Monthly Notices
of the
ROYAL ASTRONOMICAL SOCIETY

MNRAS **444,** 1591–1599 (2014) doi:10.1093/mnras/stu1517

A Bayesian self-clustering analysis of the highest energy cosmic rays detected by the Pierre Auger Observatory

Alexander Khanin[1]★ and Daniel J. Mortlock[1,2]

[1]*Astrophysics Group, Blackett Laboratory, Imperial College London, Prince Consort Road, London SW7 2AZ, UK*
[2]*Department of Mathematics, Imperial College London, London SW7 2AZ, UK*

Accepted 2014 July 26. Received 2014 July 11; in original form 2014 March 7

ABSTRACT

Cosmic rays are protons and atomic nuclei that flow into our Solar system and reach the Earth with energies of up to $\sim 10^{21}$ eV. The sources of ultrahigh energy cosmic rays (UHECRs) with $E \gtrsim 10^{19}$ eV remain unknown, although there are theoretical reasons to think that at least some come from active galactic nuclei (AGNs). One way to assess the different hypotheses is by analysing the arrival directions of UHECRs, in particular their self-clustering. We have developed a fully Bayesian approach to analysing the self-clustering of points on the sphere, which we apply to the UHECR arrival directions. The analysis is based on a multistep approach that enables the application of Bayesian model comparison to cases with weak prior information. We have applied this approach to the 69 highest energy events recorded by the Pierre Auger Observatory, which is the largest current UHECR data set. We do not detect self-clustering, but simulations show that this is consistent with the AGN-sourced model for a data set of this size. Data sets of several hundred UHECRs would be sufficient to detect clustering in the AGN model. Samples of this magnitude are expected to be produced by future experiments, such as the Japanese Experiment Module Extreme Universe Space Observatory.

Key words: methods: statistical – cosmic rays.

1 INTRODUCTION

Cosmic rays (CRs) are high-energy particles that flow into our Solar system and reach the Earth. They consist mainly of protons and atomic nuclei, and have energies in the range 10^9–10^{21} eV, which makes them the most energetic particles observed in nature (see e.g. Letessier-Selvon & Stanev 2011 for a review). A number of open issues remain in this field, especially with respect to ultrahigh energy cosmic rays (UHECRs) with arrival energies $E_{\rm arr} \gtrsim 10^{19}$ eV. In particular, no consensus has been reached on the sources of UHECRs. A number of candidates, such as active galactic nuclei (AGNs) and pulsars have been proposed, but lack empirical verification.

The strongest demonstration of the origin of the UHECRs would be if they could be associated with their progenitors, something which is made plausible by the fact that the most energetic CRs can only travel for cosmologically short distances before losing energy. UHECRs with energies of $E \gtrsim 5 \times 10^{19}$ eV scatter off the cosmic microwave background (CMB) radiation via the Greisen–Zatsepin–Kuzmin (GZK) effect (Greisen 1966; Zatsepin & Kuzmin 1966). The resultant energy loss is very significant: the mean free path of the GZK effect at high energies is a few Mpc and the energy loss in each collision is 20–50 per cent depending on energy (Stanev 2009). The GZK effect is expected to cause an abrupt cutoff in the flux of UHECRs at $\sim 4 \times 10^{19}$ eV, for which there has been observational support (Abraham et al. 2008; Bergman 2008). UHECRs that arrive at Earth with energies above the GZK limit can only have come from within a limited radius (the GZK horizon) of $\sim$100 Mpc.

Due to the low flux, the number of detected UHECRs is small: the largest currently available sample is the 69 events with $E \geq 5.5 \times 10^{19}$ eV recorded by the Pierre Auger Observatory (PAO) between 2004 January 1 and 2009 December 31 (Abreu et al. 2010). The low number of events is the main reason why any hypothesis about the sources is difficult to investigate.

Another difficulty is that CRs are charged particles, and so are deflected by magnetic fields. The deflection due to the extra-Galactic magnetic fields is expected to be $\sim$2 – $\sim$10 deg for the highest energy CRs (e.g. Medina Tanco, de Gouveia dal Pino & Horvath 1998; Sigl, Miniati & Enßlin 2004; Dolag et al. 2005). This complicates the study of UHECR origins because it becomes difficult to directly link arrival directions with possible sources.

Nevertheless, a number of attempts have been made to find a correlation between the arrival directions of UHECRs and catalogues of potential sources, although no clear consensus has yet been reached. The Pierre Auger Collaboration reported a strong correlation between the arrival directions of UHECRs with energies $E \geq 5.7 \times 10^{19}$ eV and the positions of nearby AGNs (Abraham et al.

★ E-mail: ak2008@imperial.ac.uk

2007). The result is supported by Yakutsk data (Ivanov 2009), but not by HiRes (Abbasi et al. 2008) or Telescope Array (Abu-Zayyad et al. 2012). A more recent analysis of a larger PAO UHECR sample has shown a much weaker correlation than before (Abreu et al. 2010).

All attempts to associate UHECRs with specific sources are hampered to some degree by large magnetic deflections, possibly transient sources and incomplete catalogues. An alternative approach is based on the idea that if the UHECR sources are distributed inhomogeneously inside the GZK horizon, it should be possible to detect self-clustering in the UHECR arrival directions, independent of any source catalogue. Examples of such work include De Domenico et al. (2011) and Abreu et al. (2012). In Abreu et al. (2012), the Pierre Auger Collaboration studied the self-clustering using three statistical methods based on correlation functions (two methods based on the two-point correlation function, one method based on a three-point correlation function, developed by Ave et al. 2009). No strong evidence of non-uniformity was found based on the p-values obtained under the null hypothesis of no clustering. The interpretation of p-values is, however, known to be problematic as they have no quantitative link to the (posterior) probability that the null hypothesis is correct (see e.g. Berger & Delampady 1987).

Whereas p-values are probabilities conditional on the null hypothesis, what is needed is a method of calculating the probability that the null hypothesis is correct. Cox (1946) proved that Bayesian inference is the only self-consistent method to make probabilistic statements about models based on observations, and Bayesian methods have previously been used to assess whether UHECRs originate from AGNs (Watson, Mortlock & Jaffe 2011; Soiaporn et al. 2012).

