to the slow decay rates corresponding to the median

value there is also a small fraction of spots with very rapid decay rates (up to 100 μHem/day or so). Therefore, is it sensible to test a broad range for r^*_{erode}, we explore the range: 100 km – 2 Mm. For example, according to equation (2), for a flux element with a radius of 25 Mm (area of 645 μHem) this range corresponds to a maximum of 5.15 – 99.1 μHem/day, which is consistent with the observed range. Note that ΔA_{erode} scales with element radius r so larger decay rates are possible for larger elements.

The area conversion efficiency, η, is not well constrained by observations. For this reason, we explore the widest plausible range: 0–1. However, one would expect that neither of the extreme limits is physically acceptable. In our model, $\eta = 0$ would mean that no bright, small-scale flux elements are produced during active region decay and thus there would be no mechanism to counteract sunspot darkening. On the other hand, $\eta = 1$ would imply a perfectly efficient area conversion process. It is possible that η is larger than 1, if the magnetic filling factor in sunspots is greater than that in faculae; however, we do not consider this case.

The fraction of the area that is lost per day from the collection of small-scale flux elements associated with each active region, λ, is also difficult to constrain observationally. Individual, isolated small-scale elements may have lifetimes of a few hours (e.g., Liu et al. 1994; Zhang et al. 1998a,b, 2003; Hagenaar & Shine 2005). However, we are interested in the lifetimes of larger scale structures that are composed of small-scale elements that can persist for days or weeks. Therefore, we would expect λ to be quite small so we test the range: 0–0.5. The lower limit is unphysical because it would imply that small-scale elements survive indefinitely. The upper limit prevents any structure composed of small-scale elements from losing more than half of its area in one day. This is reasonable for large scale structures (consisting of many small-scale elements) but may not be for smaller structures (of a few small-scale elements).

The faculae intensity contrast, α_{fac}, has been estimated from observations to have a value of about 0.036 (Chapman & Meyer 1986). Because our model does not account for the interaction and dispersal of the small-scale flux elements, the facular areas that we calculate will include elements that would in reality form the network. Therefore, we take a somewhat flexible approach to applying equation (9) to the small-scale flux elements; although, the centre-to-limb variation in equation (9) may not be appropriate for the network contribution (Foukal et al. 1991; Ermolli et al. 2003; Ortiz et al. 2002, 2006). We explore a wide range of values for α_{fac}: 0.018–0.072. The upper bound is two times the empirical value, whereas the lower bound is half of it.

The parameter that determines which large-scale elements count as sunspots for equation (7) is r^*_s. To avoid redundancy r^*_s should not be smaller than r^*_f. As discussed above, there are also physical reasons to impose this lower limit because r^*_f is supposed to represent the size where magnetic structures have a brightness comparable to the quiet-Sun. We set $r^*_s = 1$ Mm because it is roughly consistent with the smallest area value quoted in the daily sunspot observation record.

The quiet-Sun irradiance, S_Q, is quite well constrained by observations. It is, however, instrument and data-set dependent. For the composite data-set `d41_61_0702` provided by Physikalisch-Meteorologisches Observatorium Davos (hereafter, PMOD composite) the quoted average of the value at each observed solar minima is: 1365.502 ± 0.433 W m^{-2} For the purpose of validating the model we explore the range: 1365–1366 W m^{-2}.

As explained above, we investigate the possibility that the quiet-Sun irradiance has a linear trend over the period 1978–2007. There is some evidence that a downtrend is present in the PMOD composite; the quoted differences between the average quiet-Sun irradiance and that observed at each minima are: 0.067, 0.055, -0.122 W m^{-2}. It is possible that the underlying trend is more complicated than a simple linear one (e.g., Tapping et al. 2007). However, for the purposes of showing that our model could detect such a trend, we restrict ourselves to the linear case. For the intercept, $S_{Q,c}$, we explore the above range for the constant quiet-Sun irradiance, S_Q: 1365–1366 W m^{-2}. For the gradient, $S_{Q,m}$, we explore the range: [-1, 1] W m^{-2} over the 1978–2007 time interval. This corresponds to [-0.0354, 0.0354] W m^{-2} per year and encompasses the variation quoted for the PMOD composite.

The threshold radius for the collection of small-scale flux elements associated with each active region, r^*_t, is fixed. The choice of its value is less crucial than that for the threshold radius for the large-scale elements. It is primarily computational and determines the size at which our algorithm stops tracking the collection of small-scale flux elements. The only constraint is that r^*_t should be small enough that it has a negligible impact on the irradiance. We set $r^*_t = 25$ km.

3. DETERMINING MODEL PARAMETERS WITH A GENETIC ALGORITHM

Our task is now to choose numerical values for the various free parameters of the model in such as way as to provide as close a fit as possible to the TSI and daily sunspot area observations in the 1978–2007 time interval. This represents, at first glance, a classical optimization problem, which can be formulated as a least-squares minimization of the summed day-to-day squared difference between the observed time series and that produced by our fragmentation/erosion model.

For a number of reasons, this optimization problem turns out to be far from straightforward. First, the procedure used to inject active regions near the limb and on the far side of the Sun (see § 2.1) introduces an element of stochasticity into the simulation. While we expect to inject, on average, the proper number of active regions, we certainly cannot expect that they are injected on the correct day, or at the correct position. Second, the process of active region decay, as we model it (§ 2.2) also involves explicit stochastic factors, including the determination of the relative sizes of fragment, and the number of times over which fragmentation takes place. This already guarantees that our synthetic TSI (sunspot area) time series will never be able to match exactly the observed TSI (sunspot area). However, this is not critical for our purpose, since the model should still be able to reproduce adequately the mean variations on timescales of the order of active region lifetime and longer. The expected misfits between the observed and synthetic time series

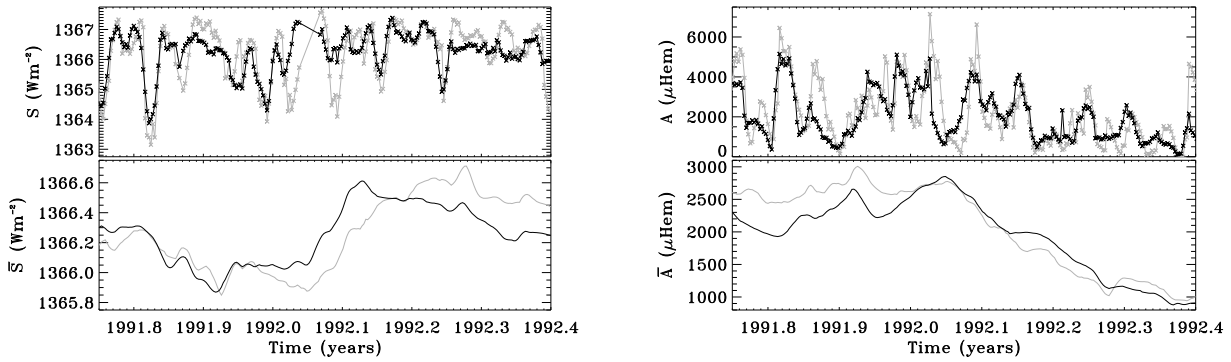

FIG. 3.— The unsmoothed and smoothed time series of the total solar irradiance and the daily sunspot area. In each plot the black curve corresponds to the observations and the grey curve corresponds to the results from a representative model with parameter settings: $p_{frag} = 0.89$, $r^*_{erode} = 105$ km, $\eta = 1$, $\lambda = 0.0115$, $\alpha_{fac} = 0.069$, and $S_Q = 1365.4$ W m^{-2}. On each side, the lower panel shows the 81-day box-car average and the upper panel shows unsmoothed (raw) results (each data point is marked with a cross). *Left:* Total solar irradiance as function of time. *Right:* Daily sunspot area as a function of time.

pose a serious obstacle toward optimizing the model parameters: is a given synthetic time series showing a strong misfit with observations because the underlying model uses intrinsically sub-optimal parameter values, or because of an unfavorable realisation of the active regions that emerge on the far side and/or the fragmentation/erosion sequence? We have no way to know on the basis of the mean-squared residuals alone.

To circumvent this issue we take a box-car average of the irradiance and sunspot area. The box-size has a maximum length of 81 days (roughly the duration of a three solar rotation periods). The number of days in the averaging kernel for the irradiance must vary because there are days where no irradiance data was recorded. Identical smoothing kernels are used to smooth both the observed and modelled time series to ensure a consistent comparison. The smoothing kernels for the irradiance and sunspot area are different because area data is available every day, although it may be zero if no spots are present.

Figure 3 shows the total solar irradiance (*left panel*) and the daily sunspot area (*right*) as a function of time for both the observations and a representative model. On each side, the upper panel shows the unsmoothed values. At certain times the discrepancy between the model and the observations can be relatively large. For example, there will be times when an active region emerges on the far side when none were actually present. This results in a large trough (peak) in the synthetic irradiance (sunspot area) and, thus, a large discrepancy between the synthetic and observed value. It is evident in the lower panels of Figure 3 that the discrepancy between the model and observations in the smoothed case is substantially less than that in the unsmoothed case (note the difference in vertical scaling for the two panels). Indeed, the average squared difference between the smoothed model and the observations is reduced by more than a factor of ten in comparison to the unsmoothed case (for this particular example over the displayed time interval).

We define the fitness of a particular model by $\mathcal{F} = 1/(\nu\chi^2_S\chi^2_A)$, where $\nu = N - n$ is the number of degrees of freedom, with n being the number of model parameters and N the number of days over which the comparison is made, and the normalised sum of the squared residuals for the irradiance is defined as

$$\chi^2_S = \frac{\sum_{i=1}^{N} \left(\bar{S} - \bar{S}_{obs}\right)^2}{\nu \left(\max(\bar{S}_{obs}) - \min(\bar{S}_{obs})\right)^2}\,, \tag{11}$$

where $\bar{S}_{obs}$ and $\bar{S}$ are the box-car averages of observed and modelled irradiance, respectively. The unsmoothed observed irradiance data, S_{obs}, is the constructed composite data-set `d41_61_0702` provided by Physikalisch-Meteorologisches Observatorium Davos (Fröhlich 2000, 2006) and $\max(\bar{S}_{obs})$ and $\min(\bar{S}_{obs})$ refer to the extrema obtained by the smoothed observations during the 1978–2007 time interval. The definition for the normalised sum of the squared residuals for the daily sunspot area, χ^2_A, is equivalent to Equation (11) with area replacing irradiance, and the unsmoothed observed sunspot area time series, A_{obs}, is obtained from the RGO and USAF/NOAA measurements as described in § 2.1.

We start the simulations 500 days prior to the first irradiance observation to allow the ensemble of magnetic structures to build up to a natural state. We end the calculation on the day of the last irradiance observation in the PMOD composite. The comparison of the irradiance and sunspot area time series are made over the same time interval, which spans the first and last irradiance observations (i.e., 1978–2007). In the above definition of χ^2_S each data point has the same weight. This is reasonable because error estimates for the composite irradiance data (and sunspot area) are not tabulated. It is less than ideal because one would expect the quality of the observations to vary with time.

Initial attempts to determine the best fit to the irradiance data alone (i.e., to minimise χ^2_S) resulted in very poor fits to the sunspot area observations. For this reason we decided to fit the irradiance and sunspot area observations simultaneously. In multi-objective optimization problems such as this there are many ways to define fitness $\mathcal{F}$. We have chosen the above definition, based on the product of reciprocals of χ^2, because it is simple and it reliably produces high quality fits to both the irradiance and sunspot area as we will demonstrate.

