Quantifying tensions between CMB and distance data sets in models with free curvature or lensing amplitude

S. Grandis,[1,2]★ D. Rapetti,[1,2,3,4] A. Saro,[1,2] J. J. Mohr[1,2,5] and J. P. Dietrich[1,2]

[1]*Faculty of Physics, Ludwig-Maximilians-Universität, Scheinerstr. 1, D-81679 Munich, Germany*
[2]*Excellence Cluster Universe, Boltzmannstr. 2, D-85748 Garching, Germany*
[3]*Center for Astrophysics and Space Astronomy, Department of Astrophysical and Planetary Science, University of Colorado, Boulder, C0 80309, USA*
[4]*NASA Ames Research Center, Moffett Field, CA 94035, USA*
[5]*Max Planck Institute for Extraterrestrial Physics, Giessenbachstr., D-85748 Garching, Germany*

ABSTRACT

Recent measurements of the cosmic microwave background (CMB) by the Planck Collaboration have produced arguably the most powerful observational evidence in support of the standard model of cosmology, i.e. the spatially flat ΛCDM paradigm. In this work, we perform model selection tests to examine whether the base CMB temperature and large scale polarization anisotropy data from Planck 2015 (P15; Planck Collaboration XIII) prefer any of eight commonly used one-parameter model extensions with respect to flat ΛCDM. We find a clear preference for models with free curvature, $\Omega_{\rm K}$, or free amplitude of the CMB lensing potential, $A_{\rm L}$. We also further develop statistical tools to measure tension between data sets. We use a Gaussianization scheme to compute tensions directly from the posterior samples using an entropy-based method, the surprise, as well as a calibrated evidence ratio presented here for the first time. We then proceed to investigate the consistency between the base P15 CMB data and six other CMB and distance data sets. In flat ΛCDM we find a 4.8σ tension between the base P15 CMB data and a distance ladder measurement, whereas the former are consistent with the other data sets. In the curved ΛCDM model we find significant tensions in most of the cases, arising from the well-known low power of the low-ℓ multipoles of the CMB data. In the flat ΛCDM+$A_{\rm L}$ model, however, all data sets are consistent with the base P15 CMB observations except for the CMB lensing measurement, which remains in significant tension. This tension is driven by the increased power of the CMB lensing potential derived from the base P15 CMB constraints in both models, pointing at either potentially unresolved systematic effects or the need for new physics beyond the standard flat ΛCDM model.

Key words: methods: statistical – cosmic background radiation – cosmological parameters – cosmology: observations – distance scale.

1 INTRODUCTION

Over the last two decades, growing observational evidence has been collected in support of a model with a flat geometry, cold dark matter (CDM) and a cosmological constant, Λ. This model has been extremely successful in the face of observational constraints from a wide variety of data sets, such as temperature and anisotropy measurements of the cosmic microwave background (CMB; Bennett et al. 2013; Hinshaw et al. 2013; Planck Collaboration I 2014; Planck Collaboration I 2015a; Planck Collaboration XIII 2015b; Planck Collaboration XI 2015c). It also accurately predicts measurements of the cosmic distance ladder (Riess et al. 2011; Efstathiou 2014; Aubourg et al. 2015; Riess et al. 2016), supernovae type Ia (Conley et al. 2011; Betoule et al. 2013), baryon acoustic oscillations (BAO; Beutler et al. 2011; Anderson et al. 2014; Delubac et al. 2015; Ross et al. 2015), cluster gas mass fraction (Allen et al. 2008; Mantz et al. 2014), cosmic shear correlation function (Kilbinger et al. 2013; Mandelbaum et al. 2013; DES Collaboration 2015), CMB lensing (Das et al. 2011; van Engelen et al. 2012; Planck Collaboration XV 2015d), and cluster number counts (Mantz et al. 2008; Vikhlinin et al. 2009; Benson et al. 2013; Hasselfield et al. 2013; Bocquet et al. 2015; Mantz et al. 2015; Planck Collaboration XXIV 2015e; de Haan et al. 2016).