In this paper, we present a Bayesian analysis of the self-clustering of the PAO UHECRs. The Bayesian method for assessing non-uniformity is explained in Section 2. In Section 3, the effectiveness of the method is discussed, based on tests of the method on simulated mock UHECR catalogues. The application of the method to data from PAO is discussed in Section 4. Our conclusions are summarized in Section 5.

2 STATISTICAL FORMALISM

Our primary aim here is to assess whether there is evidence that the distribution of UHECR arrival directions is anisotropic. We do this by using Bayesian inference in the context of two models: a uniform model, M_u, which would be the null hypothesis in a classical hypothesis test; and a non-uniform model, M_n, as yet unspecified. The posterior probability of the non-uniform model, conditional on data in the form of N UHECR arrival directions $\{\boldsymbol{r}_i\}$ (where $i \in \{1, 2, \ldots, N\}$), is given by the Bayes's theorem as

$$\Pr(M_\mathrm{n}|\{\boldsymbol{r}_i\}) = \frac{\Pr(M_\mathrm{n})\Pr(\{\boldsymbol{r}_i\}|M_\mathrm{n})}{\Pr(M_\mathrm{u})\Pr(\{\boldsymbol{r}_i\}|M_\mathrm{u}) + \Pr(M_\mathrm{n})\Pr(\{\boldsymbol{r}_i\}|M_\mathrm{n})}, \quad (1)$$

where $\Pr(M_\mathrm{u})$ and $\Pr(M_\mathrm{n})$ are the prior probabilities of the two models, and $\Pr(\{\boldsymbol{r}_i\}|M_\mathrm{u})$ and $\Pr(\{\boldsymbol{r}_i\}|M_\mathrm{n})$ are the probabilities of the data under each of the models (i.e. the likelihoods). With just two models, it is convenient to work with the ratio of the posterior probabilities, given by

$$\frac{\Pr(M_\mathrm{n}|\{\boldsymbol{r}_i\})}{\Pr(M_\mathrm{u}|\{\boldsymbol{r}_i\})} = \frac{\Pr(M_\mathrm{n})}{\Pr(M_\mathrm{u})} B, \quad (2)$$

where

$$B = \frac{\Pr(\{\boldsymbol{r}_i\}|M_\mathrm{n})}{\Pr(\{\boldsymbol{r}_i\}|M_\mathrm{u})} \quad (3)$$

is the Bayes factor. In the convention adopted here, models M_u and M_n are favoured by small and large values of B, respectively.

If a model M has an unspecified parameter θ, then $\Pr(\{\boldsymbol{r}_i\}|M)$ is the marginal likelihood,[1] which is given by

$$\Pr(\{\boldsymbol{r}_i\}|M) = \int_{-\infty}^{\infty} \Pr(\theta|M)\Pr(\{\boldsymbol{r}_i\}|\theta, M)\,\mathrm{d}\theta, \quad (4)$$

where $\Pr(\{\boldsymbol{r}_i\}|\theta, M)$ is the probability of the data for a given value of θ and $\Pr(\theta|M)$ is the prior distribution of the parameter. This distribution must be fully specified and unit-normalized, otherwise the resultant value of $\Pr(\{\boldsymbol{r}_i\}|M)$ is meaningless (Jeffreys 1961).

The next task is to specify the two models to be compared and to evaluate the marginal likelihoods for both. The null hypothesis represented by the uniform model (Section 2.1) is unambiguous and yields the marginal likelihood given in equation (5); the alternative non-uniform model (Section 2.2) is more complicated and is derived from a subset of the data, eventually yielding the marginal likelihood given in equation (9). This requirement means that both marginal likelihoods are evaluated only for the remaining data that was not used to obtain the non-uniform model.

2.1 Uniform model

In the uniform model, M_u, the probability that a UHECR arrives from direction $\boldsymbol{r}$ is constant at $\Pr(\boldsymbol{r}|M_\mathrm{u}) = 1/(4\pi)$. Hence, the marginal likelihood for a test sample of N_t UHECRs with arrival directions $\{\boldsymbol{r}_t\}$ (with $t \in \{1, 2, \ldots, N_\mathrm{t}\}$) is given by

$$\Pr(\{\boldsymbol{r}_t\}|M_\mathrm{u}) = \frac{1}{(4\pi)^{N_\mathrm{t}}}. \quad (5)$$

This simple expression is, however, valid only in the case of uniform exposure; if the exposure is non-uniform, as is always the case for real experiments, it must be modified as described in Section 2.3.

2.2 Non-uniform model

In contrast to the above uniform model, there is an infinite variety of possible non-uniform models that might explain the distribution of UHECR arrival directions. This is a significant conceptual problem: it is difficult to decide which alternative clustered model should be used. To resolve this issue, we develop a multistage, Bayesian approach by splitting the arrival directions $\{\boldsymbol{r}_i\}$ into three subsets:

(i) First, N_g generating points $\{\boldsymbol{r}_g\}$ (with $\mathrm{g} \in \{1, 2, \ldots, N_\mathrm{g}\}$) are chosen as the centres of smooth, localized kernels which can be combined into a mixture distribution on the sphere (Section 2.2.1).

(ii) Then, N_f fitting points $\{\boldsymbol{r}_f\}$ (with $f \in \{1, 2, \ldots, N_\mathrm{f}\}$) are used to obtain a distribution for the unspecified width parameter of the kernels (Section 2.2.2).

(iii) Finally, the remaining N_t testing points $\{\boldsymbol{r}_t\}$ (with $t \in \{1, 2, \ldots, N_\mathrm{t}\}$) are used to evaluate the marginal likelihood under this non-uniform model (Section 2.2.3).

The partitions of the data are chosen at random and the generating points are not linked to the putative UHECR sources in any way. This method is hence independent of any source catalogue or propagation model and, indeed, could be applied to any sample of points on the sphere. The three steps of this approach are illustrated in Fig. 1 for the three test cases described in Section 2.4.

[1] The marginal likelihood is sometimes referred to as the model-averaged likelihood or, particularly in astronomy, as the (Bayesian) evidence.

Figure 1. Full process of model creation for data sets of 69 UHECRs for three test cases: uniform arrival directions (left); three sources (middle); and AGN sources from a realistic mock catalogue (right). Three aspects of the analysis procedure are shown: (1) the full input UHECR data set; (2) partition of the data set into • generating points, • fitting points and ○ testing points; and (3) the resultant mixture distribution of vMF kernels centred on the generating points. The model is created for a range of κ values, but for each of the three test cases, only the maximum likelihood value of κ is displayed here. These highest likelihood values are $\kappa = 0$, 108 and 13 for the uniform sources, three sources, and AGN sources, respectively.