In part because of the stochastic elements present in the model, we already expect the maximization of $\mathcal{F}$ with respect to model parameter values to yield a strongly multimodal optimization problem. For this reason, we choose

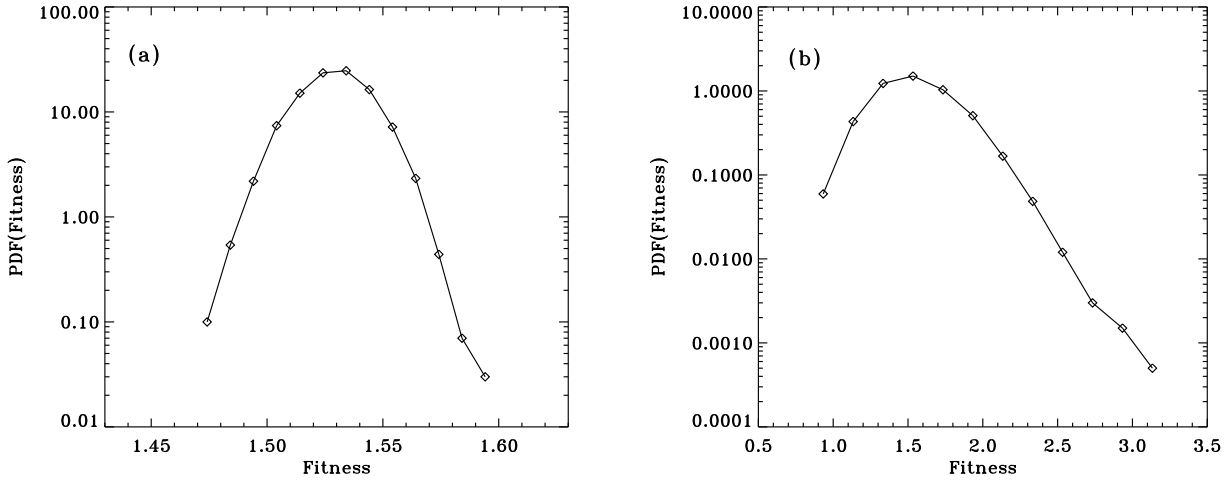

FIG. 4.— Probability density function of fitness $\mathcal{F}$ for 10^4 repeated fitness evaluations for the model with parameter settings: $p_{frag} = 0.89$, $r^*_{erode} = 105$ km, $\eta = 1$, $\lambda = 0.0115$, $\alpha_{fac} = 0.069$, and $S_Q = 1365.4$ W m^{-2}. (a) The case where the sequence of active regions that emerge on the far side is identical for repeated fitness evaluations (with a bin-size of 0.01). (b) The case where a different far side realisation is chosen for each fitness evaluation (with a bin-size of 0.2). The definition of the PDF here is equivalent to eq. (1), except fitness replaces area. The diamonds represent the PDF and the curve is to guide the eye.

to tackle it with genetic forward modelling (Charbonneau et al. 1998), an optimization scheme based on a genetic algorithm (Holland 1975; Goldberg 1989; Mitchell 1996), specifically version 1.2 of the `PIKAIA` code (Charbonneau 1995; Charbonneau & Knapp 1995; Charbonneau 2002).

In a nutshell, the idea is to produce an initial population of models with randomly chosen parameter values. Each model produces a synthetic TSI and sunspot area time series that are used to assign it with a fitness value $\mathcal{F}$. The best fitting (or perhaps more accurately at this stage, the least misfitting) models are extracted from the initial set, and used to breed a new generation of models. This is done by encoding the numerical values of the model's defining parameters as a character string, then genetically inspired operators combine spliced fragments of the defining string of two parents, with occasional random mutation of string elements, into two new strings defining a pair of offspring which replace their parents in the population. Repetition of this evaluate/select/breed/replace procedure over many generations leads to a gradual increase in the overall fitness of population members. Empirical evidence for the power of this evolutionary approach to multimodal parameter optimization problems is overwhelming (e.g., Ackley 1987; Goldberg 1989; Davis 1991). Theoretical analysis has demonstrated that this power is achieved through intrinsic parallelism, whereby the fitness of various substrings are tested simultaneously across the whole evolving population, with the occurrence frequency of advantageous substrings continuously increasing, and the combination of these advantageous substrings through breeding continuously produces ever fitter models (Holland 1975).

The stochastic components of our model introduce difficulties here as well. Consider two model runs using the same parameter set, but distinct random number realizations; both are runs of the same model, but may well end up being assigned significantly different fitness values, even under the smoothing procedure introduced above. To accommodate this difficulty we have modified the elitism option in `PIKAIA` to carry over the three best solutions of each generation into the subsequent one, rather than just the single best solution, as normally done under internal control parameter setting `ctrl(11)=1`. Likewise, adaptive adjustment of the mutation rate must be done based on metric distance in parameter space (`ctrl(5)=3` in version 1.2 of `PIKAIA`), rather than fitness differential, to avoid erratic variations of the mutation rate slowing down convergence.

Figure 4 displays how different stochastic components of our model contribute to the variation in fitness values. It shows the PDF of fitness $\mathcal{F}$ over 10^4 repeated evaluations with identical parameter settings for two cases: (1) the sequence of active regions that emerge on the far side is identical for repeated fitness evaluations (Fig. 4(a)); and (2) a different far side realisation is used for each fitness evaluation (Fig. 4(b)). Both cases appear to have similar Gaussian-like profiles with comparable average values: 1.53 and 1.66, respectively; however, the standard deviations are significantly different: 0.0157 and 0.266, respectively. This indicates that the variation introduced by the active regions that emerge on the far side dominates that introduced by the fragmentation/erosion model. Thus, any information that can determined about active regions that emerge on the far side (e.g., size and position), say from helioseismic far side imaging (e.g., Lindsey & Braun 2000; González-Hernández et al. 2007), could be highly valuable for constraining future TSI models of this type. We have also examined the individual variations in the values for χ^2_S and χ^2_A for 10^4 repeated evaluations with identical parameter settings. There is slightly more variation associated with χ^2_A. The relative variations should be similar because χ^2_S and χ^2_A are not independent and are affected by the same set of stochastic components. The variations associated with χ^2_S are reduced relative to those associated with χ^2_A because the modelled irradiance is also sensitive to the non-stochastic factors, such as the decay rate for small-scale elements, the faculae intensity contrast, and the quiet-Sun irradiance. For each fitness evaluation choosing a different realisation of the active regions that emerge on the far side makes the problem more difficult. However, we believe that finding the set of model parameters that provide the best fit to the observations over an ensemble of far side realisations is

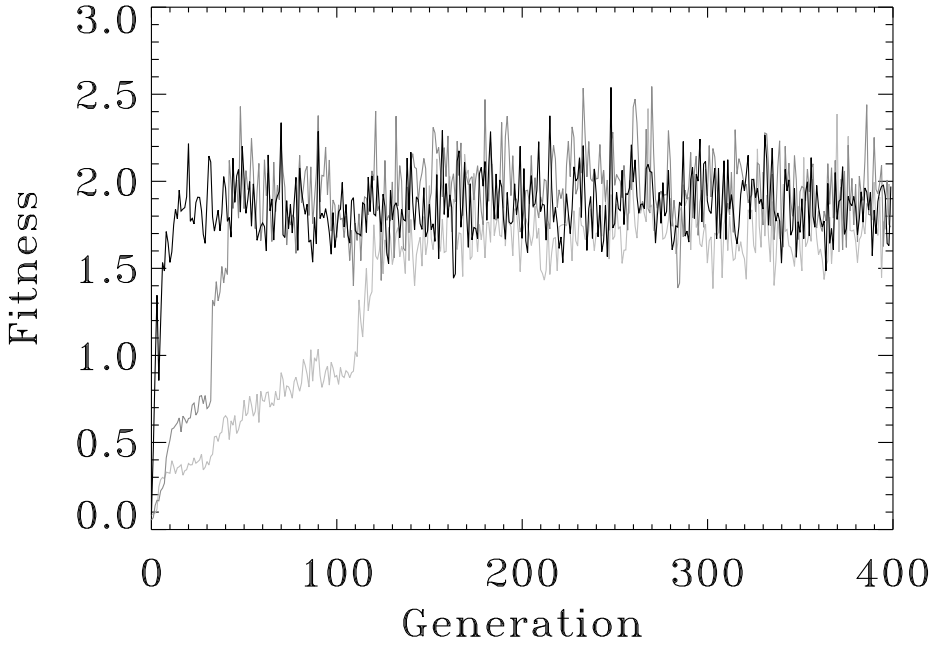

FIG. 5.— Fitness value of the best solution as a function of generation for a sample of three individual optimization runs. Each curve corresponds to a different run with a different initial population of randomly chosen parameter values.

TABLE 2
OPTIMAL PARAMETER VALUES

Symbol	Description	Optimal value
p_{frag}	Fragmentation probability	0.84 ± 0.05
r^*_{erode}	Maximum radial thickness of the eroded annulus	(171 ± 61) km
η	Area conversion efficiency	0.95 ± 0.04
λ	Decay rate for small-scale elements	0.0109 ± 0.0007
α_{fac}	Faculae intensity contrast	0.068 ± 0.003
S_Q	Quiet-Sun irradiance	(1365.42 ± 0.03) W m^{-2}

more worthwhile than doing the same for a fixed far side realisation, because we have no way to know if a particular realisation is accurate.

Figure 5 shows the evolution of the fitness value of the best solution as a function of generation for a sample of three individual optimization runs over 400 generations. These show some features typical of convergence curves produced by genetic algorithms (GAs). Convergence is quite rapid at first, but then levels off, further improvements occur as sudden increases in fitness punctuate phases during which fitness remains essentially constant. This usually reflects the presence of secondary extrema in parameter space, on which the evolving population of trial solutions remains temporarily "stuck" until a favorable mutation or breeding event leads to the discovery of an another, higher extremum. The three sample convergence curves plotted in Figure 5 are markedly different from one another in the first few tens of generations, reflecting the stochastic nature of the genetic operators, as well as the randomness of the initial population. Yet, after a few hundred generations, all solutions have converged to a similar fitness value, here approximately 1.9, offering hope — but never strictly proving — that this represents the true, global maximum in the fitness landscape. On the other hand, significant generation-to-generation fluctuations in fitness are not typical of classical GA-based optimisation using elitism. They occur here because of the additional stochasticity introduced at the level of the fragmentation/erosion model, and in particular with varying realisations of the active regions that emerge on the far side, as discussed previously.

3.1. *Results*

The genetic algorithm, `PIKAIA`, was run over 400 generations starting from 108 different initial populations of randomly chosen parameter values. The populations consist of fifty individuals. The fitness of each optimal solution found by `PIKAIA` was evaluated 100 times and the average was used to rank the final solutions. We consider the best fifty solutions to be truly optimal; these have average fitness values that lie in the range: 1.503 to 1.665. The purpose of this post-evaluation step is to ensure that the optimal solutions are indeed optimal and not solutions that got lucky during the final generations due to the aforementioned stochasticity. Table 2 summarises the parameters values for the optimal sets of solutions – for each parameter the third column shows the mean value plus or minus its standard deviation.