Despite some recently discussed tensions concerning the value of the present day Hubble parameter (see for instance Verde, Protopapas & Jimenez 2013, 2014; Bennett et al. 2014; Riess et al. 2016) and the power of scalar fluctuations as measured from the CMB and large-scale structure probes (see, among others, Hamann & Hasenkamp 2013; Battye & Moss 2014; Raveri 2015; Grandis

★E-mail: s.grandis@lmu.de

et al. 2016; Joudaki et al. 2016), the constraints from all these observations seem to agree reasonably well with each other in this model.

To test various key assumptions of the flat ΛCDM model, we consider a series of one parameter extensions to this model and investigate whether the increased complexity of the extended models is needed to improve the goodness of fit to the data. In this work we show that the temperature and large scale polarization CMB anisotropy measurements, i.e. the base CMB constraints, of the Planck Collaboration XIII (2015b) and Planck Collaboration XI (2015c) (hereafter called base P15 CMB) prefer a ΛCDM model with free curvature, $\Omega_{\rm K}$, or free lensing potential amplitude, $A_{\rm L}$. Besides model selection, tensions between different data sets within the same model can also indicate that the assumed model is not adequate. We examine the level of agreement between the base CMB constraints and different additional data sets in the flat ΛCDM, curved ΛCDM, and flat ΛCDM+$A_{\rm L}$ models. In flat ΛCDM, we find mostly consistency among the data sets we consider with the exception of a significant tension with a recent distance ladder measurement, but this is no longer true in the two extended models we examine. We find that the base P15 CMB constraints are in significant tension with most external data sets in curved ΛCDM, whereas in flat ΛCDM+$A_{\rm L}$, only the CMB lensing data show a significant disagreement with the P15 CMB constraints.

To perform these data consistency tests, we have developed different statistical tools. Generalizing the Gaussianization scheme of Schuhmann, Joachimi & Peiris (2016), for each model we consider, we find a transformation of parameter space that maps on to Gaussian distributions both the P15 CMB constraints alone and these combined with one external data set. We then measure the degree of tension introduced by these combinations using the entropy based 'surprise', which was introduced by Seehars et al. (2014, 2016) to measure the consistency of an historical sequence of CMB surveys, and employed by Grandis et al. (2016) to demonstrate the agreement of different external data sets with the *Wilkinson Microwave Anisotropy Probe* (*WMAP*; Bennett et al. 2013; Hinshaw et al. 2013). The Gaussianization procedure is crucial to test the consistency between data sets in models with strong parameter degeneracies, as it allows one to analytically approximate their constraints. This provides an important test, which we argue should be systematically performed when combining data sets.

We also investigate the statistical properties of evidence ratios, a widely used measure of data set agreement (see Marshall, Rajguru & Slosar 2006; Amendola, Marra & Quartin 2013; Heneka, Marra & Amendola 2014; Karpenka, Feroz & Hobson 2014; Martin et al. 2014; Raveri 2015). We demonstrate theoretically and with simple examples that evidence ratios can be highly biased and therefore need to be accurately calibrated. We also compare calibrated evidence ratios to the surprise results, and find that they give very comparable measures of the significance of the tension.

We organize the paper as follows. In Section 2, we discuss the statistical tools employed. In Section 3, we present the data sets used in our analysis. We then report our results on model selection and data set consistency in Section 4, discussing the impact of systematics and choices of priors in Section 5, which also contains a discussion of the physical effects responsible for the deviation from flatness or from $A_{\rm L} = 1$.