The resultant model (and marginal likelihood) is fully specified, but the algorithm for generating it has two free parameters: N_g and N_f. The relative merits of using a low or high fraction of the data to generate and fit the model (leaving, respectively, a high or low fraction to evaluate the marginal likelihood) is an important area of investigation (e.g. Spiegelhalter & Smith 1982; O'Hagan 1995), but here we take the simplest approach by using a third of the data at each step, so $N_f = N_g = \mathrm{floor}(N/3)$, leaving $N_t = N - (N_f + N_g) \simeq N/3$ testing points. The results of varying these divisions are deferred to a later paper.

The above three-step approach is novel, but similar in principle to the methods of partial or incomplete Bayes factors that have been explored by e.g. Spiegelhalter & Smith (1982), Aitkin (1991), O'Hagan (1991), O'Hagan (1995) and Ghosh, Delampady & Samanta (2006). In all cases, the aim is to evaluate the marginal likelihood for a model with unspecified parameters that do not have strongly motivated priors; and in all cases the basis of the approach is the same as is used here, namely to use part of the data to generate the parameter distributions that are necessary to evaluate the integral in equation (4).

2.2.1 *Generating a clustered model from the data*

The first step to specifying a non-uniform model is to use the N_g generating points $\{\boldsymbol{r}_g\}$ as the centres of smooth, localized kernels of an as yet unspecified angular size.

The specific kernel chosen was the von Mises Fisher (vMF) distribution, which resembles a Gaussian on the sphere and is defined by the density

$$\Pr(\boldsymbol{r}|\overline{\boldsymbol{r}}, \kappa) = \frac{\kappa}{4\pi \sinh(\kappa)} \mathrm{e}^{\kappa \boldsymbol{r}\cdot\overline{\boldsymbol{r}}}, \tag{6}$$

where $\overline{\boldsymbol{r}}$ is the central direction and κ is the concentration parameter. This is inversely related to the width of the distribution: for large values of κ the distribution is peaked over an angular scale of $\sim 1/\sqrt{\kappa}$, while if κ tends to 0 the distribution becomes uniform on the sphere. The vMF distributions were centred on the generating points to give the mixture model density

$$\Pr(\boldsymbol{r}|\{\boldsymbol{r}_g\}, \kappa) = \frac{\kappa}{4\pi N_g \sinh(\kappa)} \sum_{g=1}^{N_g} \mathrm{e}^{\kappa \boldsymbol{r}\cdot\boldsymbol{r}_g}. \tag{7}$$

2.2.2 *Obtaining a concentration distribution*

The last step to fully defining the non-uniform model is to specify a distribution for κ. This is done by using the fitting points to obtain a fully normalized posterior for κ that can be used as a parameter prior in the model comparison step. A uniform prior for $\kappa \geq 0$ is chosen in order to include models with $\kappa = 0$ (which would not be possible for, e.g. a logarithmic prior in κ). The posterior distribution that results from generating points $\{\boldsymbol{r}_g\}$ and fitting points $\{\boldsymbol{r}_f\}$ is

$$\begin{aligned}\Pr(\kappa|\{\boldsymbol{r}_g\}, \{\boldsymbol{r}_f\}) &= \frac{\Pr(\kappa)\Pr(\{\boldsymbol{r}_f\}|\{\boldsymbol{r}_g\}, \kappa)}{\int_0^\infty \Pr(\kappa')\Pr(\{\boldsymbol{r}_f\}|\{\boldsymbol{r}_g\}, \kappa')\,\mathrm{d}\kappa'} \\ &\propto \Theta(\kappa) \prod_{f=1}^{N_f} \Pr(\boldsymbol{r}_f|\{\boldsymbol{r}_g\}, \kappa) \\ &\propto \frac{\Theta(\kappa)\,\kappa^{N_f}}{\sinh^{N_f}(\kappa)} \prod_{f=1}^{N_f} \left(\sum_{g=1}^{N_g} \mathrm{e}^{\kappa \boldsymbol{r}_f\cdot\boldsymbol{r}_g} \right), \end{aligned} \tag{8}$$

where $\Theta(\kappa)$ is the Heaviside step function that encodes the fact that κ is non-negative. The posterior distribution is straightforward to normalize numerically as it is (generally) unimodal and as there is only one parameter.

The alternative, non-uniform model for the UHECR arrival directions is hence fully specified (in the sense of being usable in Bayesian model comparison). It is a sum of vMF distributions centred on the set of generating points, $\{\boldsymbol{r}_g\}$, and with the distribution of vMF concentration parameter κ given by equation (8).

2.2.3 *Evaluating the marginal likelihood*

Having specified the non-uniform model, M_n, with the generating points, $\{\boldsymbol{r}_g\}$ and obtained the distribution $\Pr(\kappa|M_n)$ by using the fitting points, $\{\boldsymbol{r}_f\}$, it is now possible to use the remaining data, the testing points $\{\boldsymbol{r}_t\}$, to evaluate the marginal likelihood. From equation (4) this is

$$\Pr(\{\boldsymbol{r}_t\}|M_n) = \int_0^\infty \Pr(\kappa|M_n)\Pr(\{\boldsymbol{r}_t\}|\kappa, M_n)\,\mathrm{d}\kappa, \tag{9}$$

where $\Pr(\kappa|M_n) = \Pr(\kappa|\{\boldsymbol{r}_g\}, \{\boldsymbol{r}_f\})$ is given in equation (8) and now plays the role of the prior distribution for κ, and the likelihood for the testing points is (cf. equation 7)

$$\begin{aligned}\Pr(\{\boldsymbol{r}_t\}|\kappa, M_n) &= \Pr(\{\boldsymbol{r}_t\}|\{\boldsymbol{r}_g\}, \kappa) \\ &= \prod_{t=1}^{N_t} \Pr(\boldsymbol{r}_t|\{\boldsymbol{r}_g\}, \kappa) \\ &= \frac{\kappa^{N_t}}{[4\pi N_g \sinh(\kappa)]^{N_t}} \prod_{t=1}^{N_t} \left(\sum_{g=1}^{N_g} \mathrm{e}^{\kappa \boldsymbol{r}_t\cdot\boldsymbol{r}_g} \right). \end{aligned} \tag{10}$$

The one-dimensional integral in equation (9) is, once again, straightforward to evaluate numerically. This then gets further modified by the non-uniform exposure, as described in Section 2.3.