Both of the parameters that control the decay of large-scale flux elements, p_{frag} and r^*_{erode}, have optimal values

that are a reasonable distance from the limits of their respective allowed ranges (see Table 1). This indicates that the imposed ranges for these parameters were sensible and that the fragmentation/erosion scheme adopted here is capturing the average behaviour of active region decay with some success, despite reducing this complex process to a two-parameter model. The optimal values for p_{frag} lie nearer to the largest allowed value with a fairly small standard deviation. For r^*_{erode} the mean optimal value is nearer to the smallest allowed value. The parameter r^*_{erode} represents the upper-bound for the erosion rate, if we take half of its mean value as a representative setting for r_{erode} the corresponding area decay rate (cf. eq [2]) for a flux element with a radius of 10 Mm (area of 103 μHem) is 1.76 μHem/day and for a flux element with a radius of 40 Mm (area of 1652 μHem) it is 7.05 μHem/day. These values lie at the low-end of the spectrum determined by Martinez Pillet et al. (1993); however, the standard deviation of the optimal values for r^*_{erode} is relatively large, thus, greater decay rates will be produced by the best fit models. Figure 6(a) shows a scatterplot of the optimal pairs of values for p_{frag} and r^*_{erode}. Clearly, there is significant correlation between these two parameters over the displayed range, with high fragmentation rates corresponding to lower erosion rates and vice versa. This indicates that there are multiple pathways that will lead to the same overall decay rate (on average). For example, larger fragmentation rates result in a more fragments and, therefore, a greater combined surface area that is subject to boundary erosion. Consequently, when the fragmentation rate is high, smaller erosion rates are required to maintain a fixed overall decay rate. Conversely, when the fragmentation rate is low, larger erosion rates are required.

Of the various parameters, the quiet-Sun irradiance, S_Q, has the smallest range of optimal values in relative terms. The mean value agrees very well with the average of the value at each observed solar minima quoted for the PMOD composite: 1365.502 ± 0.433 W m^{-2}.

The range of optimal values for the decay rate of small-scale elements, λ, is quite small. Its average value corresponds to an e-folding time-scale of roughly 92 days, which is large but not implausible for aggregates of small-scale flux elements like faculae (plage) and network; for example, Foukal (1998) quotes faculae lifetimes of tens of days, up to about fifty days for large faculae. The faculae intensity contrast, α_{fac}, also has a fairly small range of optimal values; although its mean value is nearly two times the empirical value of 0.036 (Chapman & Meyer 1986). Likewise, the optimal range for the area conversion efficiency η is quite small, with a mean value close to the largest allowed value. Broadly speaking, to accurately fit the observed irradiance it is apparent that the contribution from the bright, small-scale flux elements, requires substantial enhancement to counteract the darkening induced by the model sunspots: almost complete conversion of the area from large- to small-scale elements ($\eta \approx 1$), long lifetimes for the small-scale elements (small λ), and large intensity contrast α_{fac}. It is therefore unsurprising that there are some, albeit weak, correlations between these three parameters. For example, Figure 6(b) shows that there is a weak correlation between optimal values for η and λ. The exact nature of this relationship is difficult to characterise, but there may be a threshold where, for a fixed value of η, there is an upper bound on the optimal value for λ that increases with η as shown in Figure 6(b). Similar relationships seem to exist between the pair η and α_{fac} and the pair λ and α_{fac} (not shown).

The enhanced contribution from the bright, small-scale flux elements – that is required for our simulation results to match the observed irradiance – suggests that our model is neglecting a significant source of excess brightness. Regarding the contribution from surface magnetism, the obvious candidates are: (1) the contribution from the magnetic network, and (2) the injection of small-scale elements supplied by ephemeral regions, which is mostly independent of active region decay and is not incorporated into our model at present; this could contribute significantly to both the network and faculae. However, we cannot rule out possibilities unrelated to surface magnetism.

Various results from one of the best fit models are presented in Figures 7 and 8. Figure 7 displays the 81-day box-car averaged irradiance and daily sunspot area. It shows that the model does quite well at reproducing the observations over nearly three solar cycles. The fit to the daily sunspot area is particularly impressive. There are important differences between the observed irradiance time series and that produced by the model: (1) the model consistently underestimates the irradiance during the ascending phase of a cycle (for both observed instances); (2) for the period 2000–2003, around solar maximum, some of the large peaks and troughs in the TSI are not well replicated by the model; (3) the model overestimates the irradiance during the early stages of the observation record: 1980 and before; and (4) during the period approaching the recent solar minimum, from 2003 and beyond, the observed irradiance is noticeably smaller than that predicted by the model. Point (4) is possibly associated with point (2) and it may indicate the presence of an underlying long-term downtrend in the PMOD composite, because the observed irradiance for 2006 and beyond is substantially below that during the previous two minima. It is not entirely clear if the cause of this is solar in nature, or if it is due to instrument degradation, or if it is an artifact of the composite construction process. Other than an underlying long term trend in the irradiance, it is difficult to identify an unmodelled mechanism that could explain the discrepancy referred to in point (4). It is worth noting that a similar discrepancy is also evident when comparing the PMOD composite TSI to other activity indices, such as a 3-component proxy model based on a MgII index (see e.g., Fröhlich 2006, and references therein). In each of these four time periods, when there is a noticeable difference between the synthetic and observed irradiance, it is significant that the observed and modelled sunspot area times series do not suffer the same disagreement.

In Figures 7(b) and (c) we examine the average discrepancy between the 81-day averaged synthetic and observed time series. The vertical grey lines correspond to times where the discrepancy changes sign; these help to identify time periods where the discrepancy changes sign frequently or is persistently of the same sign. For the daily sunspot area (Fig. 7(c)) it is evident that the sign of the discrepancy changes sign frequently during times of high activity, with

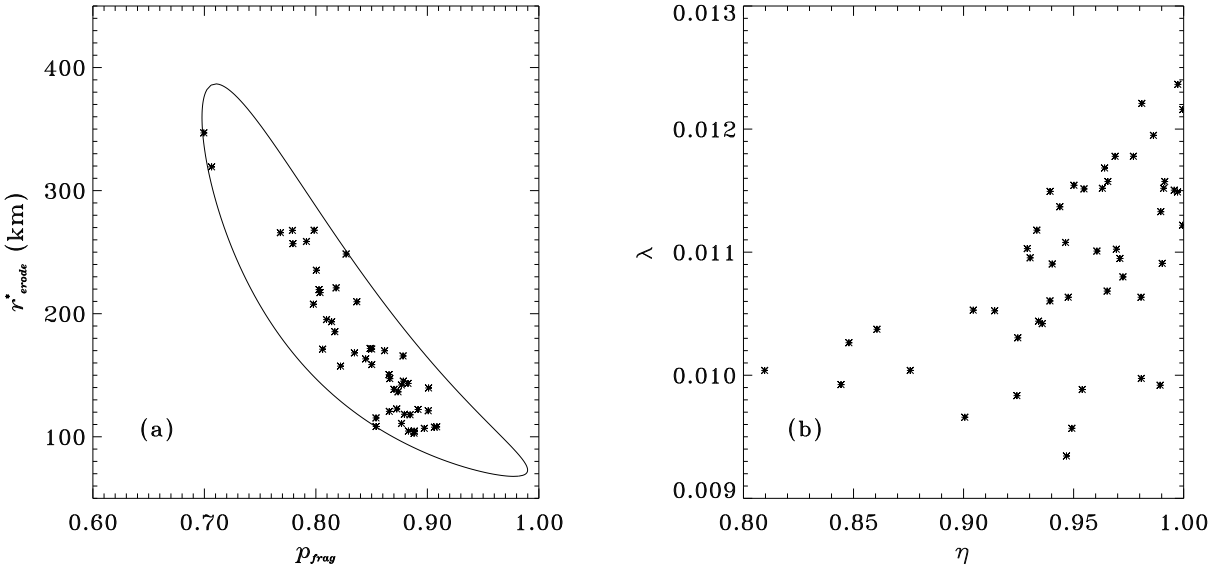

FIG. 6.— Scatterplot of pairs of optimal parameters. In each panel, each pair of points corresponds to the optimal parameter values that were found by PIKAIA for one of the fifty best solutions. (a) Optimal values for p_{frag} and r^*_{erode} are plotted. The "ellipse" is used to account for the correlation between these parameters when extracting values from their optimal distributions, see § 4. (b) Optimal values for η and λ are plotted.

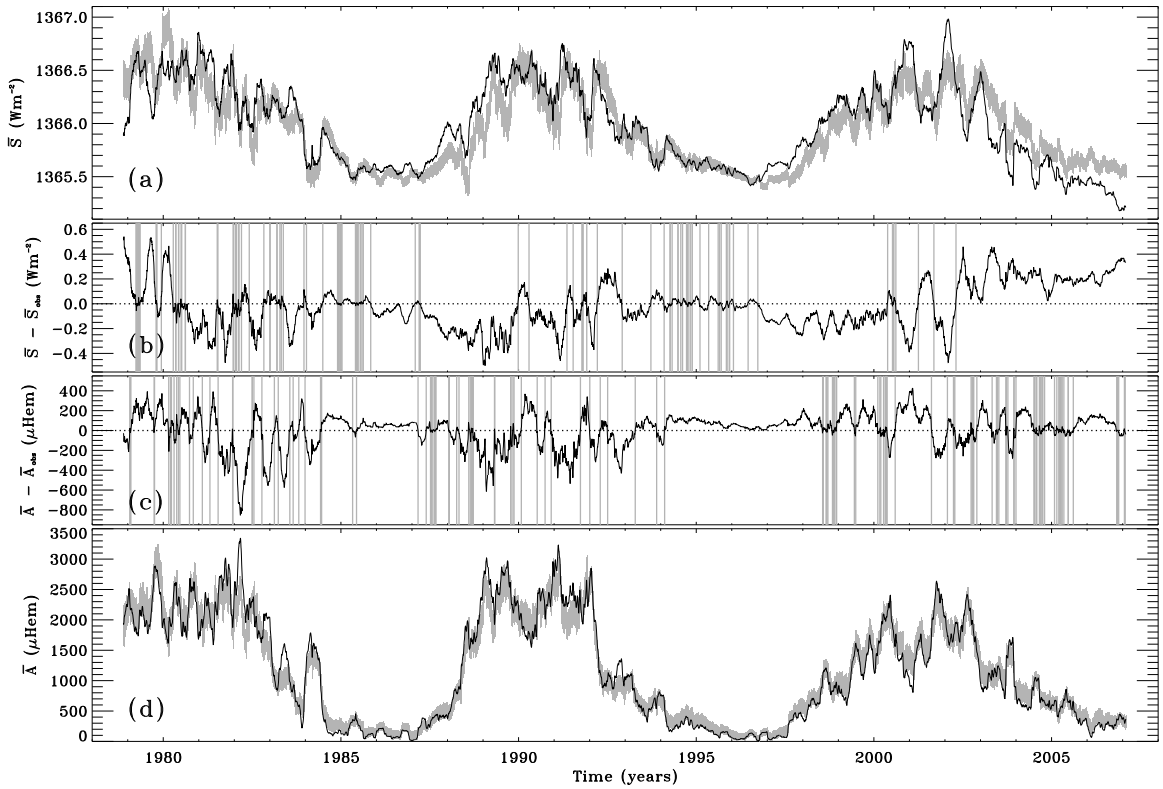

FIG. 7.— (a) Smoothed total solar irradiance, $\bar{S}$, as a function of time. The black curve is the 81-day box-car average of the observed irradiance time series. For the best fit model, with parameter settings: $p_{frag} = 0.89$, $r^*_{erode} = 105$ km, $\eta = 1$, $\lambda = 0.0115$, $\alpha_{fac} = 0.069$, and $S_Q = 1365.4$ W m^{-2}, the irradiance time series was calculated 1000 times, each with different far side realisations and different random decay sequences. On each day, the mean and standard deviation of the corresponding 81-day box-car average was determined. The grey shaded regions correspond to the mean plus one standard deviation and the mean minus one standard deviation. (b) The average discrepancy between the synthetic and observed irradiance, $\bar{S} - \bar{S}_{obs}$, as a function of time. For each of the 1000 realisations of the best fit model, the discrepancy between the synthetic and the observed irradiance is determined. The black curve corresponds to the mean value of the discrepancy on each day. The vertical grey lines indicate times where the discrepancy changes sign. (c) and (d) Same as (b) and (a), respectively, except for the smoothed daily sunspot area, $\bar{A}$.

