CS-BIGS 3(2): 108-122
http://www.bentley.edu/csbigs/documents/boland.pdf

Derivation of a Solar Diffuse Fraction Model in a Bayesian Framework

Philippe Lauret
Laboratoire de Physique du Bâtiment et des Systèmes, La Réunion, France

John Boland
University of South Australia, Adelaide, Australia

Barbara Ridley
University of South Australia, Adelaide, Australia

We propose a Bayesian statistical approach to deriving a simple logistic hourly diffuse fraction model. The model is calibrated with data that include northern as well as southern hemisphere sites. An independent dataset comprising seven worldwide locations is used to compare the model against several previous models such as Skartveit et al. (1998), Reindl et al. (1990), Erbs et al. (1982), and a version of the logistic model derived by Boland et al. (2008). On an overall basis, the new model performs better than the models of Erbs or Reindl, and exhibits similar performance to the Skartveit model but with a much simpler expression. In addition, the use of a Bayesian criterion for model selection confirms that the new proposed model achieves the best trade-off between goodness-of-fit and model complexity. Finally, it is shown that the use of Bayesian methods instead of classical statistical techniques lead to a less-biased model. Our presentation is accessible to readers with an intermediate level of statistics.

Keywords: *Diffuse fraction model, Bayesian statistics, prior information, model comparison, information criterion*

1. Introduction

Knowledge of solar radiation (both direct and diffuse) is essential for the proper modelling of solar energy oriented applications. More precisely, the evaluation of the performance of a solar collector such as a solar hot water heater or photovoltaic cell requires knowledge of the amount of solar radiation incident upon it. Solar radiation measurements are usually for global radiation on a horizontal surface. These global values comprise two components : the direct and the diffuse. The diffuse component takes into account the additional irradiance reflected from the clouds and the clear sky. However, it must pointed out that the diffuse part is not generally measured. Consequently, a method must be derived to estimate the diffuse radiation on a horizontal surface from the global radiation upon that surface. Numerous researchers have studied this problem and have been successful to varying degrees (Boland et al. 2001). Indeed, many models have been developed for determining the fraction of the global which is direct or diffuse (Wong and Chow 2001). One class of these models called decomposition models is based on correlations between the dimensionless hourly clearness index $k_t = I_G / I_0$ and the hourly diffuse dimensionless fraction $k_d = I_d / I_G$, where I_G, I_d, and I_0 are the global, diffuse and extraterrestrial radiation integrated over the hour in question (Orgill and Hollands 1977; Erbs et al. 1982 ; Spencer 1982; Reindl et al. 1990). For example, Reindl et al. (1990) initially proposed a first model that consists of 3 equations (each correlation for a specific bin of clearness index) that relate k_d to k_t. In a similar way, Erbs et al. (1982) developed polynomials up to order 4 to derive the diffuse fraction from the clearness index. In order to improve the models, additional geometrical and meteorological input variables have been proposed (Reindl et al. 1990; Skartveit et al. 1998). For instance, Reindl et al. (1990) have derived a second model that consists of piecewise linear correlations that compute the hourly diffuse fraction as a function of the hourly clearness index and solar elevation. A third model has been also designed when measurements of temperature and relative humidity are available. However, it may be worth noticing that measurements such as temperature or relative humidity are not always readily available in solar radiation series. As a consequence, some authors like Skartveit et al. (1998), Perez et al. (1992), Gonzales and Calbo (1999) proposed models that make use of information that is solely extracted from hourly global irradiance series. For instance, Skartveit et al. (1998) introduced, in addition to the clearness index and the solar altitude, a third variable called the hourly variability index[1] (σ_3), which is defined as the root mean squared deviation between the clear sky index of the hour in question and, respectively, the preceding hour and the succeeding hour. This additional predictor is intended to account for the effect of variable/inhomogeneous clouds. In their survey, the authors have compared extensively their model against models such as Erbs et al. (1982), Maxwell (1987) and Perez et al. (1992). While the Skartveit model exhibits very good performance, it is quite complicated as the model is phrased with a set of analytical expressions designed to assign the diffuse fraction into four bins of clearness index and to take into account the case of invariable hours $(\sigma_3 \approx 0)$, and variable hours $(\sigma_3 > 0)$.

Further, it must be outlined that most of the above cited models have been tuned or calibrated with data from Europe or North America and consequently may prove to be inadequate for southern hemisphere sites like Australia. However, some attempts have been made to build diffuse fraction models from southern sites data. As an illustration, Spencer (1982) developed a model suited for Australian locations in the latitude range 20-45° S. Boland et al. (2001) developed a validated model for Australian conditions using a logistic function of the form $k_d = \frac{1}{1+\exp(\alpha_0 + \alpha_1 k_t)}$ instead of piecewise linear or nonlinear correlations. More recently, Boland et al. (2008) used sound statistical techniques to justify the use of the logistic function. While the model has been constructed with data from various locations including northern as well as southern hemisphere sites, the authors concluded their work by the need for developing a new logistic model that takes into account (in addition to the clearness index) other predictors to enhance the fit.

Finally, it must also stressed that, to the best of our knowledge, the construction of all above models was made by using the classical least-squares technique. In this work, we propose to approach the problem from a different perspective by using Bayesian methods in order to derive a new logistic model. Indeed, Bayesian probability theory is currently experiencing an increase in popularity in the sciences as a means of probabilistic inference (Malakoff, 2005). Bayesian inference has already been applied successfully to complex models in the fields of physics, astronomy, medical statistics, financial modelling and genetics (Congdon, 2001). In the realm of solar energy-oriented applications, Lauret et al. (2006a) proposed a Bayesian approach to estimating convective heat transfer coefficients of a roof-mounted radiant barrier and also designed a neural network based-model in order to estimate the direct solar irradiance (Lauret et al., 2006b).

The goals of this paper can be stated as follows:

i) to propose a new logistic diffuse fraction model that is constructed from information readily available in global irradiance series,
ii) to investigate the use of a rather new (in the realm of solar radiation modelling) statistical method to design the model (Bayesian parameter estimation),
iii) to use a Bayesian criterion in order to quantitatively select a diffuse fraction model (i.e. Bayesian model selection)

The remainder of this paper is organised as follows. Section 2 discusses the datasets and the procedure used to calibrate the models and to assess the models' performance. Section 3 describes the proposed model. Section 4 introduces Bayesian techniques regarding the

[1] Perez *et al.* (1992) also introduced such a variability index in a similar manner.

two levels of inference: parameter estimation and model selection. Section 5.1 presents the results of the Bayesian parameter estimation while section 5.2 contains a detailed evaluation of the proposed diffuse fraction model. Section 5.3 deals with the Bayesian model selection. Finally, Section 6 gives some concluding remarks.

2. Datasets and Procedure

In an attempt to construct a generic model that could be applied to any location, we used hourly data from nine worldwide locations listed in Tables 1 and 2. The proposed dataset is supposed to cover a variety of climates and environments in Europe, Africa, Australia and Asia. Detailed information about the collected data for most of the locations can be found in Boland et al. (2008).

Note that our goal is not to build correlations that take into account seasonal dependency but instead yearly diffuse fraction correlations. Consequently, whenever possible, complete years of data were used to ensure that all seasons were treated equally well. It should also be noted that the calculation of the clearness index includes an element of deseasoning.

We divided the data into two subsets: the calibration or training set and the test set. Our model will be fitted using the calibration set (see Table 1) but for comparison purposes with the other existing models, a second dataset called test or validation set (see Table 2) will be used. Because this test set is not used in the construction of the model, it will provide an unbiased estimate of the performance of the models. To obtain the training dataset, a detailed study of the distribution of the diffuse fraction values for the different sites led us to consider 3 sites for the southern hemisphere and two stations for the northern hemisphere. This choice gave the best coverage of the diffuse fraction values. In other words, the training dataset (11074 data points) was constructed so as the dataset encompasses the whole spectrum of diffuse fraction values.