2.3 Non-uniform exposure

When studying the measured arrival directions of CRs in a real experiment, the non-uniform exposure of the observatory needs to be taken into account. This is characterized by the relative exposure per unit solid angle, $\mathrm{d}\epsilon/\mathrm{d}\Omega$, defined such that $\int (\mathrm{d}\epsilon/\mathrm{d}\Omega)\,\mathrm{d}\Omega = \epsilon_{\mathrm{tot}}$ is the total exposure.[2] The relative exposure is proportional to $\Pr(\mathrm{det}|\boldsymbol{r})$, the probability that a UHECR arriving from direction $\boldsymbol{r}$ is detected. The distribution of arrival directions of detected CRs is then given by the Bayes theorem as

$$\Pr(\boldsymbol{r}|\mathrm{det}) \propto \Pr(\boldsymbol{r})\Pr(\mathrm{det}|\boldsymbol{r}) \propto \Pr(\boldsymbol{r})\,\frac{\mathrm{d}\epsilon}{\mathrm{d}\Omega}, \tag{11}$$

where $\Pr(\boldsymbol{r})$ is the distribution of arrival directions of all CRs, irrespective of whether they are detected.

For uniform UHECR arrival directions discussed in Section 2.1, $\Pr(\boldsymbol{r}|M_u) = 1/(4\pi)$, so that $\Pr(\boldsymbol{r}, \mathrm{det}|E)$ simply becomes

$$\Pr(\boldsymbol{r}|\mathrm{det}, M_u) = \frac{1}{\epsilon_{\mathrm{tot}}}\frac{\mathrm{d}\epsilon}{\mathrm{d}\Omega}. \tag{12}$$

For the non-uniform UHECR arrival directions discussed in Section 2.2, $\Pr(\boldsymbol{r}|\kappa, M_n)$ is given in equation (7), so that

$$\Pr(\boldsymbol{r}|\mathrm{det}, \kappa, M_n) \propto \frac{\mathrm{d}\epsilon}{\mathrm{d}\Omega} \sum_{g=1}^{N_g} \mathrm{e}^{\kappa \boldsymbol{r}\cdot\boldsymbol{r}_g}, \tag{13}$$

[2] The units of the total exposure are km^2 sr yr.

where the normalization depends on the position of the generating points, $\{\boldsymbol{r}_g\}$, the relative exposure and κ, and must be calculated numerically.

2.4 Illustration of the multistep method

Fig. 1 illustrates how the multistep Bayesian method works for several simple test cases: a uniform source distribution; a model with three sources; and a model based on the AGN simulations described below in Section 3.2. The total number of UHECRs is 69 in all cases. The associated κ posteriors and the resultant distribution of Bayes factors are shown in Fig. 2.

The first test case was a very simple scenario: the UHECRs were simulated with isotropic arrival directions, for the case of uniform exposure. The κ posterior for the uniform case has its maximum very close to 0, and declines rapidly, because the vMF distributions that are fitted to the data are almost uniform. The Bayes factors for this case are small: the uniform model is favoured in 74.1 per cent of the simulations.

The second test case is a simple model of non-uniform arrival directions: the UHECRs were sampled from three vMF distributions, representing three UHECR sources. The concentration parameter κ of the vMF distributions was taken as 90. The κ posterior for this case is systematically peaked at higher values, as can be seen in Fig. 2 A. It is peaked at a value higher than the input value of κ because each of the three original kernels is now accounted for by multiple narrower kernels that are slightly off-centre. The Bayes factors are very large: the non-uniform model is favoured in more than 99.9 per cent of the simulations and the average Bayes factor is $\sim$40.

For the case of three sources, it was also possible to apply an idealized form of the multistep method: instead of using one third of the full data set as the generating points, the generating points were taken as the actual positions of the sources of the UHECRs. In this way, the idealized method does not share the catalogue-independence of the full three-step method described in Section 2.2. For this idealized case, the κ posterior is consistent with the input value, because the three original kernels are accounted for by three kernels located on the original kernel positions. This is also the reason why the Bayes factors are even larger than for the ordinary case. The idealized form of the multistep method is useful to see the potentially strong impact the lack of knowledge about the source positions can have, although it hence cannot be used to analyse real data.

The third test case was the case for UHECRs generated by AGNs, simulated with the realistic model described in Section 3.2. The input value of $\kappa = 360$ was chosen to give the strongest plausible signal, but the resultant posterior is peaked close to $\kappa = 0$. The reason is that there are now so many sources compared to the number of UHECRs that the source distribution is undersampled. This is an indication that, given the weak (projected) clustering expected of nearby AGNs, a significantly larger UHECR sample would be needed for their self-clustering to be apparent. More realistic tests that are documented in Section 3 confirm this result.

3 APPLICATION TO SIMULATED UHECR SAMPLES

To investigate the effectiveness of the multistage Bayesian method described above, it was applied to realistic mock catalogues of UHECRs. Catalogues were created for two different UHECR scenarios: isotropic (Section 3.1) and AGN centred (Section 3.2). The samples of incoming UHECRs were then subjected to the PAO measurement process (Section 3.3). The distributions of Bayes factors for the resultant observed samples are analysed in Section 3.4.

3.1 Isotropic distribution of sources

The application of the multistep method to uniform UHECR distributions acted as a false positive test. Computing large numbers of Bayes factors for uniform UHECR distributions can be used to establish how often the null hypothesis is wrongly rejected.

3.2 AGN sources

Simulated UHECR catalogues were created for the case of UHECRs originating in AGNs. The simulation encompassed two main components: the injection of the UHECRs at the sources and a propagation model.

3.2.1 Injection at the sources

The AGN sources were drawn randomly from the simulated Las Damas 'Consuelo' catalogues,[3] following a similar procedure to Berlind et al. (2011).

Two source densities were used: $10^{-3.5}$ and $10^{-4.5}\,\mathrm{Mpc}^{-3}$. These are the highest and lowest source densities available in the Consuelo catalogues, and represent a reasonable range of possible source densities.