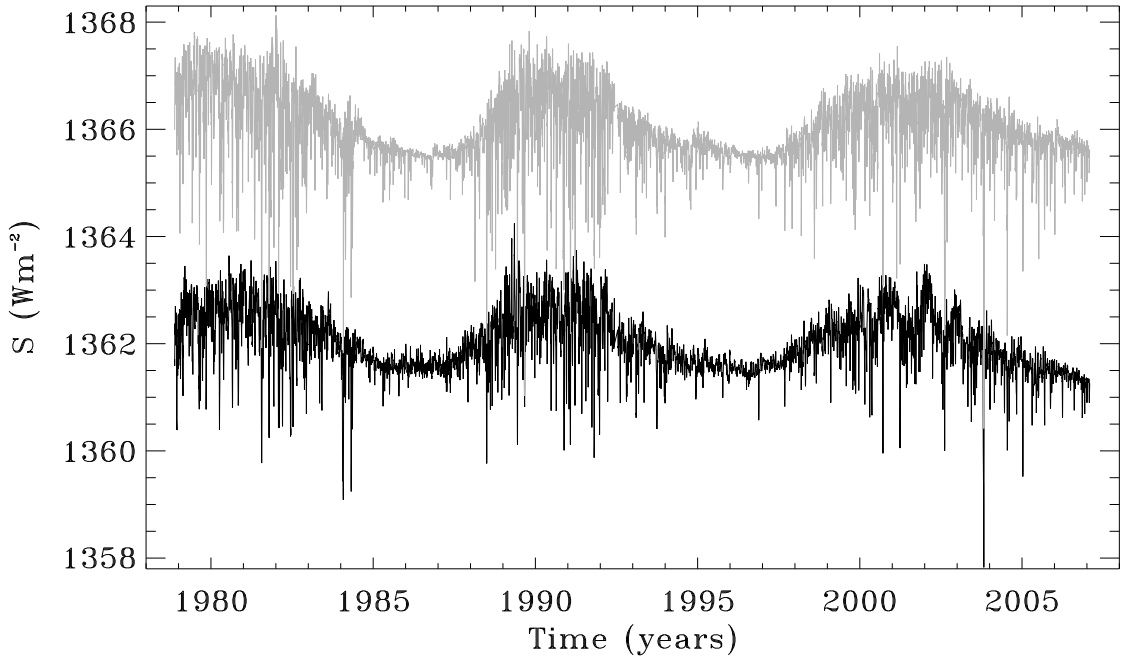

FIG. 8.— Total solar irradiance as a function of time. The black curve is the observed irradiance minus 4 W m^{-2}. The grey curve is the synthetic irradiance for one representative realisation of the best fit model with parameter settings: $p_{frag} = 0.89$, $r^*_{erode} = 105$ km, $\eta = 1$, $\lambda = 0.0115$, $\alpha_{fac} = 0.069$, and $S_Q = 1365.4$ W m^{-2} (the same case as Fig. 7).

especially high frequency alternations occurring during the ascending phase of the cycles (where the irradiance fits are poor). At solar minimum and at the end of the descending phase, the model is consistently overestimating the sunspot area (this is not immediately obvious in Fig. 7(d)); although persistent for a long period, the discrepancy is relatively small in magnitude and does not appear to manifest itself in a predictable way in the corresponding irradiance. This area discrepancy may be due to an inadequacy in the procedure for injecting synthetic active regions that emerge on the far side and near the limb. For the irradiance discrepancy (Fig. 7(b)), two time intervals are particularly conspicuous: (1) the ascending phases when the model consistently underestimates the irradiance; and (2) 2003–2007 when the model consistently overestimates the irradiance. In both of these cases there is no accompanying period of a persistent sign in the area discrepancy. The underestimates in the TSI during the ascending phase of a cycle are preceded by overestimates of the area (at solar minimum); however, it is unlikely that there is a causal relationship between these as there is no plausible mechanism for a link over such a long time. More likely, the underestimates of the irradiance during the ascending phases by the model, which are significant in both magnitude and duration, are indicative of a neglected source of irradiance enhancement during these phases. A possible source is the contribution to faculae and network from the small-scale flux elements supplied by ephemeral regions, which start emerging 2-3 years before the first sunspots of a given cycle (e.g., Martin & Harvey 1979; Solanki et al. 2002; Preminger & Walton 2005).

Figure 8 shows a comparison between the unsmoothed irradiance observations and model results. The general behaviour of the synthetic time series is very similar to the observations. The differences that were apparent in Figure 7 are again apparent here: the model overestimates the irradiance for periods before 1980 and after 2003; some of the broad, large observed peaks and troughs during the solar maximum around 2000–2003 are absent from the model results; and the model underestimates the irradiance during the ascending phase of a cycle. Additionally, it is evident that the observed irradiance fluctuates more than the synthetic irradiance at times around solar minimum. This may be related to deficit in the synthetic irradiance during the ascending phase of a cycle and could be caused by the aforementioned components missing from our model: the magnetic network and the injection of small-scale elements supplied by ephemeral regions that could feed network and faculae production (independent of sunspot decay).

4. DETERMINING THE TOTAL SOLAR IRRADIANCE FOR 1874–1978

Judging from the results presented in the previous section the model does a very good job of reproducing the observed irradiance and sunspot area over the period 1978–2007. In this section we use the distribution of optimal values for each parameter to estimate the total solar irradiance over the interval 1874–1978, taking observed emerging active regions as model input. This period pre-dates the recording of irradiance observations. We could extend these reconstructions further by employing the sunspot number record; this would require determining an accurate relationship between the active region emergence rate and sunspot number, which is not undertaken at this time. The basic idea is to choose a value for each model parameter by drawing from the distribution of its optimal values, as determined in Section 3 by calibrating the model over the period 1978–2007.

In practice, for each parameter the distribution of optimal values is decomposed into ten equally sized bins (ranging from the minimum to the maximum of the optimal range). The PDF is then determined along with the corresponding

cumulative distribution function, CDF. The CDF can then be used to transform a random number, uniformly distributed over the range 0–1, into one that is statistically consistent with the optimal distribution – the central value of the selected bin is then used for the parameter value. This approach assumes that each parameter is independent of the other parameters (i.e., not correlated); however, as was discussed in Section 3, this assumption is not entirely correct (see Fig. 6).

The most significant parameter correlation is between p_{frag} and r^*_{erode} (Fig. 6(a)). Other correlations exist (e.g., Fig. 6(b)) but these are weaker than the correlation between p_{frag} and r^*_{erode} and, therefore, are not considered here. To account for the correlation between p_{frag} and r^*_{erode}, an ellipse is constructed to enclose all of the pairs of optimal values for p_{frag} and r^*_{erode} in the encoded (p_{frag}, x)-space, where $\log_{10} r^*_{erode} = 5 + 1.30103x$ and $0 \leq x \leq 1$. This enclosing ellipse is displayed in Figure 6(a), but does not appear elliptical in physical space as shown in the figure. A value for p_{frag} is calculated from its distribution of optimal values (as described above). For that value of p_{frag} the enclosing ellipse is used to determine the allowed range for the corresponding value of r^*_{erode}. The value of r^*_{erode} is calculated from its distribution of optimal values (as described above), but only values that lie within the allowed range are accepted.

Once a set of parameters is determined we calculate the corresponding irradiance and sunspot area for each day over the interval 1874–2007. We repeat this process 1000 times. Thus, on each day there will be 1000 different sunspot area and irradiance values. We then determine the mean and standard deviation of the distribution of 81-day box-car averaged values on each day. In Figure 9 we plot in grey the region spanned by the mean value minus one standard deviation and the mean plus one standard deviation over the time interval used for the model calibration. Figure 9 also displays the 81-day box-car average of the observed time series. It is evident that the model results do very well at bracketing the observations. In general, the important aspects of Figure 9 are similar to those of Figure 7: the model reproduces the sunspot area better than it reproduces the irradiance; the model overestimates the irradiance for periods before 1980 and after 2003; some of the broad, large observed peaks and troughs during solar maximum around 2000–2003 are absent from the model results; the model underestimates the irradiance during the ascending phase of a cycle; and the model overestimates the sunspot area during solar minimum. The good correspondence between Figures 7 and 9 provides some confidence that the procedure for determining the model parameter settings from their distribution of optimal values is robust. One would expect that the discrepancy between the observed and synthetic time series to be larger in Figure 9 than in Figure 7. Indeed the average value of the unnormalised squared residual differences between the observations and the model predictions is 3% larger for the sunspot area and 50% larger for the irradiance. We consider these changes to be reasonable. The larger change for the irradiance is to be expected because it is controlled by more parameters than the sunspot area and we have neglected some of the weaker parameter correlations.

4.1. *Sunspot area extrapolation*

Figure 10 shows the smoothed daily sunspot area for the 1874–2007 time interval, where parameter settings have been extracted from the distributions of optimal values. In this case observations are available to compare with the model predictions. Qualitatively, the model appears to bracket the observations quite well over the entire time interval. This would suggest that the sunspot decay model is capturing the essential physics and that the procedure for injecting active regions on the far side and near the limb is reasonable. Because the fragmentation/erosion model was calibrated over the 1978–2007 interval we expect the discrepancy between the model and the observations to be larger during the 1874–1978 interval. The average value of the unnormalised squared residual differences between the observations and the model predictions is approximately 70% larger during the 1874–1978 interval than during the 1978–2007 interval. Again, the model overestimates the sunspot area during solar minimum for 1874–1978, but also tends to underestimate it during solar maximum. The increase in residual differences is quite large and may suggest that the two-parameter active region decay model, as calibrated over the 1978–2007 interval, is missing the information required to accurately capture the active region decay process over the 1874–1978 interval. Another possibility is the difference in the two data sets for the active region observations. The model has been calibrated with USAF/NOAA data, whereas the majority of the data for 1874–1978 was collected by the RGO. We have attempted to correct for the change (as described in § 2.1) but this may be insufficient.

The distribution of sunspot areas produced by our model is consistent with a lognormal distribution. This is to be expected because the imposed distribution of (actual and synthetic) emerging active region areas is consistent with a lognormal distribution (as described in § 2.1 and shown in Fig. 2). For the best fit models, it turns out that the synthetic distribution is different from the observed distribution. A detailed investigation of the source of this discrepancy is beyond the scope of this paper, but preliminary numerical experiments show that better agreement between the observed and synthetic distributions can be obtained with lower values of p_{frag} and larger values of r^*_{erode} (than those optimal values determined by `PIKAIA`). The distribution of synthetic sunspot areas is essentially independent of solar cycle phase and is constant from one cycle to the next, which is consistent with the findings of Bogdan, Gilman, Lerche, & Howard (1988) for sunspot umbral areas.