3. Proposed Model

As mentioned in the introduction, a logistic model of the form given by Eq. (1) was proposed by Boland et al. (2008) as to estimate the diffuse fraction:

$$k_d = \frac{1}{1+\exp\left(\alpha_0+\alpha_1 k_t\right)} \qquad (1)$$

By amalgamating data from seven locations (northern and southern hemisphere sites) and by using classical statistical techniques (Least-squares fit), Boland et al. (2008) obtained the following parameter values α_0=-5.00 and α_1 = 8.60.

Table 1. Training dataset: data used to calibrate or tune the models

	Site	Location	Year	#data
Southern sites	Adelaide (Australia)	34° 56' S, 138° 36' E, 48m a.s.l	2003	2,434
	Darwin (Australia)	12° 28' S, 130° 51' E, 30m a.s.l	2001, 2002	1,418
	Reunion (Reunion Is.)	20° 52' S, 55° 28' E, 25m a.s.l	2002	2,125
Northern sites	Camborne (UK)	50° 13' N, 5° 19' W, 88m a.s.l	2001	1,460
	Lisbon (Portugal)	38° 42' N, 9° 05' W, 56m a.s.l	1980	3,637

Table 2. Test dataset: data used to test the models (Validation dataset)

	Site	Location	Year	#data
Southern sites	Adelaide (Australia)	34° 56' S, 138° 36' E, 48m a.s.l	2004	2,307
	Darwin (Australia)	12° 28' S, 130° 51' E, 30m a.s.l	2003, 2005	1,179
	Maputo (Mozambique)	25° 58' S , 32° 35' E, 39 m a.s.l	1970	3,548
	Reunion (Reunion Is)	20° 52' S, 55° 28' E, 25m a.s.l	2000	2,045
Northern sites	Bracknell (UK)	51° 25 ' N, 0° 46' W , 77m a.s.l	1972	3,613
	Macau (China)	22° 11' N ,113° 33', 10m a.s.l	1985	3,513
	Uccle (Belgium)	50°48' N, 4°21' E, 104m a.s.l	1990	3,639

The use of such a logistic function brings a clear advantage as only one closed form equation is used to model the diffuse fraction. Indeed, in addition to being well suited for the S-shaped form of the diffuse fraction-clearness index relationship, there is no need here to split the model into different clearness index bins. Furthermore, and contrary to the models of Reindl et al. (1990), the values of the modelled diffuse fractions are always in the range (0-1). Note also that a formal statistical argument was presented to support the form of the model (Boland et al. 2008).

In an attempt to enhance the performance of the model, we propose in this paper an extension of this previous logistic model (denoted herein 'old' logistic model) by using the following variables: apparent solar time (AST), solar elevation α (in degrees), daily clearness index K_T and persistence index ϕ. This new logistic model, denoted herein the BRL model (for Boland-Ridley-Lauret model), is given by Eq. (2):

$$k_d = \frac{1}{1+\exp(\alpha_0 + \alpha_1 k_t + \beta_1 AST + \beta_2 \alpha + \beta_3 K_T + \beta_4 \phi)} \qquad (2)$$

The choice of these additional variables was made upon the following considerations: AST, unlike the solar elevation, is asymmetric about solar noon and may explain differences in the atmosphere between morning and afternoon (Boland et al., 2001). The daily clearness index can also be used as a predictor as the whole day may have a common characteristic. As mentioned above, some researchers (Skartveit et al. 1998; Gonzales and Calbo, 1999) demonstrated the improvement brought by adding a variability index into the list of the input parameters. Thus, we consider the same type of predictor variable as Skartveit et al. (1998) but in a different and simpler form. Indeed, we take as an extra predictor both a lag (at hour h-1) and a lead (at hour h+1) of the clearness index and average them, namely we define

$$\begin{gathered} \phi = (k_{t-1} + k_{t+1})/2 \text{ for sunrise} < \text{t} < \text{sunset}, \\ \phi = k_{t+1} \text{ for t=sunrise, } \phi = k_{t-1} \text{ for t=sunset.} \end{gathered} \qquad (3)$$

This input variable ϕ is deemed to account for the persistence of the sky conditions. Boland (2008) showed that the hourly deseasoned global radiation follows a first order autoregressive process, AR (1), commonly called a Markov Chain. Since k_t is a deseasonalized variable, it should possess the Markovian property and we have used this persistence index to capture this.

We wish to point out that in Ridley et al. (2009), we showed that if the extra variables were added separate to the logistic formulation, the result was diffuse fraction values >1 and <0. Including all variables within the logistic function constrains the diffuse fraction values.

Unlike Reindl et al. (1990), we will not consider variables such as ambient temperature and relative humidity. Indeed, there are many locations for which the humidity and ambient temperature would not be recorded, particularly when the solar radiation is estimated from satellite data. Our goal is to be able to predict the diffuse fraction with information solely extracted from the hourly global irradiance series.

Regarding the choice of the additional input variables, when building a solar radiation model, a term like the clearness index is necessarily included, but the inclusion of the other terms is open to doubt. This raises the question of selection among a range of possible variables or equally of possible models. For instance, Reindl et al. (1990) used stepwise regression techniques to reduce a set of 28 potential predictor variables to only four significant ones. For a set of v potential variables, there are 2^v potential models. If we restrict ourselves to the four preceding variables or their equivalent correlation coefficients $(\beta_i)_{i=1,2,3,4}$, this leads to $2^4 = 16$ possible choices. But, among the set of possible predictor variables, which ones should we choose? To put it in other words, which variables (and hence model) should we select?

Before proceeding further, it must be stressed that we have applied a Bayesian methodology for variable selection suggested by George and McCulloch (1993). This work (not presented in this paper) strongly led to the selection of the 'full' model i.e. the one that contains the four potential variables namely (AST), solar elevation α, daily clearness index K_T and persistence index ϕ.

The BRL model will be compared against four previous models : the 'old' logistic model (Eq. (1)), the model of Erbs et al. (1982) , the model of Skartveit et al. (1998) and to the second form of the model of Reindl et al. (1990) (i.e. the one that includes, in addition to the clearness index, the solar elevation).

Finally, it must also be noted that the present work aims to complement the work of Ridley et al. (2009) by estimating the parameters of the model in a Bayesian framework. The next section is devoted to a brief introduction of the Bayesian inference while Appendix A will give further details about the Bayesian computations.

4. Bayesian Inference

Bayesian probability theory is currently experiencing an increase in popularity in the sciences as a means of probabilistic inference (Malakoff, 2005). Cox (1946) showed that any method of scientific inference that satisfies simple rules of logical and consistent reasoning must be equivalent to the use of ordinary probability theory as originally developed by Bayes (1763) and Laplace (1812). Two of these simple rules of probability theory are the sum rule and product rule (where prob stands either for a probability or a probability density function (pdf)):

$$prob(x \mid I) + prob(\overline{x} \mid I) = 1 \qquad (4)$$

$$prob(x, y \mid I) + prob(x \mid y, I) \times prob(y \mid I) \qquad (5)$$

where $\overline{x}$ represents the proposition that x is false, the vertical bar "|" means "given" and the comma is read as the conjunction "and" . Two useful relationships are derived from these basic rules, namely, Bayes's theorem and the marginalization relationship:

$$prob(x \mid y, I) = \frac{prob(y \mid x, I) \times prob(x \mid I)}{prob(y \mid I)} \tag{6}$$

$$prob(x \mid I) = \int prob(x, y \mid I) dy \tag{7}$$

where the symbol *I* denotes the relevant background and assumptions. Notice that, for sake of clarity, the relevant background I will be omitted in the subsequent formulae related to the pdf's.

In the Bayesian context, a probability represents a degree-of-belief (or encodes a state of knowledge); that is, how likely something is to be true based on all the relevant information at hand. In other words, in the Bayesian context, a probability evaluates (quantitatively) the veracity of a hypothesis and this on the basis of all the available information. The name given to this approach comes from the key role played by Bayes's theorem. The latter is used to update the probabilities in the light of new data. Thus, the Bayesian approach is very close to the basics of scientific reasoning. Indeed, from a set of initial hypotheses, we carry out observations which enable us to deduce (or to infer) other conclusions or to update our initial beliefs.