The injection spectrum of the UHECRs at the sources is assumed to be a power-law of the form $Q(E) \propto E^{-\alpha}$, where $Q(E)\,\mathrm{d}E$ is the number of cosmic rays emitted with energy between E and $E + \mathrm{d}E$ per unit time, and α is the power-law index. Simulations were conducted for three realistic values of the index: 2.0, 2.3 and 2.7, spanning the range of values used in e.g. De Domenico & Insolia (2013), Abreu et al. (2013), Ahlers & Salvado (2011) and Decerprit & Allard (2011).

3.2.2 Propagation model

Both the energy loss that the UHECRs experience during propagation and their magnetic deflection must be accounted for. The deflection is not treated explicitly, but included in the observational smearing described in Section 3.3; the energy loss model is described here.

A pure proton composition of UHECRs was assumed and so the energy loss during propagation consists of three components (e.g. Stanev 2009):

(i) The GZK scattering off the CMB photons at energies above $E \gtrsim 5 \times 10^{19}\,\mathrm{eV}$;

(ii) Bethe Heitler $\mathrm{e}^+\mathrm{e}^-$ pair production (also a scattering process off the CMB radiation), which dominates at lower energies (Hillas 1968);

(iii) The adiabatic energy loss due to the expansion of the Universe.

Our implementation of this propagation model includes the BH and adiabatic losses in a continuous approximation, and treats the GZK effect as a stochastic process.

[3] http://ls.phy.vanderbilt.edu/lasdamas. The catalogues have been compiled by the Las Damas collaboration, Andreas Berlind et al.

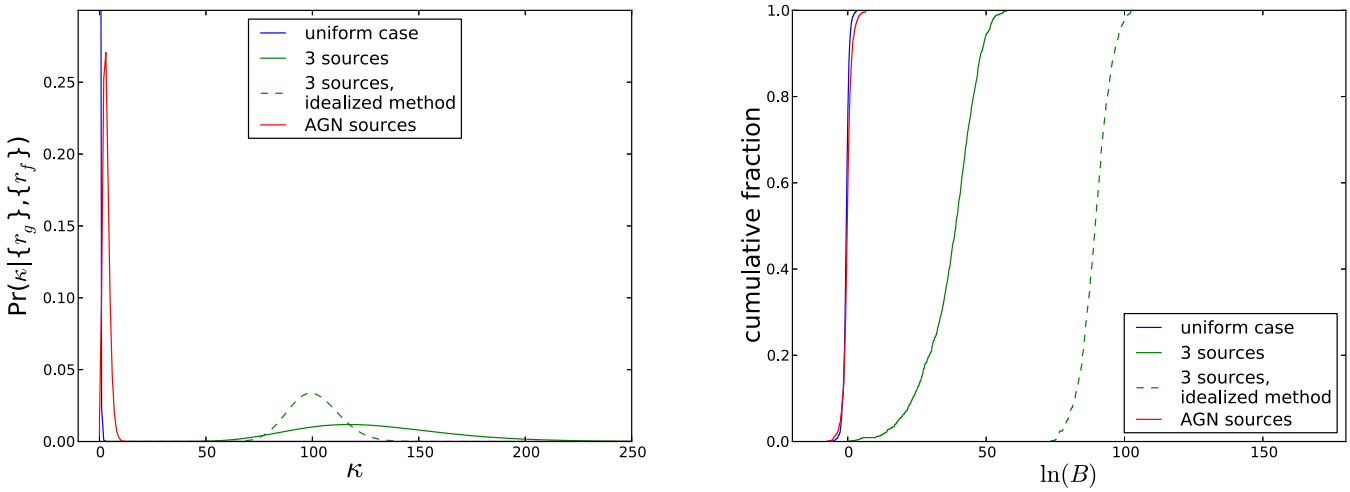

Figure 2. (A) Kappa posteriors and (B) cumulative fractions of Bayes factors, produced by the application of the multistep method to test cases of 69 UHECR events. Three test cases are considered: uniform UHECRs; UHECRs generated from three sources; and UHECRs generated by AGNs from a realistic catalogue. In the case of three sources, in addition to the conventional application of the multistep method, the results for an idealized method are displayed. In the idealized application of the method, the generating points are taken as the true centres of the vMF distributions that generate the UHECRs, rather than as a random subset of the data.

3.3 Measurement

All of the simulations were done for a PAO-like experiment, three aspects of which were modelled explicitly:

(i) PAO's non-uniform exposure was taken into account by accepting arriving UHECRs with a probability proportional to the relative exposure $\mathrm{d}\epsilon/\mathrm{d}\Omega$ defined in Section 2.3.

(ii) The error in PAO's energy measurement is about 12 per cent (Letessier-Selvon et al. 2013), and was included in the model. This is significant as only UHECRs that have an observed energy above a fixed threshold are included in the simulated samples.

(iii) The angular resolution of PAO varies from about 2.2 deg to about 1 deg for the lowest and highest energies, respectively (Abreu et al. 2012). The magnetic deflection that the UHECRs experience during propagation also means that their arrival directions are offset from the source. The magnitude of this effect is uncertain, the estimates of typical deflection angles ranging from ∼2 to ∼10 deg for the highest energy UHECRs (e.g. Medina Tanco et al. 1998; Sigl et al. 2004; Dolag et al. 2005). These two effects are simulated together by drawing a measured arrival direction from a vMF distribution centred on the source. We used three different values for the concentration parameter κ of the vMF distributions: 30, 90 and 360, which correspond to average angular deviations of approximately 10, 6 and 3 deg, respectively.

We treated the magnetic deflection as a simple smearing, rather than including detailed simulations of the Galactic and extra-Galactic magnetic fields, because our aim was to assess the arrival directions without reference to a particular physical model. Detailed models of the magnetic fields are available (De Domenico & Insolia 2013; Farrar 2014 and references therein), and a formalism for incorporating these into a Bayesian UHECR analysis has been developed in Soiaporn et al. (2012).

3.4 Results of the simulations

Simulations were performed and Bayes factors evaluated for the isotropic model, and for the AGN-centred model with 18 combinations of the above parameters:

(i) source densities of $10^{-3.5}$ Mpc^{-3} and $10^{-4.5}$ Mpc^{-3};
(ii) injection parameters α of 2.0, 2.3, 2.7;
(iii) concentration parameters κ of 30, 90, 360.