2 STATISTICAL METHODS

Cosmological constraints on a specific model, M, derived from astrophysical data, D, are usually expressed as a posterior distribution $p(\boldsymbol{\theta}|D, M)$ on the space of cosmological parameters $\boldsymbol{\theta}$. Posterior distributions can be obtained by using the Bayes' Theorem as

$$p(\boldsymbol{\theta}|\, D, M) = \frac{L(D|\,\boldsymbol{\theta}, M)}{E(D|\, M)}\, p(\boldsymbol{\theta}), \tag{1}$$

where $p(\boldsymbol{\theta})$ is a prior, $L(D|\,\boldsymbol{\theta},\ M)$ the likelihood and $E(D|M)$ the evidence.

2.1 Gaussianization

In some models, the posterior distribution displays significant departures from Gaussianity. This complicates both a possible analytic approximation of the posterior as well as the comparison with other posterior distributions. However, as explicitly shown by Schuhmann et al. (2016), a suit of optimized transformations of the parameters can efficiently map a generic uni-modal distribution on to a Gaussian distribution. This allows one to analytically approximate the distribution, significantly speeding up its evaluation. For details on the precision of this approximation, see Appendix A.

Here we generalize the Gaussianization method proposed by Schuhmann et al. (2016) to simultaneously Gaussianize two distributions. Such a joint Gaussianization will allow us to compare the two distributions analytically. For details, see also Appendix A. In the following, we present the statistical tools we employ to quantify comparisons between data sets (Section 2.2) and between models (Section 2.3).

2.2 Quantifying tension

Given the variety of cosmological data sets, it is of great importance to assess their mutual agreement. The absence of this agreement is usually referred to as 'tension' between data sets. We first discuss an entropy based method to measure these tensions and then an evidence ratio based one.

2.2.1 Entropy-based method

To quantify the consistency of a data set D_1 with another data set D_2 we can use the Kullback–Leibler divergence, also called relative entropy, introduced by Kullback & Leibler (1951),

$$KL[D_2|\, D_1] = \int \mathrm{d}^d\boldsymbol{\theta}\ p(\boldsymbol{\theta}|\, D_1, D_2, M) \ln\left(\frac{p(\boldsymbol{\theta}|\, D_1, D_2, M)}{p(\boldsymbol{\theta}|\, D_1)}\right), \tag{2}$$

where $p(\boldsymbol{\theta}|\, D_1)$ is the posterior distribution of the data set D_1, which we employ as a prior for updating the joint posterior of the two data sets $p(\boldsymbol{\theta}|\, D_1, D_2, M)$.

As discussed elsewhere (Seehars et al. 2014, 2016; Grandis et al. 2016), the relative entropy depends on the data sets D_1 and D_2, and as such has an expected value $\langle KL\rangle_{D_2|D_1}$ and a mean fluctuation around this value $\sigma(KL)$, which depends on the expected distribution of the data set D_2 given the prior $p(\boldsymbol{\theta}|\, D_1)$. The difference between the actual relative entropy and the expected relative entropy is defined by Seehars et al. (2014) as the *surprise* $S = KL[D_2|\, D_1] - \langle KL\rangle_{D_2|D_1}$. If the surprise is negative, $S < 0$, the data set D_2 is in better agreement with the prior than expected; if the surprise is positive, the data set D_2 is in worse agreement with the prior than expected. Comparing the surprise S to its expected fluctuation $\sigma(KL)$ allows one to estimate the significance of the underlying tension (see Seehars et al. 2014, 2016, for more details).

The relative entropy is invariant under transformations in parameter space (for proof see appendix B in Grandis et al. 2016), and it

is analytic if prior and posterior are multivariate Gaussian distributions (see Seehars et al. 2014). Thus, it can be easily estimated for two generic distributions after a joint Gaussianization. As shown by Seehars et al. (2014, 2016), in this case it will be given by

$$S = \frac{1}{2}\mathbf{\Delta\mu}^T \mathbf{C}_{\rm pr}^{-1}\mathbf{\Delta\mu} - \frac{1}{2}\mathrm{tr}\left(\mathbb{I} - \mathbf{C}_{\rm po}\mathbf{C}_{\rm pr}^{-1}\right), \quad (3)$$