However, this concept seems too vague and too subjective to the school of conventional statistics (i.e. a frequentist approach) which defined probability as the long-run relative frequency with which an event occurred, given infinitely many repeated experimental trials. Indeed, the concept of degree-of-belief is criticized by the school of frequentists as it leads to subjectivity (because my belief could be different from yours). Although the frequency definition appears to be more objective, it fails to tackle most real-life scientific problems. Further, from a Bayesian viewpoint, all probabilities are always conditional (i.e. based on all the relevant background) and as stated by (Jaynes 2003), objectivity requires only that two people having the same information should assign the same probability. A good review of the Bayesian approach is given by (Jaynes 1986) and (Loredo 1990).

Two levels of inference are involved in the task of data modelling. At the first level, we suppose that a particular model is true (i.e. the structure of the model is deemed correct) and we fit that model to the data i.e. we infer what values its free parameters should plausibly take, given the data. This step is repeated for each model. The second level of inference is the task of model comparison or model selection. This step consists in ranking the alternative models in the light of the data. Bayesian methods are able consistently and quantitatively to solve both the inference tasks (MacKay 2003). Let us write Bayes' rule for the two levels described above.

4.1. Bayesian Parameter Estimation

Assume the model has a vector of m parameters $\Theta = (\theta_1, \theta_2, ..., \theta_m)$. Bayesian inference deals with the estimation of the values of m model parameters about which there may be some prior beliefs. These prior beliefs can be expressed as a probability density function (pdf) called prior, $p(\Theta)$ and may be interpreted as the probability placed on all possible parameter values before collecting any new data. The dependence of the n observations (or measurements) $D = (d_1, d_2, ..., d_n)$ on the p parameters can also be expressed as a pdf: $p(D \mid \Theta)$, called the likelihood function. The latter is used to update the prior beliefs about Θ , to account for the new data D . This updating is done through Bayes's theorem:

$$p(\Theta \mid D) = \frac{p(D \mid \Theta) p(\Theta)}{p(D)} \tag{8}$$

where $p(\Theta \mid D)$ represents the posterior pdf and expresses the values of the parameters after observing the new data. In other words, the prior is modified by the likelihood function to yield the posterior. A major difference between Bayesian and frequentist (or classical) methods is that the Bayesian inference offers a framework (through the use of prior information) to continuously update our posterior beliefs. In other words, all previous work is not wasted as the preceding model parameters can be used as prior information for the derivation of the parameters estimates of the next (new) model.

Bayesian models or more precisely the computation of the posterior given by Eq. (8) cannot be done analytically. Under some simplifying assumptions (unimodal distribution, Gaussian approximation for the posterior distribution), the calculation of the posterior distributions is readily done by using Laplace's method (see MacKay 2003 for details). Unfortunately, most multi-dimensional models, like for instance the one under consideration here, do not fall into this specific class. Consequently, Bayesian analysis usually requires numerical methods for calculating the posteriors of interest. Any algorithm that generates samples from a distribution function could be used. In this survey, we propose to work with Monte Carlo methods which do not make any simplifying assumptions in the process of Bayesian data analysis.

Progress in Bayesian posterior computation is due undoubtedly to Markov Chain Monte Carlo (MCMC) methods. In this work, the Bayesian analysis is conducted by using the statistical software package WinBUGS (Winbugs 1989). WinBUGS (for Bayesian Inference Using Gibbs Sampling) is an easy-to-learn and easy-to-use software that implements the Gibbs sampler (Thomas et al. 1992; Geman and Geman 1984) for generating samples from a Markov Chain whose equilibrium distribution is the posterior distribution. The interested reader is referred to Appendix A for a detailed survey of Bayesian computations.

4.2. Bayesian model selection

At the second level of inference, the problem consists in inferring which model is most plausible given the data. The posterior probability of each model is as follows:

$$p\left(M_k \mid D\right) = \frac{p\left(D \mid M_k\right) p\left(M_k\right)}{p(D)} \qquad (9)$$

The data-dependent term $p\left(D \mid M_k\right)$ is called the evidence or the marginal likelihood for model M_k. The quantity $p\left(M_k\right)$ represents a prior belief for model M_k. If we have no particular reason to prefer one model over another, then we will assign equal priors to all models. Since the denominator does not depend on the model, one can see that the different models are ranked according to the evidence term $p\left(D \mid M_k\right)$. Notice that the evidence term is obtained from the likelihood $p\left(D \mid \Theta_k, M_k\right)$ by averaging over the priors of model parameter Θ_k for model M_k,

i.e. $p\left(D \mid M_k\right) = \int p\left(D \mid \Theta_k, M_k\right) p\left(\Theta_k \mid M_k\right) d\Theta_k$.

There are, however, practical problems associated with the computation of the above model probabilities. Indeed, a practical difficulty is that marginal likelihoods and the corresponding model probabilities are very sensitive to the choice of model parameters priors. To circumvent this problem, we propose the use of information criteria. Different model selection criteria exist. Among them, one can cite the Bayesian information criterion (BIC) (Schwarz 1978) and the Akaike information criterion (AIC) (Akaike 1974).

In the Bayesian framework, a new criterion called the Deviance Information Criterion (DIC) (Spiegelhalter et al. 2002) can be used. But contrary to the former criteria, the DIC is easy to calculate from the samples generated by a Markov Chain Monte Carlo simulation and is applicable to a wide range of statistical models. Moreover, it overcomes the necessity of identifying the numbers of parameters in the model which is required for calculation of BIC and AIC. Models with smaller DIC are better supported by the data. Therefore, DIC will provide a means to rank the different diffuse fraction models in the light of data.

Note that a DIC module that automatically computes this criterion is implemented in the latest version of WinBUGS. The interested reader is referred to Appendix B for a detailed explanation related to the DIC calculation.

5. Results

5.1. Results of the Bayesian Inference (Parameter Estimation)

In this section, we focus on the results of the first level of the Bayesian inference (i.e parameter estimation) but again, let us recall that we have also used a Bayesian methodology for variable selection (George and McCulloch 1993). Indeed, prior to the estimation of the models' parameters, all 16 potential models were evaluated and it was found that the inclusion of all predictors was necessary. The MCMC simulation runs came to the selection of the 'full' model (i.e. the one that includes all the predictors). The great number of data values led to strong evidence for the full model. If less data were available for computing the models' probabilities, the results could be quite different.

As mentioned above, Appendix A details the Bayesian computations related to the BRL model. It must be stressed, however, that we used a Student's t-distribution for the likelihood distribution (as opposed to a Gaussian distribution that is related to the classical Least-squares techniques). The Student's t-distribution has in general longer 'tails' than a Gaussian. This gives the important property of robustness which means that it is less sensitive than the Gaussian distribution to the presence of a few data points which are outliers.

Regarding the problem of parameter estimation, 30.000 iterations (for each chain) of the MCMC sampler under WinBUGS led to the statistics given in Table 3. Several checks were made in order to verify the convergence of the algorithm and the (good) exploration of the model's parameter space by the sampler.

The MC error is an estimate of how much of the variation in the posterior sample is due to the noise generated in the sampler. Its value should be very small relative to the standard deviation Sd (as seen in Table 3).

The 2.5% and 97.5% percentiles define the 95% credibility interval (not to be confused with the confidence interval used in the frequentist approach) for the parameter of interest i.e. for instance the probability that the value of the parameter α_0 lies between -5.403 and -5.244 is 95%, given the observed data and the prior belief.