For each of the 18 combinations of parameters, 1000 samples of 69 UHECRs were created (matching the size of the PAO data set). For each sample, Bayes factors were computed for each of three energy thresholds: 5.5×10^{19} eV, 8.0×10^{19} eV and 10×10^{19} eV. Including the 1000 realizations of the isotropic model, 55 000 Bayes factors were computed in total.

The results of these simulations are shown as cumulative distributions of Bayes factors in the top half of Fig. 3. These are compared to similar cumulative distributions for the case of uniformly distributed UHECRs.

The Bayes factors tend to be larger for the source-centred case than for the uniform case. The difference between the results for uniform and non-uniform UHECRs is greater for the case of low source density, as for higher source density the UHECR distribution would eventually tend to a uniform distribution.

Furthest away from the uniform case is the model with the lowest source density, highest κ and highest α. Higher κ means that the UHECR arrival directions are more closely correlated with the positions of the sources. High α reduces the GZK horizon, meaning fewer contributing AGN sources and hence more non-uniformity.

The threshold energy value does not have a substantial effect on the distribution of Bayes factors. It is difficult to predict the effect of the threshold energy qualitatively, because there are two competing effects: a lower threshold would increase the sample size, which makes the non-uniformity more apparent; a higher threshold decreases the effective GZK horizon, which would increase the non-uniformity signature. This means that there is some ideal threshold that gives the greatest chance of detecting whatever anisotropy is present.

While the results for the uniform and non-uniform cases are clearly different, the difference is not very significant. If we take a threshold value of $\ln(B) = 5$ to represent a decisive detection, then anisotropy is detected only for 0.002 per cent and 5 per cent of the samples for source densities of $10^{-3.5}$ Mpc^{-3} and $10^{-4.5}$ Mpc^{-3},

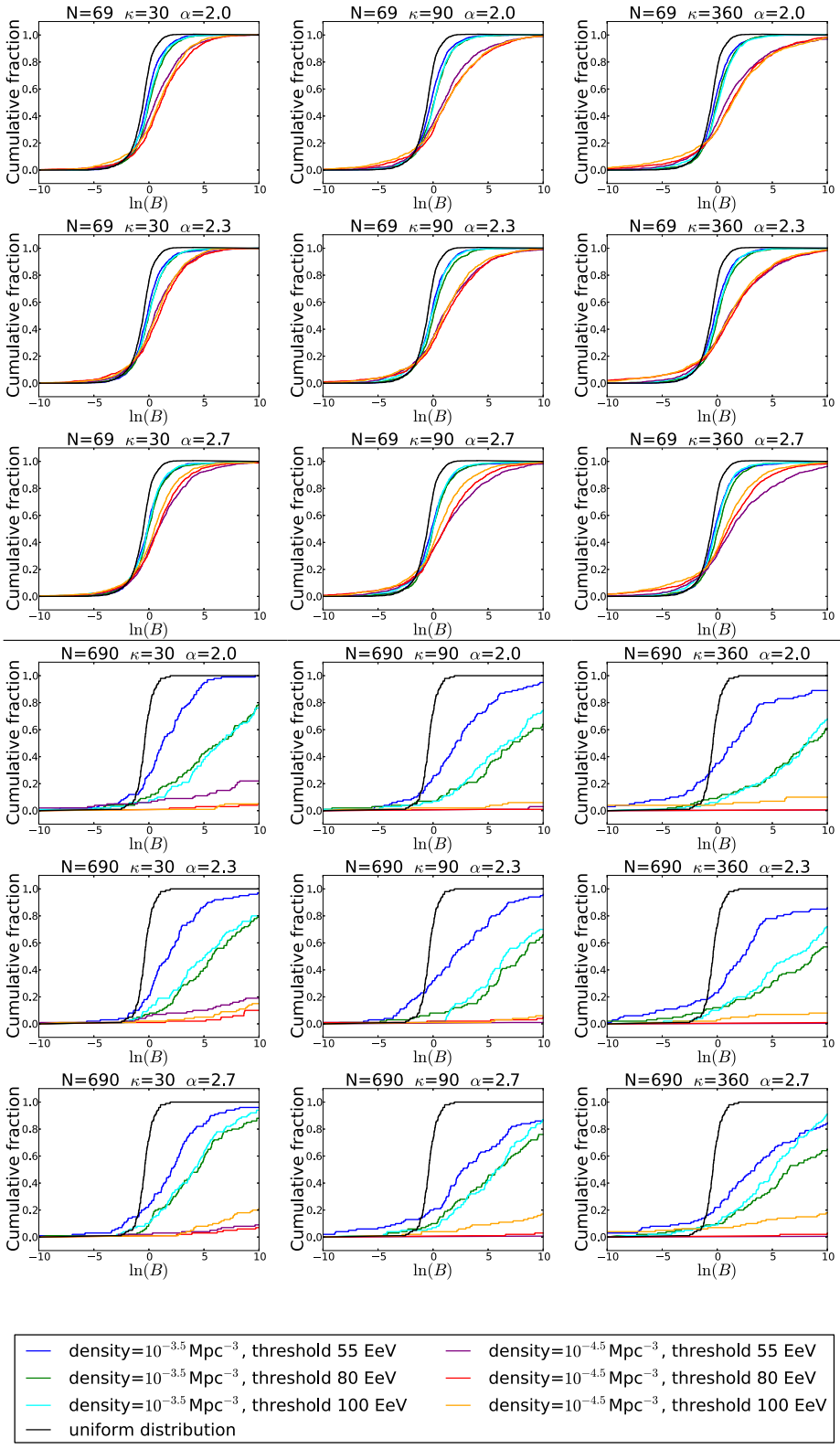

Figure 3. Results of the multistep method applied to mock UHECR catalogues. Cumulative distributions of Bayes factors have been produced for three energy thresholds, two source densities, and for different values of the sample size N, the injection parameter α and the concentration parameter κ, as indicated above.

respectively. The conclusion is that the clustering expected from a realistic model of AGN-sourced UHECRs is too weak to be detected from a sample of 69 events. This is consistent with the results of Abreu et al. (2012).