where $\mathbf{\Delta\mu}$ is the difference in means of the transformed distributions, $\mathbf{C}_{\rm pr}$ and $\mathbf{C}_{\rm po}$ the covariances of the transformed prior and posterior respectively, 'tr' stands for trace, and $\mathbb{I}$ is the identity matrix. In this case, the variance of the relative entropy is given by $\sigma^2(KL) = \mathrm{tr}[(\mathbf{C}_{\rm pr}^{-1}\mathbf{C}_{\rm po} - \mathbb{I})^2]/2$. Thus, given estimates of covariances and means for prior and posterior, these quantities can be easily estimated. Note that all entropy based results are given in units of 'bits' by normalizing with $\ln 2$.

As can be seen from equation (3), the surprise measures the shift in the mean values $\mathbf{\Delta\mu}$ created by the update of $p(\boldsymbol{\theta}|\, D_1)$ with D_2, and asses how significant this shift is by comparing it to the expected fluctuation $\sigma(KL)$. Consequently, it is well suited to test whether D_2 should be added to the constraints of D_1.

2.2.2 *Calibrated evidence ratio*

A standard way (see Marshall et al. 2006; Amendola et al. 2013; Heneka et al. 2014; Karpenka et al. 2014; Martin et al. 2014; Raveri 2015) of assessing the degree of agreement between two data sets D_1, D_2 is given by the so called *evidence ratio*

$$R = \frac{E(D_1, D_2)}{E(D_1)\,E(D_2)}, \quad (4)$$

where $E(D_1, D_2)$ is the joint evidence of the two data sets D_1 and D_2, and $E(D_1)$ and $E(D_2)$ are the evidences of the individual data sets.

This ratio is interpreted using the Jeffreys' scale introduced by Jeffreys (1961), where $\ln R > 0$ indicates agreement and $\ln R < 0$ indicates inconsistency. However, as pointed out by Seehars et al. (2016), this interpretation does not take into account the statistical behaviour of the evidence ratio. For this sake, in Appendix B1 we compute the expected evidence ratio $\langle \ln R\rangle$ and its variance $\sigma^2(R) = \langle(\ln R - \langle \ln R\rangle)^2\rangle$ for the case of data described by a Gaussian likelihood under the assumption of a linear model and flat priors, and define the *calibrated evidence ratio* $\ln R - \langle \ln R\rangle$. The latter allows a more quantitative measurement of tension than the somewhat heuristic Jeffreys' scale, and avoids biasing the results. For other details on our treatment of the evidence ratio see Appendix B.

2.3 Model selection

To determine whether a given data set, D, prefers a model M_1 or model M_2, we rely on the *deviance information criterion* (hereafter DIC). Considering the generalized chi-squared $\chi^2(\boldsymbol{\theta}) = -2\ln L(D|\,\boldsymbol{\theta}, M_i)$, the mean goodness of fit over the posterior volume can be estimated as $\langle\chi^2\rangle = -2\langle \ln L(D|\,\boldsymbol{\theta}, M_i)\rangle$. A model which fits the data better will have a lower $\langle\chi^2\rangle$. Motivated by information theory, Spiegelhalter et al. (2002) define the DIC as

$$\mathrm{DIC}(M_i) = \langle\chi^2\rangle + p_D. \quad (5)$$

This balances the mean goodness of fit $\langle\chi^2\rangle$ with the Bayesian complexity p_D, which measures the effective complexity of the model and is given by

$$p_D = \langle\chi^2\rangle - \chi^2(\tilde{\boldsymbol{\theta}}), \quad (6)$$

Table 1. Interpretation of the difference in deviance information criterion, ΔDIC, using the Jeffreys' scale as proposed by Spiegelhalter et al. (2002). For nested models described by uncorrelated Gaussian likelihoods, ΔDIC can be straightforwardly related to the deviation of the additional parameter in the more complex model w.r.t. its fixed value in the simpler one. As a reference, for a given ΔDIC we calculate the offset of an additional parameter measured in standard deviations σ, and in the corresponding p-value.