4.2. *Faculae area extrapolation*

Figure 11 shows the synthetic smoothed daily faculae area for the 1874–2007 time interval, where parameter settings have been extracted from the distributions of optimal values. Observations are available for part of this interval: the observed daily plage area from Mount Wilson spectroheliogram data obtained in the K line of Ca II between 1915 and 1984 (see Foukal 1996, for details) and the observed white-light faculae area measured by the Royal Greenwich

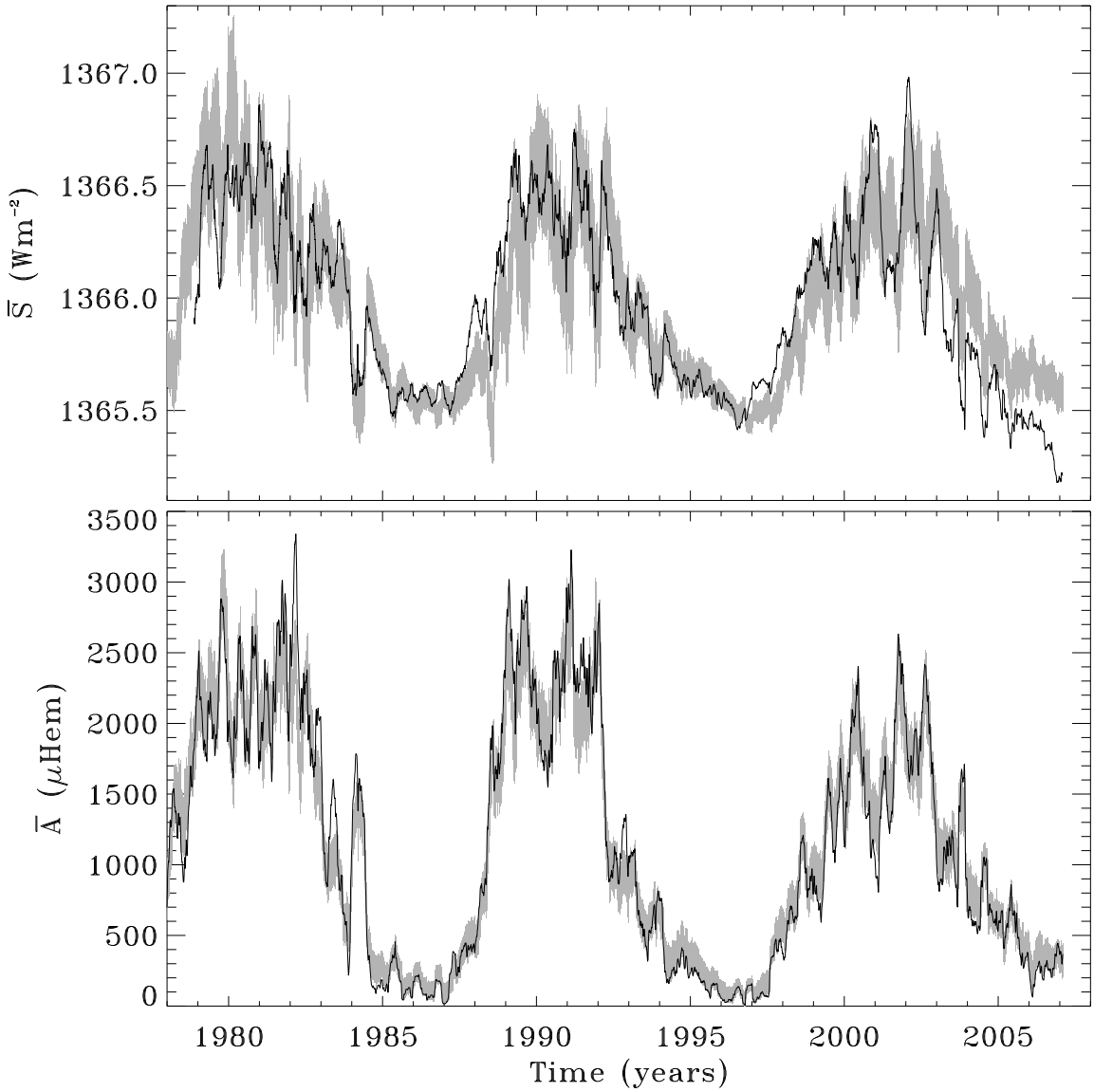

FIG. 9.— *Top:* Smoothed total solar irradiance as a function of time. *Bottom:* Smoothed daily sunspot area as a function of time. In both panels the black curve is the 81-day box-car average of the observed time series. For 1000 realisations, each drawing parameter values from the distribution of optimal values, we calculate the irradiance and daily sunspot area. On each day, the mean and standard deviation of the corresponding 81-day box-car average was determined. The grey shaded areas correspond to the mean plus one standard deviation and the mean minus one standard deviation of the 81-day box-car average.

Observatory over the intervals 1874–1877 and 1886–1976 (see Brown & Evans 1980; Foukal 1993, for details). In Figure 11 we multiply the white-light faculae area by a factor of five. Although faculae and plage represent the same magnetic structure, only at different heights in the solar atmosphere, the white-light faculae areas are roughly ten times smaller than the Ca K plage areas (Foukal 1993). This is mainly due to the higher photometric contrast of Ca K plage (white-light faculae are visible only near the limb). Also Ca K plage data captures large-area, old plages whose contrast is too low to be detected in white-light (Foukal 1998). It is also worth noting that a different morphological definition of plage can alter the observational results significantly; for example, Worden et al. (1998) find plage areas that are nearly three times greater than the Ca K plage areas displayed in Figure 11 because their definition of plage includes the fainter outer regions.

The daily faculae area produced by our model is not affected by contrast or visibility issues. However, our definition of faculae includes all small-scale elements that are produced by the active region decay process which, in reality, would contribute to both faculae/plage and the network (for observational estimates of the network area, see Worden et al. 1998; Foukal & Milano 2001). Therefore, the faculae area time series produced by our model should more closely resemble the observed Ca K plage area time series. In fact, the synthetic time series is consistently less than the Ca K plage area. This further clarifies the cause of the large values for the faculae intensity contrast, α_{fac}, found by `PIKAIA`: large values of η and small values of λ act to increase the synthetic faculae area, but this alone does not produce enough faculae coverage and therefore is insufficient to provide the brightness enhancement necessary to match the observed irradiance. The cycle-to-cycle variation in the amplitude of the peaks of the synthetic smoothed daily faculae area

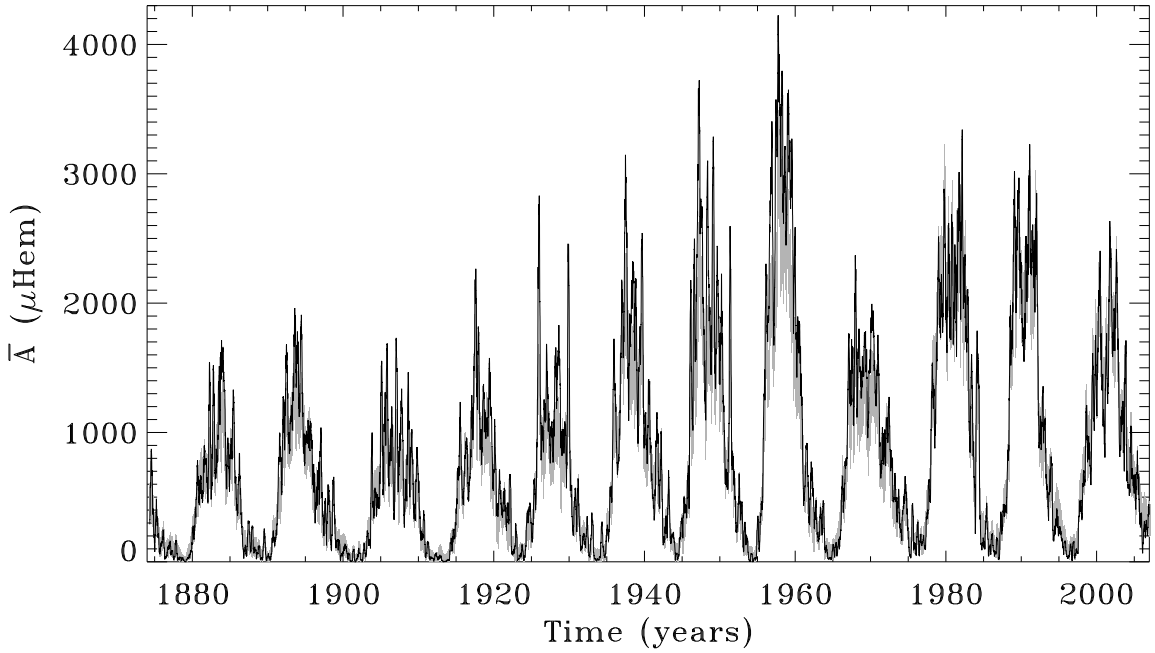

FIG. 10.— Smoothed daily sunspot area as a function of time. This plot is the same as the bottom panel of Figure 9, except that results for the 1874–2007 time interval are displayed.

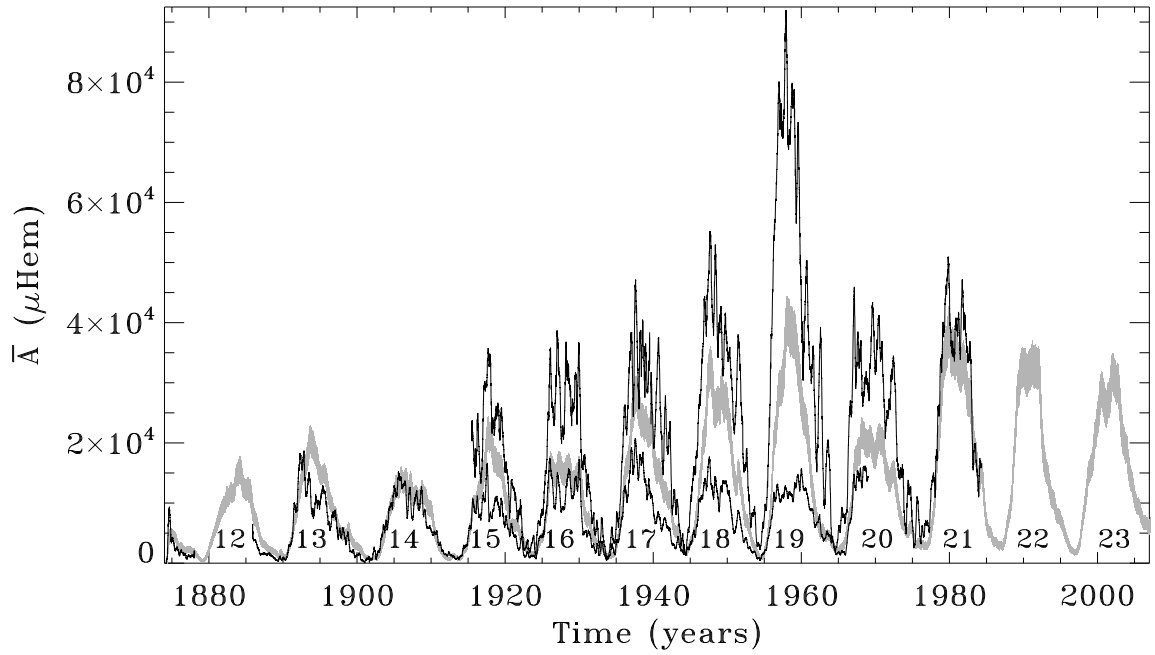

FIG. 11.— Smoothed daily faculae and plage area as a function of time. The upper black curve is the 81-day box-car average of the observed daily plage area data from Mount Wilson spectroheliograms obtained in the K line of Ca II between 1915 and 1984 (Foukal 1996). The lower black curve is five times the 81-day box-car average of the observed white-light faculae area measured by the Royal Greenwich Observatory over the intervals 1874–1877 and 1886–1976 (Brown & Evans 1980; Foukal 1993). For 1000 realisations, each drawing parameter settings from the distribution of optimal values, we calculate the daily faculae area. The grey shaded areas correspond to the mean plus one standard deviation and the mean minus one standard deviation of the 81-day box-car average of the daily faculae area.

tends to closely follow that of the sunspot area (Fig. 10) and to a lesser degree the observed Ca K plage area (although the model faculae area for cycle 16 appears to be lower than expected). For discussion of the unusual behaviour of the cycle-to-cycle variation in the observed white-light faculae area, most notably in cycles 17, 18, and 19, see Brown & Evans (1980) and Foukal (1993, 1996, 1998).