The simulations were repeated for 100 samples of $N = 690$ UHECRs (i.e. $10 \times$ the PAO sample). The results are shown in the bottom half of Fig. 3. The difference between the uniform and non-uniform cases becomes very apparent for all combinations of parameters. For source densities of $10^{-3.5}\,\mathrm{Mpc}^{-3}$ and $10^{-4.5}\,\mathrm{Mpc}^{-3}$, 22 and 93 per cent of the Bayes factors are above the threshold of $\ln(B) = 5$. UHECR samples of 690 events are sufficient to detect self-clustering for a realistic model.

We assume a pure proton composition of UHECRs, which is consistent with the results of HiRes (Abbasi et al. 2005), but not fully consistent with the results of PAO, which indicate a more complex mixed nuclear composition (Unger 2008), including heavier nuclei such as iron. For iron, the magnetic deflection angle would be increased by a factor of 26, leading to a deflection of $\sim 50 - \sim 250$ deg. This makes it more difficult to associate the UHECRs with specific sources. However, the detection of clustering is also made easier by the fact that heavier nuclei lose more energy through additional scattering processes, which reduces the GZK horizon and thus the number of candidate sources. The energy loss length for cosmic rays with $E \gtrsim 5 \times 10^{20}$eV is reduced from $\sim$10 MeV for protons to $\sim$2 MeV for iron, which reduces the GZK horizon by a factor of $\sim$5 (Stanev 2009). The net effect of these two factors will need to be established through additional simulations.

4 ANALYSIS OF THE PAO DATA

We now apply the multistage Bayesian method described in Section 2 to the PAO data set in order to assess the uniformity of the measured UHECR arrival directions. This data set consists of 69 events observed from 2004 January 1 to 2009 December 31, and is described in full by Abreu et al. (2010). As the results depend to some extent on the way the data are split into the three subsets, Bayes factors were calculated for 1 000 different random, but equal sized, partitions. The cumulative distribution of Bayes factors is shown in Fig. 4.

The Bayes factors are calculated for different partitions of the same sample. Apart from the distribution for the PAO data, Fig. 4 also shows the distribution for a uniform sample of 69 UHECRs, as well as the distribution for a UHECR sample generated from a realistic AGN catalogue (with a source density of $10^{-3.5}\,\mathrm{Mpc}^{-3}$, $\kappa = 30$ and $\alpha = 2.0$). The results shown here differ from those shown in Fig. 3, insofar as they result from different random partitions of a single sample (i.e. PAO, uniform or AGN-sourced) rather than being drawn from completely independent samples. However, the distributions produced using these two methods are comparable and the main conclusions remain unchanged.

A sensible way of dealing with the range of Bayes factors is to characterize their distribution by the arithmetic or geometric mean. There is no compelling reason to choose one over the other (see e.g. O'Hagan 1997), but the fact that the logarithm of the Bayes factor is symmetric between the two models suggests that the geometric mean is more natural. The geometric mean was 0.57 and the arithmetic mean was 1.26. From equation (1), if we assume a prior probability of 0.5 for both models, we calculate mean posterior probabilities for the clustered model of 0.37 and 0.56 for the respective means. Thus, there is no clear preference for either of the models, and the data are consistent with both. We do not detect evidence for self-clustering. Fig. 4 shows that for data sets of this size, the distributions of Bayes factors for the uniform and AGN-centred cases cannot be clearly distinguished. This is consistent with the results of Abreu et al. (2012).

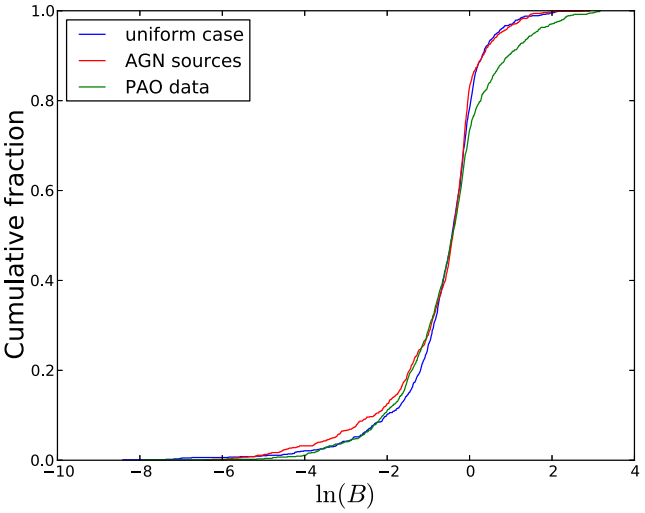

Figure 4. Cumulative fractions of Bayes factors, produced by the application of the multistep method to 1 000 partitions of: (a) the PAO data; (b) 69 simulated UHECRs from uniform sources; and (c) 69 simulated UHECRs from a realistic mock catalogue of AGNs.

5 CONCLUSIONS

We have developed a Bayesian method for the analysis of the self-clustering of points on a sphere and applied it to the 69 highest energy UHECRs detected by PAO up until 31 December 2009.

The method is a three-step Bayesian approach, in which the data are divided into three subsets: the first two subsets of the data are used to generate a model of self-clustered UHECRs; the third subset is used to perform Bayesian model comparison between this self-clustered model and a uniform model of UHECRs. This approach is an extension of the Bayesian model comparison methods that were developed by Spiegelhalter & Smith (1982), Aitkin (1991), O'Hagan (1991) and O'Hagan (1995). Like the multistep method that is presented here, those approaches are aimed to evaluate the marginal likelihood in cases when there is weak prior information on the model parameters. The method we have presented here is not specific to the UHECR problem in question and could be applied to anisotropy searches in other areas of astronomy, such as the search for angular anisotropies in the distribution of gamma-ray bursts described by e.g. Balazs, Meszaros & Horvath (1998) and Magliocchetti, Ghirlanda & Celotti (2003).

There is some ambiguity in the partitioning of the full data set. In the present implementation, the total data set is divided into three subsets of equal size. However, it is possible that a different partitioning, or perhaps an average over partitions could make this method more effective. These issues will be explored in future work.

We tested our model comparison method on mock catalogues of UHECRs. The results for uniform UHECR arrival directions were compared to the results for UHECRs originating in AGNs from a realistic mock catalogue. UHECR clustering in a realistic AGN centred model is too weak to be detected in a sample of 69 events, but would be detectable in samples of 690 events. This is consistent with the results of Abreu et al. (2012).