ΔDIC	Preference	σ	p-value
$(-2, 0)$	insignificant	(1.41, 2.00)	(2.28e − 2, 7.93e − 2)
$(-5, -2)$	positive	(2.00, 2.65)	(4.02e − 3, 2.28e − 2)
$(-10, -5)$	strong	(2.65, 3.46)	(2.70e − 4, 4.02e − 3)
$(-\infty, -10)$	decisive	(3.46, ∞)	(0, 2.70e − 4)

where $\tilde{\boldsymbol{\theta}}$ denotes the maximum likelihood point. A lower DIC means either that the model fits the data better (lower $\langle\chi^2\rangle$) or that it has a lower level of complexity, p_D. A higher complexity, such as additional model parameters, can only be compensated if they allow a sufficient improvement of the goodness of fit.

For model selection, the difference $\Delta\mathrm{DIC} = \mathrm{DIC}(M_2) - \mathrm{DIC}(M_1)$ is interpreted using the Jeffreys' scale (see Table 1). $\Delta\mathrm{DIC} = 0$ means that the data provide no preference for one model over the other, $-2 < \Delta\mathrm{DIC} < 0$ that there is 'no significant' preference for M_2, $-5 < \Delta\mathrm{DIC} < -2$ a 'positive' preference, $-10 < \Delta\mathrm{DIC} < -5$ 'strong', and $-10 < \Delta\mathrm{DIC}$ 'decisive'. The same values but positive indicate a preference for M_1 instead.

As an example of model selection with the DIC, consider the following case. Let the data be described by a standardized Gaussian likelihood $-2\ln L = \sum_{i=1}^{n}\theta_i^2$, where θ_1, ..., θ_n are the model parameters of M_2, and let the simpler model M_1 derive from M_2 by setting one of the model parameters θ_j to the value σ. In this case, assuming flat priors, the ΔDIC can be calculated analytically as $2 - \sigma^2$ and the p-value of the offset σ can be computed from the fact that the posterior of θ_j given the data D, $p(\theta_j|D, M_2)$, is Gaussian. For reference, we present these results in Table 1.

As discussed by Spiegelhalter et al. (2002), the DIC can also deal with strong parameter degeneracies, such as the geometrical degeneracy of the CMB data in curved models. It takes also into account 'parameter volume effects', as it considers the goodness of fit averaged over the posterior volume. Furthermore, this measure can be easily computed from a posterior sample, which saves the values $\ln L(D|\,\boldsymbol{\theta}, M_i)$ in every point, making it more versatile than the evidence ratio (for applications of this measure to astrophysics and cosmology, see Porciani & Norberg 2006; Liddle 2007; Mantz et al. 2010; Joudaki et al. 2016).

3 COSMOLOGICAL DATA

3.1 *Planck* data

We employ the `TT_lowTEB` constraints from the Planck Collaboration XIII (2015b) of the temperature and large scale polarization anisotropies in the CMB, which we also refer to as 'base P15 CMB'. When considering the full Planck 2015 temperature and polarization measurements, we use the `TTTEEE_lowTEB` sample, which we will also refer to as 'full P15 CMB'. We also use the CMB lensing constraints (Planck Collaboration XV 2015d) included in the `TT_lowTEB+lensing` samples, referring to them as 'CMB lens'. The Monte Carlo Markov Chain (MCMC) CMB samples analysed in this work were downloaded from

the Planck Legacy Archive[1] and subsequently Gaussianized as described in Section 2.1.

3.2 Additional geometrical probes

Given an analytic expression for the base P15 CMB constraints derived from the Gaussianization process described in Section 2.1 and Appendix A, we can easily combine them with measurements from geometrical probes. This has the advantage that the prominent geometrical degeneracy of the CMB data in curved models can be broken (see e.g. Bond, Efstathiou & Tegmark 1997; Zaldarriaga, Spergel & Seljak 1997). We compute the theoretical distance predictions using CAMB (Lewis, Challinor & Lasenby 2008)[2] and sample the joint constraints with the parallelized MCMC engine COSMOHAMMER (Akeret et al. 2012).[3] In the following we present the additional geometrical data sets we used in this work.