Table 3. Results of the Bayesian inference

Parameter	Mean	SD	MC error	2.5%	median	97.5%
α_0	-5.323	0.040	7.78E-4	-5.403	-5.323	-5.244
α_1	7.279	0.074	0.002	7.135	7.280	7.423
β_1	-0.030	0.002	3.78E-5	-0.033	-0.030	-0.026
β_2	-0.005	3.29E-4	7.06E-6	-0.005	-0.005	-0.004
β_3	1.719	0.049	0.001	1.621	1.720	1.812
β_4	1.082	0.067	0.002	0.950	1.082	1.214

The appraisal of the BRL model and comparisons against previous models will be made using the following equation that summarizes the BRL model:

$$k_d = \frac{1}{1+\exp\left(-5.32+7.28k_t-0.03AST-0.0047\alpha+1.72K_T+1.08\phi\right)} \quad (10)$$

Diffuse fractions predicted by the above model are plotted in Figure 1 against clearness index for respectively 3 AST values, 3 solar elevations, 4 daily clearness index and 4 persistence values.

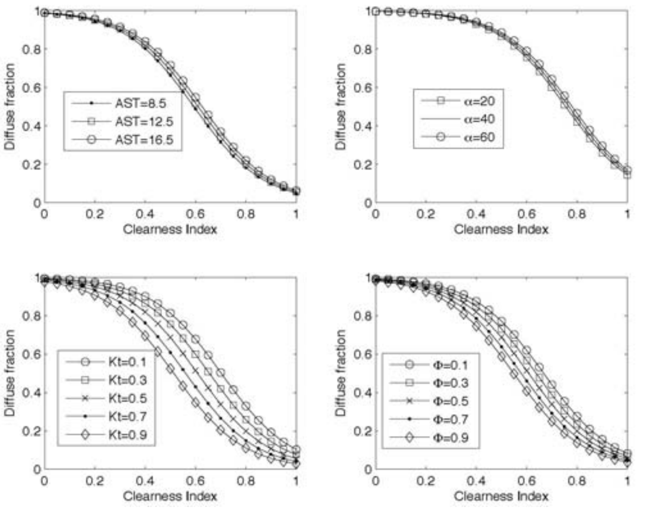

Figure 1. Diffuse fraction vs clearness index from the BRL diffuse fraction model for 3 AST values, 3 solar elevations, 4 daily clearness index and 4 persistence values.

Note that the sensitivity to the daily clearness index and the persistence is particularly significant in the k_t central interval (0.25 – 0.75). Thus, the model is quite able to deal with the large variety of sky conditions for a similar value of clearness index. Conversely, the sensitivity of the model output to the AST and to the solar elevation appears to be very small (as confirmed by the small values of the coefficients related to these 2 input variables). Hence, one may ask if it is worth including these two variables. In order to verify this point, and to confirm the results given by the work related to the Bayesian method for variable selection (briefly introduced in section 3), we fit a model by omitting these 2 variables. The results (not shown in this paper) were not as good as with the full model. A more important fact to note is that our model is not designed to take into account the usually observed increase in the diffuse fraction as $k_t > 0.8$. Further, such a trend is reinforced in this range of clearness index for decreasing or low solar elevations (as seen for instance in the model of Skartveit et al., 1998 or Reindl et al., 1990). However, the small amount of data in this region of k_t (2% of the present test dataset) and our will to build a single closed-form equation (instead of a set of complicated analytical expressions) led us to disregard this particular point. Further, one may notice that, in that region, the amount of incident energy is low, since it always occurs near sunrise or sunset. Note also that Erbs et al. (1982) also chose to disregard this fact arguing that these points are not understood well enough to justify fitting a curve to them.

5.2. Appraisal of the BRL model

The two classical statistical indicators Root Mean Squared Error (RMSE) and Mean Bias Error (MBE= Modelled -Observed) are used to evaluate the BRL logistic model against the preceding models. Tables 4 and 5 give an overall comparison of the different models for the different test sites. The LS column corresponds to the model tuned to the training dataset (see Table 1) by using the classical least-squares technique (see section 5.4 and Eq. (11)). The old logistic model is given in Eq. (1). Number of hours (N) and observed average diffuse fraction (Mean) are also given.

Note that, for the moment, we restrict the comparisons to the first five models. A special discussion related to the LS model will be given in section 5.5. Regarding the diffuse fraction, the RMSE for the BRL model are lowest or second lowest (just after the Skartveit model for the location of Adelaide and Maputo). The only exception to this ranking is related to the Uccle station where the Reindl model performs better than the BRL model. On

Table 4. MBE and RMSE of hourly diffuse fraction for all models.

Location	N	Mean		BRL	Skartveit	Reindl	Erbs	Old	LS
Adelaide	2,307	0.376	RMSE	0.092	0.091	0.103	0.104	0.099	0.096
			MBE	-0.011	0.013	0.001	0.003	-0.001	-0.001
Darwin	1,179	0.184	RMSE	0.084	0.096	0.098	0.091	0.094	0.090
			MBE	0.012	0.038	0.057	0.034	0.039	0.033
Maputo	3,548	0.403	RMSE	0.104	0.102	0.111	0.110	0.114	0.110
			MBE	0.004	0.024	0.034	0.008	0.015	0.013
Reunion	2,045	0.430	RMSE	0.131	0.138	0.143	0.149	0.149	0.126
			MBE	0.031	0.055	0.054	0.033	0.034	0.035
Bracknell	3,613	0.793	RMSE	0.100	0.111	0.114	0.114	0.111	0.101
			MBE	-0.023	-0.035	-0.040	-0.035	-0.040	-0.033
Macau	3,513	0.669	RMSE	0.100	0.102	0.103	0.110	0.111	0.101
			MBE	0.016	0.022	0.024	0.003	0.001	0.013
Uccle	3,639	0.699	RMSE	0.106	0.093	0.100	0.113	0.114	0.102
			MBE	0.032	0.015	0.016	0.022	0.019	0.023
All sites	19,844	0.562	RMSE	0.104	0.105	0.111	0.114	0.114	0.104
			MBE	0.008	0.014	0.015	0.005	0.005	0.008

Table 5. MBE and RMSE of hourly diffuse irradiance for all models.

Location	N	MEAN		BRL	Skartveit	Reindl	Erbs	Old	LS
		($W.m^{-2}$)	($W.m^{-2}$)						
Adelaide	2,307	166	RMSE	56	53	62	68	62	59
			MBE	-6	7	3	3	-1	3
Darwin	1,179	115	RMSE	55	62	65	59	59	59
			MBE	8	22	38	22	23	23
Maputo	3,548	170	RMSE	61	57	69	65	67	64
			MBE	2	12	22	1	4	10
Reunion	2,045	183	RMSE	63	67	72	70	71	64
			MBE	9	23	25	4	6	16
Bracknell	3,613	172	RMSE	40	41	44	50	47	39
			MBE	-11	-14	-17	-18	-18	-12
Macau	3,513	195	RMSE	47	51	53	54	54	48
			MBE	7	10	11	-3	-2	9
Uccle	3,639	150	RMSE	34	33	35	39	38	33
			MBE	7	3	4	1	2	7
All sites	19,844	169	RMSE	50	51	56	57	56	52
			MBE	1	6	9	-2	0	6

average, and for all the sites (see last line of Table 4), the BRL model reduces the RMSE of the diffuse fraction by 1% , 6%, 9% and 9% when compared respectively to the Skartveit model, the Reindl model, the Erbs model and the old logistic model (see Eq. (1)).

In a similar way, the performance of the BRL model regarding the MBE is quite good (apart again the Uccle site). In addition, the BRL model produces better MBE at southern sites like Darwin, Maputo and Reunion than the other models. Overall, the results seem to be consistent between all the models. Indeed, all the models tend to overestimate the diffuse fraction (with only one exception for the Adelaide station) and a closer look at Table 4 reveals that all the models underestimate the diffuse fraction for the Bracknell data.

In terms of RMSE of hourly diffuse irradiance (see Table 5), on an overall basis, we can state that the BRL model performs better (or equally well as the Skartveit model) than the others models. Indeed, for all the sites, the BRL model diminished the RMSE by 1% when compared to the Skartveit model. Conversely, its overall performance against the Reindl model or the Erbs model is much better (11% and 12% of RMSE improvement respectively). In addition, our goal to enhance the fit of the old logistic model (see Eq. (1)) is also reached as the BRL model increase by a factor of 11% the performance of the previous model.