We assumed a pure proton composition of the cosmic rays, but there are some indications that heavier nuclei are also part of the

composition (Unger 2008). The effect of including heavier nuclei will be investigated through additional simulations.

For the PAO data, Bayes factors were calculated for different random partitions of the data. The geometric and arithmetic means of the Bayes factors were 0.57 and 1.26, respectively, corresponding to posterior probabilities of 0.37 and 0.56 for the clustered model. Thus, we did not find strong evidence for clustering in the PAO data, although the data are also consistent with the AGN-centred simulations.

It is expected that future experiments will produce data sets that will be sufficiently large for our Bayesian method (and other statistical approaches; see e.g. Rouillé d'Orfeuil et al. 2014) to detect even the weak clustering expected if the UHECRS have come from nearby AGNs. PAO is continuing to take data and is expected to produce a sample of ∼250 UHECRs over its first decade of operations. Looking further ahead, the planned Japanese Experiment Module Extreme Universe Space Observatory (JEM-EUSO; Adams et al. 2013) on the International Space Station is scheduled for launch in 2017 and is expected to detect ∼200 UHECRs annually over its five year lifetime. These data sets should be sufficiently large to detect the self-clustering of UHECRs independent of the source population.

ACKNOWLEDGEMENTS

We thank Andreas Berlind and Glennys Farrar for making their mock catalogues public and Todor Stanev for providing the results of his UHECR propagation models. We thank an anonymous referee for making suggestions which allowed us to improve this paper. AK was supported by a Science & Technology Facilities Council studentship.

REFERENCES

Abbasi R. U. et al., 2005, ApJ, 622, 910
Abbasi R. U. et al., 2008, Astropart. Phys., 30, 175
Abraham J. et al., 2007, Science, 318, 938
Abraham J. et al., 2008, Phys. Rev. Lett., 101, 061101
Abreu P. et al., 2010, Astropart. Phys., 34, 314
Abreu P. et al., 2012, J. Cosmol. Astropart. Phys., 4, 40
Abreu P. et al., 2013, J. Cosmol. Astropart. Phys., 05, 009
Abu-Zayyad T. et al., 2012, ApJ, 757, 26
Adams J. H., Jr et al., 2013, Proc. 33rd International Cosmic Ray Conference. Rio de Janeiro, Brazil, preprint (arXiv:1307.7071)
Ahlers M., Salvado J., 2011, Phys. Rev. D, 84, 085019
Aitkin M., 1991, J. R. Stat. Soc. B, 53, 111
Ave M. et al., 2009, J. Cosmol. Astropart. Phys., 7, 23
Balazs L. G., Meszaros A., Horvath I., 1998, A&A, 339, 1
Berger J. O., Delampady M., 1987, Stat. Sci., 2, 317
Bergman D. R., 2008, in Caballero R., D'Olivo J. C., Medina-Tanco G., Nellen L., Sáánchez F. A., Valdés-Galicia J. F., eds, Proc. 30th Int. Cosmic Ray Conf., Vol. 4, Observation of the GZK Cutoff by the HiRes Experiment. Universidad Nacional Autonoma de Mexico, Mexico City, Mexico, p. 451
Berlind A., Busca N., Farrar G. R., Roberts J. P., 2011, preprint (arXiv:1112.4188)
Cox R. T., 1946, Am. J. Phys., 14, 1
De Domenico M., Insolia A., 2013, J. Phys. G: Nucl. Phys., 40, 015201
De Domenico M., Insolia A., Lyberis H., Scuderi M., 2011, J. Cosmol. Astropart. Phys., 3, 8
Decerprit G., Allard D., 2011, A&A, 535, A66
Dolag K., Grasso D., Springel V., Tkachev I., 2005, J. Cosmol. Astropart. Phys., 1, 9
Farrar G. R., 2014, C. R. Phys., 15, 339
Ghosh J., Delampady M., Samanta T., 2006, An Introduction to Bayesian Analysis: Theory and Methods. Springer-Verlag, Berlin
Greisen K., 1966, Phys. Rev. Lett., 16, 748
Hillas A. M., 1968, Can. J. Phys., 46, 623
Ivanov A. A., 2009, Nucl. Phys. B, 190, 204
Jeffreys H., 1961, Theory of Probability. Oxford Univ. Press, Oxford
Letessier-Selvon A., Stanev T., 2011, Rev. Mod. Phys., 83, 907
Letessier-Selvon A. et al., 2013, Proc. 33rd International Cosmic Ray Conference, Rio de Janeiro, Brazil, preprint (arXiv:1310.4620)
Magliocchetti M., Ghirlanda G., Celotti A., 2003, MNRAS, 343, 255
Medina Tanco G. A., de Gouveia dal Pino E. M., Horvath J. E., 1998, ApJ, 492, 200
O'Hagan A., 1991, J. R. Stat. Soc. B, 53, 136
O'Hagan A., 1995, J. R. Stat. Soc. B, 57, 99
O'Hagan A., 1997, TEST, 6, 101
Rouillé d'Orfeuil B., Allard D., Lachaud C., Parizot E., Blaksley C., Nagataki S., 2014, A&A, 567, A81
Sigl G., Miniati F., Enßlin T. A., 2004, Phys. Rev. D, 70, 043007
Soiaporn K., Chernoff D., Loredo Th., Ruppert D., Wasserman I., 2012, Ann. Appl. Stat., 7, 1249
Spiegelhalter D. J., Smith A. F. M., 1982, J. R. Stat. Soc. B, 44, 377
Stanev T., 2009, New J. Phys., 11, 13
Unger M., 2008, in Caballero R., D'Olivo J. C., Medina-Tanco G., Nellen L., Sáánchez F. A., Valdés-Galicia J. F., eds, Proc. 30th Int. Cosmic Ray Conf., Vol. 4, Study of the Cosmic Ray Composition above 0.4 EeV using the Longitudinal Profiles of Showers observed at the Pierre Auger Observatory. Universidad Nacional Autonoma de Mexico, Mexico City, Mexico, p. 373
Watson L. J., Mortlock D. J., Jaffe A. J., 2011, MNRAS, 418, 208
Zatsepin G., Kuzmin V., 1966, JETP Lett., 4, 78

This paper has been typeset from a TEX/LATEX file prepared by the author.