4.3. *Total solar irradiance extrapolation*

Figure 12 shows our total solar irradiance reconstruction for the interval 1874–2007. To compute this reconstruction parameter settings have been extracted from the distributions of optimal values. As one would expect, the behaviour

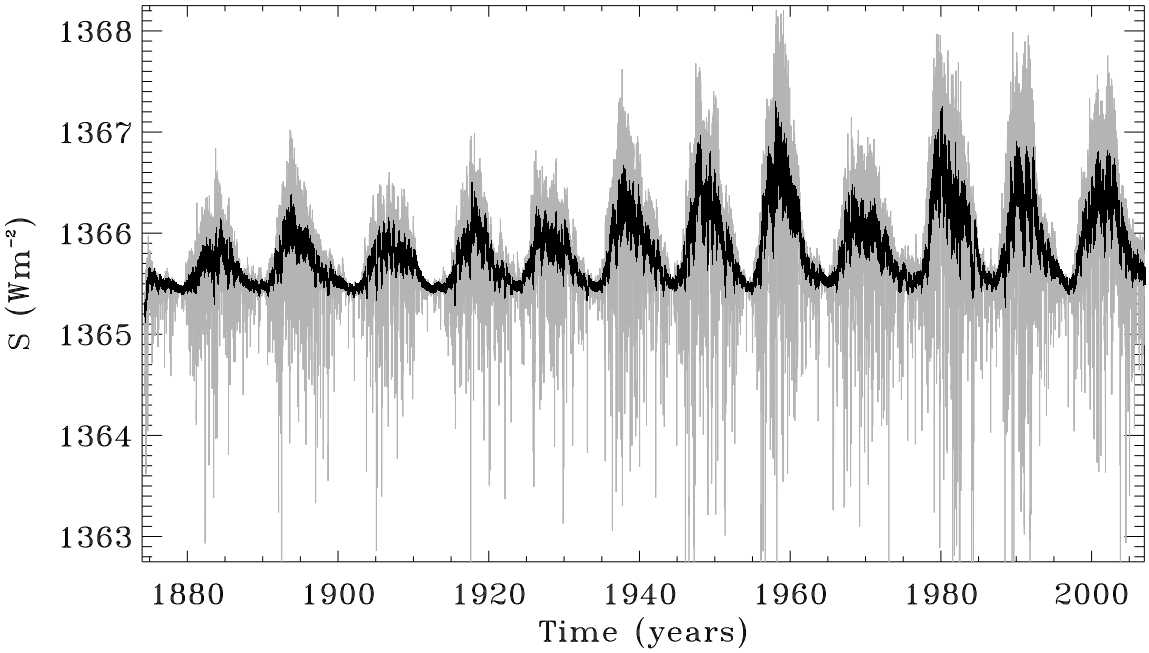

FIG. 12.— Total solar irradiance as function of time for the interval 1874–2007. For 1000 realisations, each drawing from the distribution of optimal parameter values, we calculate the irradiance and its 81-day box-car average. The black shaded area corresponds to the mean plus one standard deviation and the mean minus one standard deviation of the 81-day box-car average. The thin grey curve is one realisation of the modelled irradiance (unsmoothed).

of the irradiance closely tracks that of the active region emergence rate (Fig. 1) and the daily sunspot area (Fig. 10). This is evident, for example, in the variation of the magnitude of the peaks at each solar maximum. It is impossible to judge the accuracy of this irradiance reconstruction. As stated previously, for the daily sunspot area we know that the average value of the unnormalised squared residual differences between the observations and the model predictions is approximately 70% larger during the 1874–1978 interval than during the 1978–2007 interval. The irradiance is subject to more parameters than the sunspot area so the corresponding change in residual differences would almost certainly be greater than 70%. We discussed in detail the limitations of the irradiance model in different parts in the calibration period, 1978-2007; these will also be embodied in the extrapolations displayed in Figure 12.

The strength of this extrapolation is that it is a physical rather than statistical extrapolation. By "physical" we mean that the physical mechanisms through which large- and small-scale magnetic flux elements decay (for example) are universal, in the sense that they do not depend explicitly on time or the level of magnetic activity. One major limitation of this extrapolation is the assumption that the quiet-Sun irradiance, S_Q, is constant. There is evidence in the PMOD composite that S_Q varies during the 1978–2007 interval. Therefore, it is reasonable to assume that such variations are possible over longer time intervals. We could relax the assumption that S_Q is constant over 1874–2007 but it is difficult to determine the long-term variability in S_Q from the 1978–2007 irradiance time series without additional input physics (e.g., Lean, Beer, & Bradley 1995; Solanki & Fligge 1998; Wang, Lean, & Sheeley 2005; Krivova et al. 2007). Recently, Tapping et al. (2007) put forward an irradiance model similar ours in that it is based on the cascade of magnetic flux components from large to small scales. In that model the long term variability in the irradiance is assumed to be caused by a reservoir of subphotospheric magnetic flux that modulates energy flow through the convection zone; the influence of this reservoir is estimated by applying a low-pass filter to the slowly-varying component of the 10.7 cm solar radio flux. The irradiance reconstructions produced by Tapping et al. have solar minimum values that increase substantially from 1900 to 1950 and decrease gradually for later times, in rough correspondence with the cycle-to-cycle variation in solar activity (cf. Fig. 1). For example, their solar minimum values around 1900 are slightly more than 0.5 W m^{-2} below the values shown in Figure 12. The irradiance time series produced by our model (Fig. 12) should be interpreted as the contribution by the rapidly varying (cyclic) component of the irradiance, which may be modulated on longer timescales by a contribution from a slowly varying (secular) component. In the next section we will demonstrate that the rapidly varying component of the irradiance displayed in Figure 12 would not be strongly affected by the presence of a weak, long-term modulation, provided that such a modulation could be characterised as a slowly varying component of the quiet-Sun irradiance, S_Q.

5. A POSSIBLE LINEAR TREND DURING THE PERIOD 1978–2007

In this section, we explore the possibility that the quiet-Sun irradiance, S_Q, varies gradually over longer timescales. The purpose of this exercise is not to speculate on the cause of any trend, but to simply demonstrate that it is possible for our algorithm to detect such a trend if one is present in the observational data. For the PMOD composite a downtrend is apparent in the quiet-Sun irradiance observed at each solar minimum. For simplicity, we restrict our attention to a linear trend (cf. eq. [10]) and repeat the process outlined in § 3 to determine the optimal set of model

TABLE 3
Optimal parameter values with a linear trend in S_Q

Symbol	Description	Optimal value
p_{frag}	Fragmentation probability	0.85 ± 0.04
r^*_{erode}	Maximum radial thickness of the eroded annulus	(166 ± 54) km
η	Area conversion efficiency	0.97 ± 0.03
λ	Decay rate for small-scale elements	0.0115 ± 0.0007
α_{fac}	Faculae intensity contrast	0.069 ± 0.003
$S_{Q,c}$	Quiet-Sun irradiance (vertical intercept)	(1365.54 ± 0.05) W m^{-2}
$S_{Q,m}$	Gradient of Quiet-Sun irradiance linear trend	(-0.19 ± 0.08) W m^{-2} over 1978–2007 (-0.0069 ± 0.0027) W m^{-2} per year

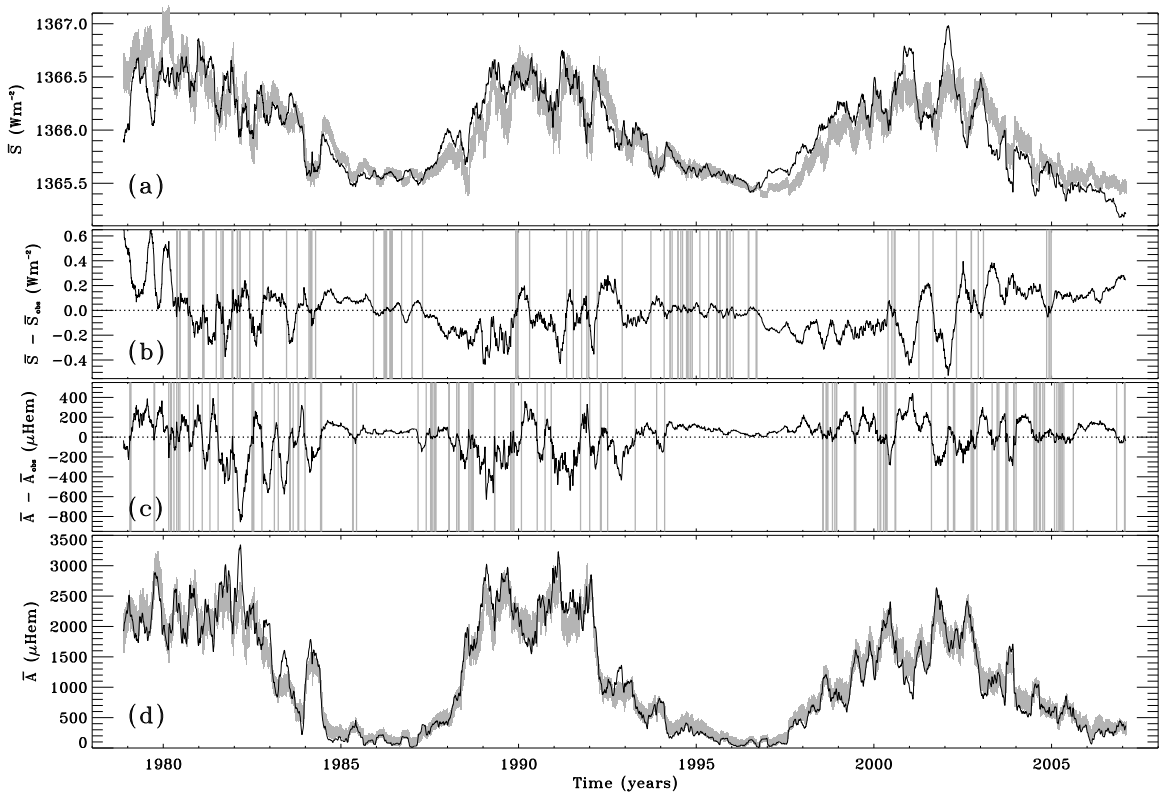

Fig. 13.— Same as Fig. 7, except the best fit model has parameter settings: $p_{frag} = 0.80$, $r^*_{erode} = 210$ km, $\eta = 1$, $\lambda = 0.012$, $\alpha_{fac} = 0.072$, $S_{Q,c} = 1365.52$ W m^{-2}, and $S_{Q,m} = -0.007$ W m^{-2} per year.

parameters (taking the best fifty solutions from 111 different PIKAIA runs). The results are summarised in Table 3.