The better performance of the BRL model and the Skartveit model against the Erbs or the Reindl model confirm (see Skartveit et al., 1998; Gonzales and Calbo, 1999) the improvement brought by adding further predictors such as the variability index σ_3 or the persistence variable ϕ.

On an overall basis, the performance of the BRL regarding the MBE of the diffuse irradiance is better than that of the other models. Again, it is also worth noticing that this better performance is more pronounced for the southern sites. Further, irrespective of the location, and unlike the other models, the variation of the MBE for the BRL model remains in the range $[-10;+10]$ W.m^{-2} (see also Figure 2).

Finally, in order to better appreciate the overall performance of the BRL model against the other models, group mean values of measured diffuse irradiance data (sorted by clearness index) and model response were calculated (see Figure 3).

As seen in Figure 3, the better overall performance of the BRL model against the models of Reindl et al. (1990) or Erbs et al. (1982) are clearly exhibited. Additionally, the performance of the BRL against the model of Skartveit et al. (1998) is slightly better. However, one may keep in mind that the expression of the Skartveit model is much more complicated than the BRL model.

Nonetheless, in order to get a more detailed comparison between these 2 best performers, we report in Table 6 the overall performance (RMSEs and MBEs of hourly diffuse irradiance) of the 2 models for the 7 sites of the test dataset (19844 data points) sorted by clearness index and solar elevation. Table 7 also gives the number of validation data together with the mean diffuse irradiance in each bin.

Hence, one may notice that the small edge of the BRL vs the Skartveit model comes from a better treatment of the intermediate ($0.25 < k_t \leq 0.75$) clearness index range (which corresponds to 67% of the test data). More precisely, in this central bin, the BRL reduced the RMSE by 4% when compared to the Skartveit model. At the opposite, and to confirm our preceding discussion about the high range of clearness index, the BRL yields a worse performance in the high clearness index range ($k_t > 0.75$). The disagreement is reinforced at decreasing solar elevations. But, again, as stated in section 5.1, we chose to not treat this special case of high range of clearness index, especially since it is dominated by the incidence of low energy values. Moreover, in our opinion, the small amount of data in this bin (2% of all the data) is not commensurate with the extra effort required to derive a specific correlation.

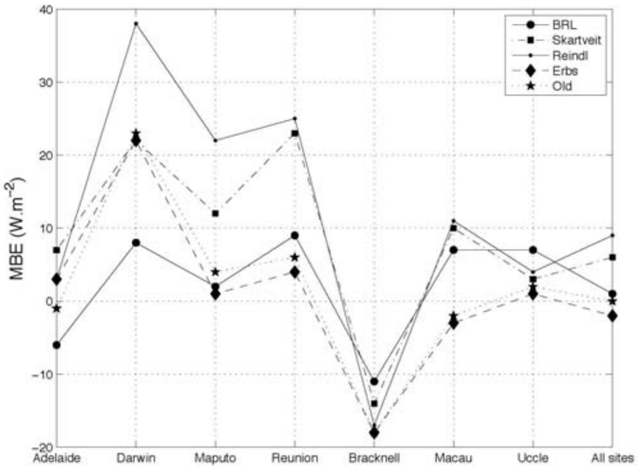

Figure 2. Variations of models MBE of diffuse solar irradiance as a function of location.

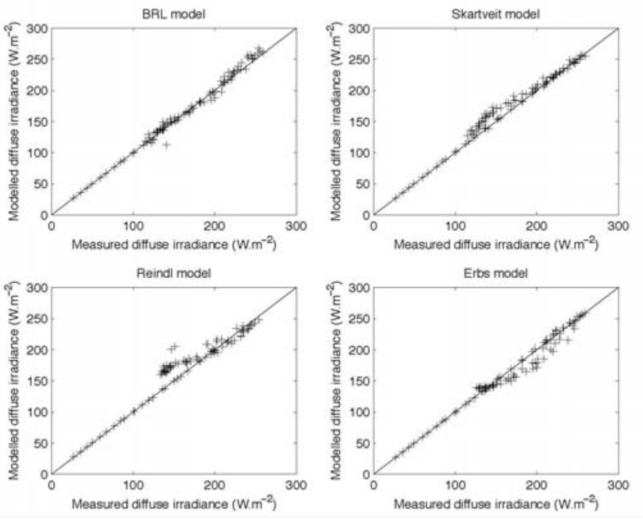

Figure 3. Modelled vs measured group mean hourly diffuse irradiance for the all the test sites. Modelled data are obtained from group mean values of clearness index.

As mentioned above, the BRL model performs slightly better against the Skartveit model. However, the expression of the Skartveit model is much more complicated than the BRL model. Therefore, in order to be assessing quantitatively this previous statement, we propose in the next section to calculate the DIC for each model.

5.3. DIC as a model selection criterion for diffuse fraction models

Table 8 lists the DIC for each model. As mentioned in Appendix B, DIC is allowed to be negative and only differences in DIC are important (its absolute size is irrelevant). The idea is that models with smaller DIC should be preferred to models with larger DIC.

Table 6. BRL and Skarveit models' overall RMSEs $\left(\text{W.m}^{-2}\right)$ and MBEs $\left(\text{W.m}^{-2}\right)$ for the 7 sites of the test dataset (19844 points) as a function of clearness index and solar elevation.

BRL Model	$\alpha \le 20°$		$20° < \alpha \le 40°$		$40° < \alpha \le 60°$		$\alpha > 60°$		$\forall\ \alpha$	
	RMSE	MBE	RMSE	MBE	RMSE	MBE	RMSE	MBE	RMSE	MBE
$k_t \le 0.25$	3	0	4	0	6	0	11	0	5	-1
$0.25 < k_t \le 0.75$	28	4	43	5	62	-1	88	-2	54	3
$k_t > 0.75$	95	0	56	-1	53	0	74	1	62	-2
$\forall\ k_t$	23	6	39	3	56	-1	77	-2	50	1
SKARTVEIT Model	$\alpha \le 20°$		$20° < \alpha \le 40°$		$20° < \alpha \le 40°$		$\alpha > 60°$		$\forall\ \alpha$	
	RMSE	MBE	RMSE	MBE	RMSE	MBE	RMSE	MBE	RMSE	MBE
$k_t \le 0.25$	3	0	4	0	5	0	11	0	5	1
$0.25 < k_t \le 0.75$	27	1	46	10	65	6	87	0	56	7
$k_t > 0.75$	70	0	49	1	51	4	67	1	57	-11
$\forall\ k_t$	22	2	41	9	58	8	74	3	51	6

Table 7. Number of validation data points (# Occ.) and mean diffuse irradiance in each clearness index-solar elevation bin. Note that the number of occurrences is 439 for the range $k_t > 0.8$ (2% of all the data).

#Occurrences / Mean $\left(\text{W.m}^{-2}\right)$	$\alpha \le 20°$		$20° < \alpha \le 40°$		$20° < \alpha \le 40°$		$\alpha > 60°$		$\forall\ \alpha$	
	#Occ.	Mean	#Occ.	Mean	#Occ.	Mean	#Occ.	Mean	#Occ.	Mean
$k_t \le 0.25$	1251	11	1437	24	845	23	359	12	3892	98
$0.25 < k_t \le 0.75$	2098	35	5436	160	4229	196	1396	8	13159	198
$k_t > 0.75$	13	1	381	7	1341	29	1058	29	2793	128
$\forall\ k_t$	3362	75	7254	144	6415	210	2813	250	19844	169

As seen in Table 8, according to the DIC, the BRL model achieves the best trade-off between goodness-of-fit and model complexity. The sizeable difference in DIC between the BRL model and the Skartveit model gives strong evidence for the selection of the former. Moreover, the ranking obtained from the DIC results seems quite consistent with the previous results (see section 5.2). The problem of finding the optimal complexity for a model provides an example of Occam's razor (MacKay 2003). This is the principle that one should prefer simpler models to more complex models, and that this preference should be traded off against the extent to which the model fits the data. To put it in other words, an increase in the number of predictors will always improve the fit. Information criteria reflect the concept that an extra predictor should give a commensurate improvement in goodness-of-fit.