3.2.1 Data sets

Various recent constraints on the Hubble constant H_0 exist in the literature (Riess et al. 2011, 2016; Bennett et al. 2014; Efstathiou 2014; Aubourg et al. 2015). In the present work, we use the latest result by Riess et al. (2016, hereafter R16), who obtain $H_0 = 73.02 \pm 1.79\,\mathrm{km\,s^{-1}\,Mpc^{-1}}$. As a consistency check, we also use the constraint $H_0^{\mathrm{E14}} = 70.6 \pm 3.3\,\mathrm{km\,s^{-1}\,Mpc^{-1}}$ reported by Efstathiou (2014, hereafter E14). We use these measurements as Gaussian likelihoods. This simple form will also allow us to employ them to compute evidences as described in Appendices B1 and B2.

We also use measurements of the Hubble parameter as a function of redshift from the latest calibration of a large compilation of supernovae type Ia (SNe) data by Betoule et al. (2013). This work combines observations from the Supernovae Legacy Survey, the Sloan Digital Sky Survey (SDSS) and the *Hubble Space Telescope*, and provides a binned version of the SNe Hubble diagram with the corresponding covariance matrix. As shown in appendix E of Betoule et al. (2013), computing the luminosity distance in Mpc h^{-1}, marginalising analytically over the intrinsic luminosity of the SNe and assuming a Gaussian likelihood allows a straightforward computation of the SNe constraints.

We also include constraints from baryon acoustic oscillations (BAO) derived from galaxy correlations in the 6dF Galaxy Survey by Beutler et al. (2011), the SDSS main galaxy sample by Ross et al. (2015), and the Baryon Oscillation Spectroscopic Survey (BOSS) by Anderson et al. (2014). The Planck Collaboration XIII (2015b, see e.g. p. 24) provided samples of these BAO measurements together with the base CMB data, labelled as `TT_lowTEB+BAO`.

Delubac et al. (2015) derived BAO measurements from the Ly α forest in the Data Release 11 of BOSS. We will refer to this measurement as 'Ly α BAO'. These results are reported as $D_A(z = 2.34) = 1662 \pm 96\,\mathrm{Mpc}\,(r_\mathrm{d}/r_\mathrm{fid})$ and $H(z = 2.34) = 222 \pm 7\,\mathrm{km\,s^{-1}\,Mpc^{-1}}\,(r_\mathrm{fid}/r_\mathrm{d})$, where D_A is the angular diameter distance, $H(z)$ the expansion rate at a given redshift z, $r_\mathrm{fid} = 147.4\,\mathrm{Mpc}$ the fiducial sound horizon used by Delubac et al. (2015) and r_d the sound horizon dependent on the cosmological parameters. We assume Gaussian likelihoods for these results.

[1] http://pla.esac.esa.int/pla/#cosmology
[2] http://camb.info/
[3] https://github.com/cosmo-ethz/CosmoHammer

4 RESULTS

4.1 Which model is preferred by the P15 CMB data?

We compute the change in the deviance information criterion ΔDIC between the standard flat ΛCDM and several extended models. We consider all the one-parameter extensions for which the Planck Collaboration published `TT_lowTEB` constraints, namely: Ω_K (we refer to this model as *curved* ΛCDM); A_L, the amplitude of the CMB lensing potential (we refer to this model as flat ΛCDM+A_L); m_ν, the effective sum of neutrino masses; $\mathrm{d}n/\mathrm{d}\ln k$, the running of the spectral index of scalar perturbations; N_eff, the effective number of relativistic degrees of freedom; $r_{0.02}$, the tensor to scalar mode ratio; w, the dark energy equation of state parameter; and Y_He, the primordial Helium fraction.