As expected the optimal value for gradient $S_{Q,m}$ is negative, which confirms the presence of a downtrend in the PMOD composite. The range of values for $S_{Q,m}$ is relatively broad, indicating that it is not tightly constrained. The optimal value for the intercept $S_{Q,c}$ is slightly greater than S_Q as found in § 3, which is expected in the case of a downtrend. The optimal value for the average of S_Q (cf. eq. [10]) over the 1978–2007 interval is (1365.44 ± 0.04) W m^{-2}, which is in close agreement with the optimal value for S_Q in Table 2. The range of optimal values for $S_{Q,c}$ is larger than the range for S_Q, yet it is still relatively small. Over the ranges defined by their optimal values, it is evident that $S_{Q,m}$ and $S_{Q,c}$ are weakly correlated, with higher values of $S_{Q,c}$ corresponding to more negative values of $S_{Q,m}$ (steeper downtrends).

The optimal values of the other parameters and their respective distributions are generally very similar to those determined for the case without a linear trend in S_Q (see Table 2). There are some slight differences in the distributions for each parameter (most noticeably in those for p_{frag} and r^*_{erode}) but all optimal settings agree within their mutual mean plus or minus one standard deviation ranges. This indicates that the optimal values of the other parameters are not strongly sensitive to long-term variations in S_Q, which is important because the downtrend present in the PMOD composite could be due to either instrument degradation or the composition process (Fröhlich 2000, 2006). It is also important because it suggests that the parameter values employed in § 4 to perform the extrapolations for 1874–1978 would not be greatly modified by the presence of a weak long-term variation in the irradiance.

Figure 13 shows the results from the best-fit model for the case with a linear trend in S_Q and can be compared to Figure 7. As expected, the case with a downtrend better reproduces the observed irradiance during the period from around 2005 onward, when the observed irradiance (from the PMOD composite) decreases dramatically. It also

appears that the fit between the observed and synthetic irradiance is better for the model with a downtrend during the 1986–1989 interval (solar minimum and the ascending phase), although the fit is a little worse for the 1995–2000 interval and for the descending phase around 1985. For the period around 1978–1980 the model with a downtrend does worse; the model with constant-S_Q overestimated the irradiance during that period and the downtrend exacerbates this. Again, for the period 2000–2003 (solar maximum) some of the large peaks and troughs in the TSI are not well replicated by the model with a downtrend. Evidently, the linear trend presents some trade-offs; however, overall its inclusion improves the fit to the 1978–2007 data. The best fifty solutions have average fitness values that lie in the range: 1.617 to 1.805, compared to 1.503 to 1.665 for the case with constant-S_Q; χ^2_S is roughly 10% smaller for the model with a downtrend in the best-fit case, whereas χ^2_A is roughly 2% larger. Indeed, it is evident when comparing Figures 7(c) and (d) with Figures 13(c) and (d) that the daily sunspot area fit does not change significantly for the model with a downtrend despite the difference in the controlling parameters p_{frag} and r^*_{erode} between the two cases.

It is possible that the slight differences in the optimal parameter value distributions for the cases with and without a linear trend in S_Q could cause significant systematic modifications to the extrapolations over the 1874–1978 period. To test this we repeat the extrapolation process (as described in § 4) using the optimal parameter value distributions determined in this section, but with the linear trend turned off (as it would be unwise to extrapolate the linear trend far outside the calibration period). To turn off the linear trend we use $S_{Q,c}$ and $S_{Q,m}$ to compute the corresponding average value for S_Q (cf. eq. [10]) over the 1978–2007 interval. As found in § 3.1 the resultant S_Q values are not strongly correlated with any of the other model parameters. We treat the correlation between p_{frag} and r^*_{erode} in the same manner as in § 4 and assume that the remaining parameters are uncorrelated. Although not displayed, the resultant daily sunspot area extrapolation is almost indistinguishable from Figure 10. In fact, the average value of the unnormalised squared residual differences between the observations and the model predictions is 2% smaller over the 1874–2007 interval. For the irradiance (not shown) the difference is small but more noticeable than the sunspot area results. The peak irradiance (at solar maximum) and the standard deviation in irradiance realisations are smaller than in Figure 12; the average value of the unnormalised squared residual differences between the observations and the model predictions is approximately 10% smaller than in Figure 12 over the 1978–2007 interval. These results confirm that the extrapolations produced in § 4, that represent the rapidly varying component of the irradiance, would not be greatly modified by the presence of a weak, slowly varying component in the quiet-Sun irradiance. Of course, the net effect would introduce a gradual modulation of the irradiance time series presented in Figure 12.

6. DISCUSSION AND CONCLUSIONS

We have developed a model for active region decay and the total solar irradiance that takes the observed timing, position, and area of emerging active regions as input. The model active regions undergo a stochastic decay process that we simulate as a combination of fragmentation and boundary erosion, which produces a time-evolving distribution of magnetic flux elements over the solar surface. We assume that the small-scale magnetic elements, that are produced by active region decay, form the bright structures that counteract sunspot darkening: faculae and network. The model has several adjustable parameters that control the decay process and the irradiance contribution from various model components, such as the quiet-Sun and the small-scale flux elements. We have used a genetic algorithm to estimate the optimal values for these parameters, by fitting to the smoothed observed irradiance and daily sunspot area time series.

Each of the model parameters has a well-defined physical meaning. As such, this approach can be used to probe potentially observable but difficult to determine quantities. For example, we have computed optimal values for the boundary erosion rate of large-scale magnetic elements; the mean value corresponds to area decay rates that are at the low-end of the spectrum determined by Martinez Pillet et al. (1993). The corresponding range of area decay rates is large because the range of sunspots areas is large and the range of optimal values for the erosion rate is relatively broad. The optimal values computed for the fragmentation probability and the erosion rate are strongly correlated, with low erosion rates corresponding to high fragmentation probabilities (and vice versa). This is because in our model the erosion process works in combination with fragmentation. Consequently, the combined area decay rate for the large-scale magnetic elements is larger than the erosion rate alone.

Despite its simplicity, our model produces time series for the total solar irradiance and the daily sunspot area that agree very well with the observations over the time interval 1978–2007. The sunspot area fits tend to perform better than the irradiance fits, although the model consistently overestimates the sunspot area during solar minimum. It is evident that the model is neglecting a significant source of excess brightness, which manifests itself in three ways. First, the optimal values for the lifetime and intensity contrast of the bright, small-scale flux elements are both larger than expected. Second, the area covered by the collection of small-scale elements is significantly less than the observed Ca K plage area (Foukal 1996). Third, the synthetic irradiance time series consistently underestimates the observations during the ascending phase of a cycle, despite the daily sunspot area fitting the observations quite well during these times.

There are two important ingredients that our current model neglects which could account for the source of excess brightness. First, the dispersal and interaction of individual small-scale flux elements. The optimal values for the decay rate of the small-scale elements computed here correspond to e-folding decay times of about 92 days. Individual small-scale elements are unlikely to last this long. Thus, we conclude that interactions between the small-scale elements and, in particular, coalescence into larger, longer lifetime structures is especially important (Crouch et al. 2007). The inclusion of these effects would allow for the realistic simulation of magnetic structures composed of such elements, like faculae and network, and subsequently a more sensible multi-component irradiance model could be developed. Second,

the injection of small-scale magnetic flux elements independent of active region decay, possible sources of which are ephemeral regions and local dynamo action (e.g., Harvey & Zwaan 1993; Wang et al. 1995; Schrijver et al. 1997; Zhang et al. 1998b; Cattaneo 1999; Hagenaar 2001; De Pontieu 2002; Hagenaar et al. 2003; Abramenko et al. 2006; Zhang et al. 2006). Both of these ingredients could be incorporated into a model such as ours (e.g., Schrijver et al. 1997; Parnell 2001; Simon et al. 2001; Schrijver 2001; Krijger & Roudier 2003; Crouch et al. 2007), however, the calculation time would be considerably longer and may prohibit the genetic forward-modelling approach employed in this investigation.

Of course, the parameter estimates computed here are only as good as the model itself and the observations that are used for the fitting; for the irradiance we used the constructed composite data-set `d41_61_0702` provided by Physikalisch-Meteorologisches Observatorium Davos (Fröhlich 2000, 2006). With the genetic forward modelling approach we computed an optimal value for the constant quiet-Sun irradiance of 1365.42 W m^{-2} which agrees well with the average of the value at each observed solar minima quoted for the PMOD composite: 1365.502 ± 0.433 W m^{-2}. We also found evidence for a linear downtrend in the quiet-Sun irradiance over the 1978–2007 interval, corresponding to a decline of 0.0069 W m^{-2} per year. It is difficult to ascertain if this trend is solar in origin or if it is caused by a limitation in the observations, such as instrument degradation or the composite construction procedure; for a careful treatment of the issues associated with the latter see Fröhlich (2006). The inclusion of a downtrend in the irradiance improves the overall fit but it is not sufficient to completely explain the large discrepancy between the model predictions and the observed irradiance during the descending phase of cycle 23. The cause of this is uncertain, a deficiency in our model is the most obvious candidate but it should be noted that a similar discrepancy is also evident when comparing the PMOD composite TSI to other activity indices (e.g., Fröhlich 2006, and references therein). Nevertheless, we showed that the optimal values of the model parameters that are not directly related to the quiet-Sun irradiance are generally unaffected by the presence of a downtrend.

Assuming a constant quiet-Sun irradiance, we employed the distributions of optimal parameter values to construct an irradiance and sunspot area time series over the 1874–1978 time interval, for which observational data of emerging active regions is available. The comparison of the observed and synthetic daily sunspot area time series indicate that the quality of the model fits is noticeably inferior over 1874–1978. Because the source of the sunspot observations changes in 1976 from the Royal Greenwich Observatory to the US Air Force (USAF) and the US National Oceanic and Atmospheric Administration, it is difficult to distinguish the inhomogeneity in the observational record from a potential flaw in the model as the cause of the poorer fits. We demonstrated that the reconstruction of the rapidly varying component of the irradiance, as calculated with our fragmentation/erosion model, would not be strongly affected by the presence of a weak, long-term modulation (secular component), provided that such a modulation could be characterised as a slowly varying component of the quiet-Sun irradiance.

Because the primary input for our model is the timing, location, and area of emerging active regions, a major source of uncertainty originates from active regions that emerge on sections of the Sun that cannot be observed reliably – in the vicinity of the limb and the far side. If models of this type were to be used to forecast the irradiance on timescales of a solar rotation period, it would be worth trying to constrain the properties of active regions that emerge on the far side with techniques such as helioseismic far side imaging (Lindsey & Braun 2000; González-Hernández et al. 2007) and possibly Ly-α back-scattering from the interplanetary medium (Quémerais & Bertaux 2002).

We thank Ken Tapping for helpful discussion. This work was supported by the Natural Sciences and Engineering Research Council of Canada and the Canada Research Chairs program. The authors acknowledge receipt of the data-set `d41_61_0702` from PMOD/WRC, Davos, Switzerland and unpublished data from the VIRGO Experiment on the cooperative ESA/NASA Mission SoHO. The PMOD composite is available from `http://www.pmodwrc.ch/pmod.php?topic=tsi/composite/SolarConstant`. Daily active region area and position information were derived from data provided by the Royal Greenwich Observatory, the US Air Force and the National Oceanic and Atmospheric Administration, and are available at `http://solarscience.msfc.nasa.gov/greenwch.shtml` thanks to David Hathaway. Daily calcium plage area information from Mt. Wilson Observatory and white light faculae area information from the Royal Greenwich Observatory were obtained from the US National Oceanic and Atmospheric Administration's National Geophysical Data Center.