As seen in Table 8, according to the DIC, the BRL model achieves the best trade-off between goodness-of-fit and model complexity. The sizeable difference in DIC between the BRL model and the Skartveit model gives strong evidence for the selection of the former. Moreover, the ranking obtained from the DIC results seems quite consistent with the previous results (see section 5.2). The problem of finding the optimal complexity for a model provides an example of Occam's razor (MacKay, 2003). This is the principle that one should prefer simpler models to more complex models, and that this preference should be traded off against the extent to which the model fits the data. To put it in other words, an increase in the number of predictors will always improve the fit.

Table 8. DIC values as estimated by the WinBUGS program for each diffuse fraction model. The column Relative difference reports the difference in DIC between the BRL model and the other models.

Diffuse fraction model	DIC	Relative difference	Rank
BRL	-5524	N/A	1
Skartveit	-4932	592	2
Reindl	-4589	935	3
Erbs	-4441	1083	4
Old logistic	-4397	1127	5

Information criteria reflect the concept that an extra predictor should give a commensurate improvement in goodness-of-fit.

5.4. Comparison with the Perez model

The Perez model (Perez et. al. 1992) is well regarded in the solar radiation modelling fraternity. We have not included it in the comparisons above since it estimates direct normal rather than diffuse radiation from global and uses other variables. It should be noted that requirements for knowledge values of the three components of radiation differ between technologies. Direct normal solar irradiance (DNI) is needed for concentrated solar power systems (CSP) and global horizontal solar irradiance (GHI) for flat plate collectors such as photovoltaic cells and solar hot water heaters. In the latter case, diffuse horizontal solar irradiance is necessary to estimate the global on a tilted surface when only the GHI is known or estimated from satellite derived data. The procedure proceeds in this manner:

- the diffuse on the horizontal is estimated from GHI ;
- the direct on the horizontal is derived by subtracting diffuse from GHI ;
- trigonometric calculations are used to calculate direct on the tilted surface ;
- the algorithms derived in Perez et. al. (1990) or similar are used to estimate the diffuse on the tilted surface ;
- the global on the tilted surface is calculated by adding direct and diffuse.

Obviously the DNI as estimated using Perez et. al. (1992) could be used in this exercise, with use of trigonometry to get the direct on the horizontal to infer the diffuse on the horizontal, and also to get the direct on the tilted surface. One can, however, see the impetus for developing models for diffuse on the horizontal from GHI as well as models for DNI.

We wish to present a comparison of the BRL model to the Perez model. We do this in two ways statistically as well as presenting a visual explanation. We won't delineate the comparison in terms of diffuse radiation since that is reported in Ridley et. al. (2009). We will only summarise results from that work. In it, the Perez model was used to predict DNI and the direct on the horizontal was derived from that and then the diffuse on the horizontal derived, and compared to that estimated using the BRL model. The BRL model consistently outperformed Perez using Median Absolute Percentage Error (MeAPE) and Bayesian Information Criterion (BIC) for several locations in both Southern and Northern Hemispheres. Assessing the Median Bias Error (MeBE), we found that for only two Northern Hemisphere locations did Perez outperform BRL and then only marginally.

We next reverse the direction of estimation for comparison. We take the diffuse on the horizontal estimated from BRL and calculate the direct on the horizontal through subtraction from the GHI. We then use the equation $DNI = direct.on.horizontal / \sin\alpha$ where as stated previously, α is the solar elevation. This may seem redundant, but when comparing model results, it is best to ensure the procedures work both ways. When we compare the predicted versus actual for the Adelaide data set, we find an MeAPE of 8.77 for the BRL model and 21.06 for the Perez model. Overall, the BRL model performs better than the Perez model for Southern hemisphere locations and at least as well for Northern Hemisphere locations.

A visual comparison of the two models using data from Adelaide, a Southern Hemisphere location, is instructive, and demonstrates the motivation for the series of papers on this topic. Figure 4 gives the diffuse fraction plotted against the clearness index overlain with the estimated diffuse fraction using the Perez model. It is easily seen that the bottom right hand corner of the data is not well predicted. If we compare that to the same graph but with the predictions performed with the BRL model, we see that this discrepancy is taken care of (Figure 5). The lack of suitability of models developed with solely Northern Hemisphere data that were inadequate for Southern Hemisphere locations was what led first Spencer (1982), and subsequently Boland et. al. (2001), Boland et. al. (2008) and Ridley et. al. (2009) to revisit the problem.

5.5. Bayesian inference vs classical least-squares technique

Let us recall that the other goal of this work was to evaluate the advantage brought by the Bayesian inference over the classical Least-Squares (LS) technique.

To assess this particular point, we fit the BRL model (with the training dataset given in Table 1) by using the classical Least-Squares (LS) technique. The following model (denoted LS model) was obtained:

$$k_d = \frac{1}{1+\exp\left(-4.60+6.54k_t - 0.04AST - 0.0054\alpha + 1.71K_T + 0.85\phi\right)} \qquad (11)$$

Note that the performance of this LS model is given in Tables 4 and 5 under the LS column.

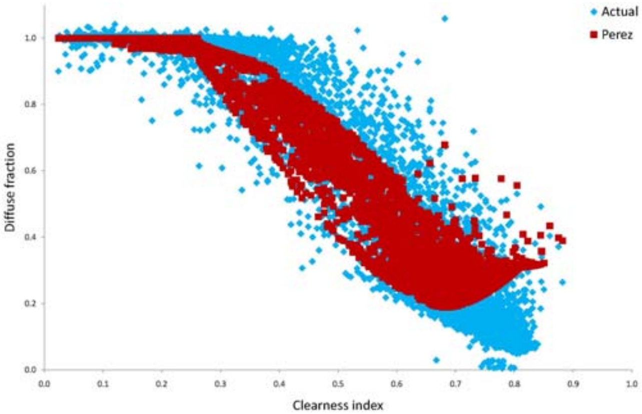

Figure 4. Diffuse fraction versus clearness index for Adelaide with Perez model predictions.

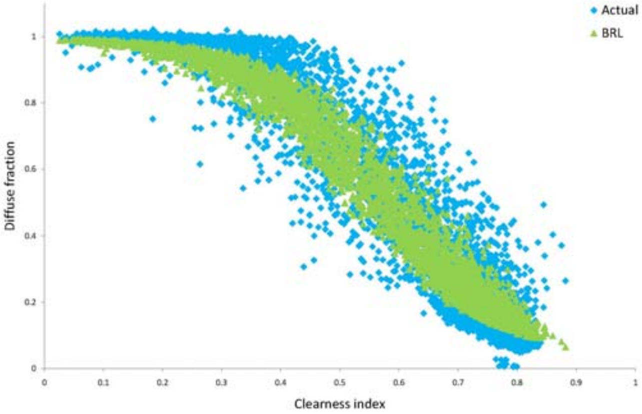

Figure 5. Diffuse fraction versus clearness index for Adelaide with BRL model predictions.

Figure 6 gives a comparison of the two approaches regarding the diffuse irradiance. As seen, the clear advantage brought by the Bayesian inference is the better MBE (particularly for the southern hemisphere sites).

This result is not surprising as we used a student-t distribution for the likelihood distribution (as opposed to a Gaussian distribution that is related to the classical Least-squares techniques). Indeed, the student-t density is a heavier tailed alternative to the Normal bell-shaped like the Gaussian but with a higher chance of extreme values in the tails. In this sense, it is a robust alternative to the Gaussian in the event of suspected outliers in the data.

Moreover, we can state that the use of southern hemisphere data to tune the model parameters (BRL and LS) naturally improve the performance of these models on the southern test stations, while performing at least as well as the best other available models for the northern hemisphere test stations. However the Bayesian method goes a step further by improving the MBE.