REFERENCES

Abramenko, V. I., Fisk, L. A., & Yurchyshyn, V. B. 2006, ApJ, 641, L65
Ackley, D. H. 1987, in Genetic Algorithms and Simulated Annealing, ed. L. Davis (London: Pitman), 170
Baumann, I., & Solanki, S. K. 2005, A&A, 443, 1061
Berger, T. E., & Title, A. M. 1996, ApJ, 463, 365
Bhatnagar, A., Jain, K., & Tripathy, S. C. 1999, ApJ, 521, 885
Bogdan, T. J., Gilman, P. A., Lerche, I., & Howard, R. 1988, ApJ, 327, 451
Brandt, P. N., Stix, M., & Weinhardt, H. 1994, Sol. Phys., 152, 119
Brown, G. M., & Evans, D. R. 1980, Sol. Phys., 66, 233
Brun, A.S., Miesch, M.S., & Toomre, J. 2004, ApJ, 614, 1073
Cattaneo, F. 1999, ApJ, 515, L39
Chae, J., Moon, Y.-J., & Pevtsov, A. A. 2004, ApJ, 602, L65
Chapman, G. A. 1980, ApJ, 242, L45
Chapman, G. A. 1987, ARA&A, 25, 633
Chapman, G. A., & Meyer, A. D. 1986, Sol. Phys., 103, 21
Chapman, G. A., & Hoffer, A. S. 2006, Sol. Phys., 237, 321
Chapman, G. A., Cookson, A. M., & Dobias, J. J. 1996, J. Geophys. Res., 101, 13541
Charbonneau, P. 1995, ApJS, 101, 309
Charbonneau, P. 2002, Release Notes for PIKAIA 1.2, NCAR Technical Note 451 + SRT, Boulder: National Center for Atmospheric Research
Charbonneau, P., & Knapp, B. 1995, A User's Guide to PIKAIA 1.0, NCAR Technical Note 418 + IA
Charbonneau, P., Tomczyk, S., Schou, J., & Thompson, M. J. 1998, ApJ, 496, 1015
Charbonneau, P., Christensen-Dalsgaard, J., Henning, R., Larsen, R. M., Schou, J., Thompson, M. J., & Tomczyk, S. 1999, ApJ, 527, 445

Crouch, A. D., Charbonneau, P., & Thibault, K. 2007, ApJ, (in press)
Davis, L. 1991, Handbook of Genetic Algorithms (New York: Van Nostrand Reinhold)
De Pontieu, B. 2002, ApJ, 569, 474
de Toma, G., White, O. R., Chapman, G. A., Walton, S. R., Preminger, D. G., & Cookson, A. M. 2004, ApJ, 609, 1140
Ermolli, I., Berrilli, F., & Florio, A. 2003, A&A, 412, 857
Ermolli, I., Criscuoli, S., Centrone, M., Giorgi, F., & Penza, V. 2007, A&A, 465, 305
Fligge, M., Solanki, S. K., & Unruh, Y. C. 2000, A&A, 353, 380
Foukal, P. 1993, Sol. Phys., 148, 219
Foukal, P. 1996, Geophys. Res. Lett., 23, 2169
Foukal, P. 1998, ApJ, 500, 958
Foukal, P. 2002, Geophys. Res. Lett., 29, 2089
Foukal, P. & Lean, J. 1986, ApJ, 302, 826
Foukal, P. & Lean, J. 1988, ApJ, 328, 347
Foukal, P., Harvey, K., & Hill, F. 1991, ApJ, 383, L89
Foukal, P., & Milano, L. 2001, Geophys. Res. Lett., 28, 883
Foukal, P., Bernasconi, P., Eaton, H., & Rust, D. 2004, ApJ, 611, L57
Foukal, P., Fröhlich, C., Spruit, H., & Wigley, T. M. L. 2006, Nature, 443, 161
Fröhlich, C. 2000, *Space Science Reviews*, 94, 15
Fröhlich, C. 2006, *Space Science Reviews*, 125, 53
Fröhlich, C., Pap, J. M., & Hudson, H. S. 1994, Sol. Phys., 152, 111
Gokhale, M. H., & Zwaan, C. 1972, Sol. Phys., 26, 52
Goldberg, D. E. 1989, Genetic Algorithms in Search, Optimization and Machine Learning (Reading: Addison-Wesley)
González-Hernández, I., Hill, F., & Lindsey, C. 2007, ApJ, accepted: ApJ preprint doi:10.1086/'521592'
Gray, D. F. & Livingston, W. C. 1997, ApJ, 474, 802
Hagenaar, H. J. 2001, ApJ, 555, 448
Hagenaar, H. J., Schrijver, C. J., & Title, A. M. 2003, ApJ, 584, 1107
Hagenaar, H. J., & Shine, R. A. 2005, ApJ, 635, 659
Harvey, K. L., & Zwaan, C. 1993, Sol. Phys., 148, 85
Harvey, K., & Harvey, J. 1973, Sol. Phys., 28, 61
Harvey, K. L., Jones, H. P., Schrijver, C. J., & Penn, M. J. 1999, Sol. Phys., 190, 35
Hathaway, D. H. & Choudhary, D. P. 2004, *Bulletin of the American Astronomical Society*, 36, 711
Holland, J. H. 1975, Adaptation in Natural and Artificial Systems (2d ed.; Cambridge: MIT)
Hoyt, D. V. & Schatten, K. H. 1993, J. Geophys. Res., 98, 18895
Hudson, H. S., Silva, S., Woodard, M., & Willson, R. C. 1982, Sol. Phys., 76, 211
Kálmán, B. 2001, A&A, 371, 731
Komm, R. W., Howard, R. F., & Harvey, J. W. 1993, Sol. Phys., 147, 207
Krause, F., & Ruediger, G. 1975, Sol. Phys., 42, 107
Krijger, J. M., & Roudier, T. 2003, A&A, 403, 715
Krivova, N. A., Balmaceda, L., & Solanki, S. K. 2007, A&A, 467, 335
Kuhn, J. R., Libbrecht, K. G., & Dicke, R. H. 1988, *Science*, 242, 908
Kuhn, J. R. & Stein, R. F. 1996, ApJ, 463, L117
Lawrence, J. K. 1988, Sol. Phys., 116, 17
Lean, J., Beer, J., & Bradley, R. 1995, Geophys. Res. Lett., 22, 3195
Lean, J. L., Cook, J., Marquette, W., & Johannesson, A. 1998, ApJ, 492, 390
Li, L. H., Basu, S., Sofia, S., Robinson, F. J., Demarque, P., & Guenther, D. B. 2003, ApJ, 591, 1267
Libbrecht, K. G., & Kuhn, J. R. 1985, ApJ, 299, 1047
Lindsey, C., & Braun, D. C. 2000, *Science*, 287, 1799
Liu, Y., Zhang, H., Ai, G., Wang, H., & Zirin, H. 1994, A&A, 283, 215
Martin, S. F., Livi, S. H. B., & Wang, J. 1985, Australian Journal of Physics, 38, 929
Martin, S. F., & Harvey, K. H. 1979, Sol. Phys., 64, 93
Martinez Pillet, V., Moreno-Insertis, F., & Vazquez, M. 1993, A&A, 274, 521
Martínez Pillet, V. 2002, Astronomische Nachrichten, 323, 342
Meyer, F., Schmidt, H. U., Wilson, P. R., & Weiss, N. O. 1974, MNRAS, 169, 35
Mitchell, M. 1996, An Introduction to Genetic Algorithms, Cambridge: The MIT Press
Moreno-Insertis, F. & Vazquez, M. 1988, A&A, 205, 289
Ortiz, A., Solanki, S. K., Domingo, V., Fligge, M., & Sanahuja, B. 2002, A&A, 388, 1036
Ortiz, A., Domingo, V., & Sanahuja, B. 2006, A&A, 452, 311
Parnell, C. E. 2001, Sol. Phys., 200, 23
Petrovay, K., & van Driel-Gesztelyi, L. 1997, Sol. Phys., 176, 249
Petrovay, K., & Moreno-Insertis, F. 1997, ApJ, 485, 398
Preminger, D. G. & Walton, S. R. 2005, Geophys. Res. Lett., 32, 14109
Preminger, D. G., & Walton, S. R. 2007, Sol. Phys., 240, 17
Quémerais, E., & Bertaux, J.-L. 2002, Geophys. Res. Lett., 29, 5
Rabin, D., Moore, R., & Hagyard, M. J. 1984, ApJ, 287, 404
Schrijver, C. J. 2001, ApJ, 547, 475
Schrijver, C. J., Title, A. M., van Ballegooijen, A. A., Hagenaar, H. J., & Shine, R. A. 1997, ApJ, 487, 424
Sheeley, N. R., Jr. 1969, Sol. Phys., 9, 347
Simon, G. W., & Leighton, R. B. 1964, ApJ, 140, 1120
Simon, G. W., Title, A. M., & Weiss, N. O. 2001, ApJ, 561, 427
Solanki, S. K. 2002, ESA SP-508: From Solar Min to Max: Half a Solar Cycle with SOHO, 11, 173
Solanki, S. K. & Fligge, M. 1998, Geophys. Res. Lett., 25, 341
Solanki, S. K., Schüssler, M., & Fligge, M. 2002, A&A, 383, 706
Spruit, H. 2000, *Space Science Reviews*, 94, 113
Steinegger, M., Vazquez, M., Bonet, J. A., & Brandt, P. N. 1996, ApJ, 461, 478
Tapping, K. F., Boteler, D., Charbonneau, P., Crouch, A., Manson, A., & Paquette, H. 2007, Sol. Phys.(accepted)
Thomas, J. H., Weiss, N. O., Tobias, S. M., & Brummell, N. H. 2002, Nature, 420, 390
Topka, K. P., Tarbell, T. D., & Title, A. M. 1997, ApJ, 484, 479
Vrabec, D. 1974, in IAU Symp. 56: Chromospheric Fine Structure, ed. R. G. Athay (Dordrecht: Reidel), 201
Wallenhorst, S. G., & Howard, R. 1982, Sol. Phys., 76, 203
Wallenhorst, S. G., & Topka, K. P. 1982, Sol. Phys., 81, 33
Walton, S. R., Preminger, D. G., & Chapman, G. A. 2003, ApJ, 590, 1088
Wang, J., Wang, H., Tang, F., Lee, J. W., & Zirin, H. 1995, Sol. Phys., 160, 277
Wang, Y.-M., Lean, J. L., & Sheeley, N. R., Jr. 2005, ApJ, 625, 522
Woodard, M. F. 1987, Sol. Phys., 114, 21
Worden, J. R., White, O. R., & Woods, T. N. 1998, ApJ, 496, 998
Zhang, J., Solanki, S. K., & Wang, J. 2003, A&A, 399, 755
Zhang, J., Lin, G., Wang, J., Wang, H., & Zirin, H. 1998a, Sol. Phys., 178, 245
Zhang, J., Lin, G., Wang, J., Wang, H., & Zirin, H. 1998b, A&A, 338, 322
Zirin, H. 1985, ApJ, 291, 858
Zhang, J., Ma, J., & Wang, H. 2006, ApJ, 649, 464