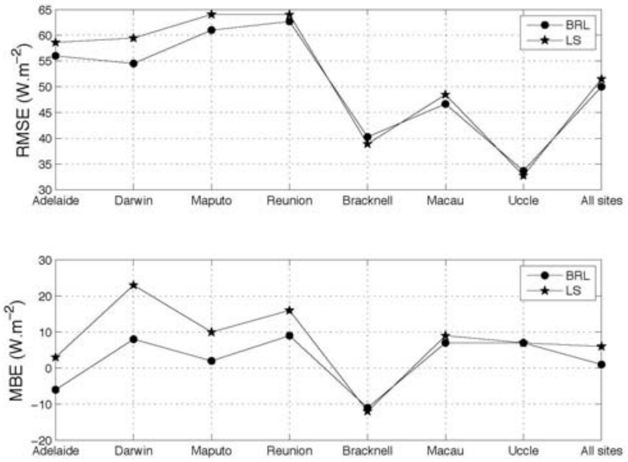

Figure 6. BRL model against LS model. Variations of models RMSE and MBE of diffuse solar irradiance as a function of location.

6. Conclusion

In this work, we have proposed a Bayesian logistic model (called the BRL model) that consists of a single equation for the entire data as opposed to a set of correlations used in models like Skartveit et al (1998), Reindl et al. (1990) and Erbs et al. (1982). A test against independent data from seven worldwide locations showed that the BRL model improved by 10% the root mean square error (RMSE) when compared to the models of Reindl and Erbs. The improvement in the RMSE is less pronounced when compared to the (much more complicated) Skartveit model. Regarding this particular point of model complexity, it has been demonstrated that using an information criterion like the DIC led clearly to the selection of the BRL model.

In addition, on an overall basis, the BRL model exhibited a better Mean Bias Error irrespective of the station than the other models. It has been shown that this characteristic comes from the use of the Bayesian inference instead of a classical approach.
The results obtained from the Bayesian inference may exhibit some sensitivity to the division of data used for training and testing the models. It would be desirable to use all the available data for estimating the parameters of the BRL model but we feel this would be unfair to the other models. If more data are available for tuning the model (or if there is need for recalibration accounting for local climatic differences) then one can refine the estimation of the parameters in the Bayesian framework. Indeed, all this previous work is not wasted as the preceding model's parameters (see Table 3) can be used as prior information for the derivation of the new model.

In our opinion, the solar radiation modelling community can greatly benefit from the Bayesian framework. Thus, it is worth using rather theoretically complicated modelling tools (though uncomplicated in practice) that could improve the quality of solar radiation models.

APPENDIX A: Bayesian Computations

A.1. The Probabilistic Model

This section gives some details about the implementation of the probabilistic model needed to fit the proposed model to the data. We successively describe each component of Eq. (8) i.e. the likelihood and the prior for the parameters.

The likelihood function

Let us consider the following form:

$$d_i = y_i + \varepsilon_i \qquad \text{(A1)}$$

where d_i represents the ith experimental measurements of the diffuse fraction , $\Theta = [\alpha_0, \alpha_1, \beta_1, \beta_2, \beta_3, \beta_4]$ the vector of model parameters, $y_i = y(\mathbf{x}_i, \Theta)$ the ideal (noiseless) model response or regression model, $\mathbf{x}_i = [k_t \quad AST \quad \alpha \quad K_t \quad \phi]_i$ the corresponding vector of predictor variables and where the noise ε_i is an expression of the various uncertainties (i.e. measurement noise plus modelling error). If the noise in the data is assumed to be Gaussian with variance σ^2 , then

$$p(\varepsilon_i) = \frac{1}{2\pi\sigma^2}\exp\left(-\frac{1}{2\sigma^2}\varepsilon_i^2\right) \qquad \text{(A2)}$$

which we can rewrite, by using Eq. (A1) as :

$$p(d_i|\Theta) = \frac{1}{2\pi\sigma^2}\exp\left(-\frac{1}{2\sigma^2}\left(d_i - y_i\right)^2\right) \qquad \text{(A3)}$$

Notice that this pdf can be written in a more compact manner as:

$$d_i : \mathrm{N}(y_i, \sigma^2) \qquad \text{(A4)}$$

which means that each measurement d_i is distributed as a Gaussian or Normal distribution with mean y_i and variance σ^2 .

Further, if we assume the noise as independent and additive then the likelihood function takes the following form:

$$p(D|\Theta) = \prod_{i=1}^{n} p(d_i|\Theta) = \frac{1}{(2\pi\sigma^2)^{n/2}}\exp\left(-\frac{1}{2\sigma^2}S(\theta)\right) \qquad \text{(A5)}$$

where n represents the number of experimental measurements and the term $S(\Theta) = \sum_{i=1}^{n}(d_i - y_i)^2$ is the sum-of -squares of the residuals.

Hence, one may notice that the classical least-squares (LS) technique is derived from the principle of maximum likelihood (ML) with a choice of a Gaussian distribution for the likelihood. Indeed, the ML principle consists in maximising the likelihood which is turn equivalent to minimising the negative log likelihood. The latter is the usual sum-of-squares.

However, in this survey, even if a quality-check has been done on the data, we chose a student's-t distribution for the likelihood function i.e. $d_i{:}\mathrm{St}(y_i, \lambda, \nu)$ where λ is the precision and ν the number of degrees of freedom that determines the extent of over-dispersion. Student's t-distribution has in general longer 'tails' than a Gaussian. This gives the important property of robustness which means that it is less sensitive than the Gaussian distribution to the presence of a few data points which are outliers (Bishop, 2006).

The prior distributions

For the BRL model, we make use of the information obtained from the previous step (See Eq. (1) and Boland et al. 2008). Indeed, as mentioned above, Bayesian inference offers a framework (through the use of prior information) to continuously update our posterior beliefs. All previous work is not wasted as the preceding model parameters can be used as prior information for the derivation of the parameters estimates of the next (new) model. As an illustration, for parameters (α_0, α_1) , we chose two normal distributions with a relatively small variance and with mean given by the values of the parameters of the first logistic model (Eq. (1)). Regarding the last four parameters $(\beta_1, \beta_2, \beta_3, \beta_4)$, we chose four Gaussian distributions with mean 0 and with large variance as we have no idea of the values they can plausibly take.

This setting is detailed in Table A1.

Table A1. Probabilistic model

Likelihood $p(D\mid\Theta)$ Student's t-distribution	Prior on parameters $p(\Theta)$ Gaussian distributions
$d_i \sim St(y_i,\lambda,\nu=2)$	$\alpha_0 \sim \mathrm{N}(-5{,}100)$
$y_i=\frac{1}{1+\exp\left(\alpha_0+\alpha_1 k_t+\beta_1 AST+\beta_2\alpha+\beta_3 K_T+\beta_4\phi\right)}$	$\alpha_1 \sim \mathrm{N}(8.60{,}100)$
$\Theta=\left[\alpha_0,\alpha_1,\beta_1,\beta_2,\beta_3,\beta_4\right]$	$\beta_1 \sim \mathrm{N}(0{,}1000000)$
$\mathbf{x}_i=\left[k_t \quad AST \quad \alpha \quad K_t \quad \phi\right]_i$	$\beta_2 \sim \mathrm{N}(0{,}1000000)$
	$\beta_3 \sim \mathrm{N}(0{,}1000000)$
	$\beta_4 \sim \mathrm{N}(0{,}1000000)$

A.3 Computation of the posterior distribution of the model parameters with MCMC numerical methods

Given the likelihood and the posterior, the MCMC implemented in WinBUGS samples the posterior given by Eq. (8). More precisely, MCMC generates samples from the posterior parameter space Θ by defining a chain $C=\{\Theta_1,\Theta_2,\Theta_3,\cdots,\Theta_i\cdots\}$. At each iteration i, candidate values Θ^* are generated randomly for each of the parameters. An acceptance probability is calculated, and the chain either moves to Θ^* ($\Theta_{i+1}=\Theta^*$) or stays at its current value ($\Theta_{i+1}=\Theta_i$). In the long run, the distribution of points in the chain C approximates the posterior distribution. When the chain converges, the posterior distribution of each parameter of interest can be drawn or summarized by the following statistics: mean, standard deviation, 95% credibility interval, etc.

APPENDIX B: Deviance Information Criterion (DIC)

The deviance information criterion (DIC) is particularly useful in Bayesian model selection problems where the posterior distributions of the models have been obtained by Markov chain Monte Carlo (MCMC) simulation (Spiegelhalter et al., 2002). Indeed, the deviance defined as $D(\Theta)=-2\log\left[p(D|\Theta)\right]$where $p(D|\Theta)$ is the likelihood function is automatically calculated under WinBUGS (Winbugs, 1989).

The DIC is the sum of two terms: a first term, the posterior mean of the deviance, that measures goodness of fit (adequacy) of the model, and a second term of penalty, P_D which estimates the complexity of the model:

$$DIC=\bar{D}(\Theta)+P_D \qquad \text{(B1)}$$

The first component is defined as the posterior mean of the deviance:

$$\bar{D}(\Theta)=E\left[D(\Theta)\right] \qquad \text{(B2)}$$

The expectation $\bar{D}(\Theta)$ is a measure of how well the model fits the data; the larger this is, the worse the fit.

The second term measures the complexity of the model by the effective number of parameters, defined as the difference between the posterior mean of the deviance and the deviance evaluated at the posterior mean $\bar{\theta}$ of the parameters:

$$P_D=\bar{D}(\theta)-D(\bar{\theta}) \qquad \text{(B3)}$$

As seen, P_D represents the decrease in the deviance expected from estimating Θ and can be used as a measure of the effective number of parameters (comparatively, one has to specify the number of model parameters when using information criteria such as BIC or AIC). The larger P_D is, the easier it is for the model to fit the data.

The idea is that models with smaller DIC should be preferred to models with larger DIC. Indeed, models are penalized both by the value of $\bar{D}(\Theta)$ but also by the effective number of parameters. Since $\bar{D}(\Theta)$ will decrease as the number of parameters in a model increases, the term P_D compensates for this effect by favoring models with a smaller number of parameters. DIC is allowed to be negative as a probability density $p(D|\Theta)$ can be greater than 1 if has a small standard deviation. Only differences in DIC are important. Its absolute size is irrelevant.

REFERENCES

Akaike, H. 1974. A new look at statistical model identification. *IEEE Transactions on Automatic Control* 19, 716-723.

Bayes, T. 1763. An essay towards solving a problem in the doctrine of chances. *Philos. Trans. R. Soc. London* 53, 370–418.

Bishop, C. M. 2006. *Pattern recognition and machine learning.* Springer, pp. 102-105.

Boland, J. 2008. Time Series Modelling of Solar Radiation. In: Badescu, V. (Ed), *Modeling Solar Radiation at the Earth's Surface*, Springer Verlag, pp. 283-312.

Boland, J., Ridley, B. & Brown, B. 2008. Models of diffuse solar radiation. *Renewable Energy* 33, 575-584.

Boland, J., Scott, L. & Luther, M. 2001. Modelling the diffuse fraction of global solar radiation on a horizontal surface. *Environmetrics*, 12, 103-116.

Congdon, P. 2001. *Bayesian statistical modelling*, Wiley and sons.

Cox, R. T. 1946. Probability, frequency and reasonable expectation. *Am. J. Phys.* 14, 1–13.

Erbs, D. G., Klein, S.A. & Duffie J.A. 1982. Estimation of the diffuse radiation fraction for hourly, daily and monthly-average global radiation. *Solar Energy* 28, 293-302.

Geman S. & Geman D. 1984. Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. *IEEE Trans. Pattn. Anal. Mach.Intel* 6, 721-741.

George, E. I. & McCulloch, R. E. 1993. Variable selection via Gibbs sampling. *J. Amer. Statist. Assoc.* 88, 881-889.

Gonzales, J. A. & Calbò, J. 1999. Influence of the gobal radiation variability on the hourly diffuse fraction correlations. *Solar Energy* 65, 119-131.

Jaynes, E. T. 2003. *Probability Theory -The Logic of Science*, Cambridge University Press, Cambridge, UK.

Jaynes, E. T. 1986. *Bayesian Methods: General Background. Maximum Entropy and Bayesian Methods in Applied Statistics*, Cambridge University Press, Cambridge, pp. 1–25.

de Laplace, P. S. 1812. *Théorie Analytique des Probabilités*, Courcier Imprimeur, Paris (in French).

Lauret, P., Miranville, F., Boyer, H., Garde, F. & Adelard, L. 2006a. Bayesian Parameter Estimation of Convective Heat Transfer Coefficients of a Roof-Mounted Radiant Barrier System. *Transactions of ASME, International Journal of Solar Energy Engineering* 128, 213-225.

Lauret, P., David, M., Fock, E., Bastide, A. & Riviere, C. 2006b. Bayesian and Sensitivity Analysis Approaches to Modeling the Direct Solar Irradiance. *Transactions of ASME, International Journal of Solar Energy Engineering* 128, 394-405.

Loredo, T. J. 1990. From Laplace to Supernova SN 1987A: *Bayesian Inference in Astrophysics. Maximum Entropy and Bayesian Methods,* P. Fougere, ed., Kluwer Academic Publishers, Dordrecht, pp. 81–142.

MacKay, D. J. C. 2003. *Information Theory, Inference, and Learning Algorithms,* Cambridge University Press, Cambridge, UK.

Malakoff, D. M. 2005. Bayes offer 'new' way to make sense of numbers. *Science* 286, 1460-1464.

Maxwell, E. L., 1987. A quasi-physical model for converting hourly global horizontal to direct normal insolation. Report SERI/TR-215-3087. Solar Energy Research Institute, Golden, CO.

Orgill, J. F. & Hollands, K. G. T. 1977. Correlation equation for houry diffuse radiation on a horizontal surface. *Solar Energy* 19, 357-359.

Perez R., Ineichen P., Seals R., Michalsky J. & Stewart R. 1990, Modeling Daylight Availibility and Irradiance Components From Direct and Global Irradiance. *Solar Energy* 44(5), 271-289.

Perez, R., Ineichen, P., Maxwell, E., Seals, R. & Zelenka, A. 1992. Dynamic global to direct irradiance conversion models. *ASHRAE Transactions, Research* 3578, 354-369.

Reindl, D. T., Beckman, W. A. & Duffie, J. A. 1990. Diffuse fraction correlations. *Solar Energy* 45, 1-7.

Ridley, B., Boland, J. & Lauret, P. 2009. Modelling of diffuse solar fraction with multiple predictors. *Renewable Energy*, (published online Sept 12 2009).

Schwarz, G., 1978. Estimating the dimension of a model. *Annals of Statistics* 6, 461-464.

Skartveit, A., Olseth, J. A. & Tuft, M. E., 1998. An hourly diffuse fraction model with correction for variability and surface albedo. *Solar Energy* 63, 173-183.

Spencer, J. W. 1982. A comparison of methods for estimating hourly diffuse solar radiation from global solar radiation. *Solar Energy* 29, 19-32.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P. & Van der Linde, A. 2002. Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society, Series B*, 64, 583–639.

Thomas, A., Spiegelhalter, D. J. & Gilks, W. R. 1992. BUGS: A program to perform Bayesian inference using Gibbs sampling. In: Bernardo, J.M., Berger, J.O., Dawid, A.P., Smith, A.F.M. (Eds), *Bayesian Statistics*, Clarendon Press, pp. 837-842.

WinBUGS 1989. http://www.mrc-bsu.cam.ac.uk/bugs/

Wong, L. T. & Chow, W. K. 2001. Solar Radiation Models. *Applied Energy* 69, 191-224.

Correspondence:philippe.lauret@univ-reunion.fr