In Fig. 1 and Table 2, we show the differences between the DIC of flat ΛCDM and those of the extended models as calculated from the publicly available samples. We find that the P15 CMB data favour most the curved ΛCDM model ($\Delta\mathrm{DIC} = 6.02$), followed by the model with free A_L ($\Delta\mathrm{DIC} = 4.12$). For the other model extensions we find no significant preference over flat ΛCDM. We also find that flat ΛCDM is preferred over a model with free tensor to scalar ratio, $r_{0.02}$.

The clear preferences for curved ΛCDM and flat ΛCDM+A_L are related to the fact that both Ω_K and A_L deviate more than 2σ from their assumed value in flat ΛCDM (see also discussion on p. 24 and 38 of Planck Collaboration XIII 2015b). For the case of

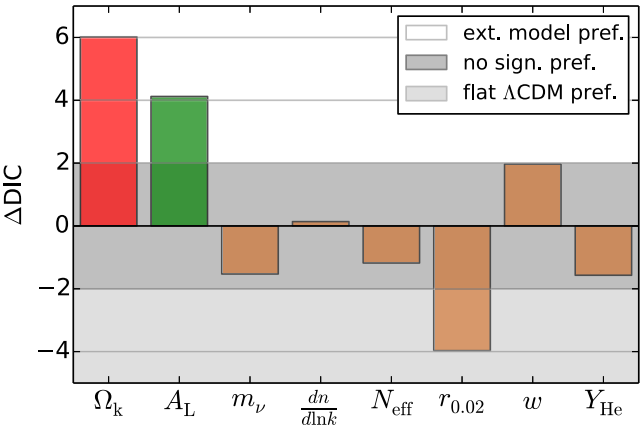

Figure 1. Differences in deviance information criterion, ΔDIC, between flat ΛCDM and various one-parameter extensions of this model. These results are estimated from the publicly available `TT_lowTEB` constraints. The ranges $-2 < \Delta\mathrm{DIC} < 2$, $\Delta\mathrm{DIC} > 2$ and $\Delta\mathrm{DIC} < -2$ indicate no significant preference for either model, a preference for the extended model, or that the data prefer the simpler model, respectively. Remarkably, we find clear preferences for two of the extended models, flat ΛCDM+A_L (green) and curved ΛCDM (red).

Table 2. ΔDIC between flat ΛCDM and various one-parameter extensions of this model, for the base P15 CMB constraints. The ranges $-2 < \Delta\mathrm{DIC} < 2$, $\Delta\mathrm{DIC} > 2$ and $\Delta\mathrm{DIC} < -2$ indicate no significant preference for either model, a preference for the extended model, or that the data prefer the simpler model, respectively.

Ω_K	A_L	m_ν	$\mathrm{d}n/\mathrm{d}\ln k$	N_eff	$r_{0.02}$	w	Y_He
6.02	4.12	−1.53	0.14	−1.18	−3.97	1.97	−1.57

curved ΛCDM, we find a preference for a closed Universe ($\Omega_K < 0$), with a p-value of

$$P(\Omega_K \geq 0) = \int_0^\infty p(\Omega_K|\,\texttt{TT\_lowTEB})\,\mathrm{d}\Omega_K = 0.0033, \quad (7)$$

corresponding to a 2.7σ significance. For the flat ΛCDM+A_L model, we find a preference for A_L larger than 1, with p-value

$$P(A_L \leq 1) = \int_{-\infty}^1 p(A_L|\,\texttt{TT\_lowTEB})\,\mathrm{d}A_L = 0.0098, \quad (8)$$

which corresponds to a 2.3σ deviation from the theoretically expected value $A_L = 1$.

Even though the P15 CMB likelihood is more complex than the one we use to calculate the reference results presented in Table 1, the significances of the offsets and the related p-values we obtain for the curved ΛCDM and flat ΛCDM+A_L models are consistent with the corresponding ΔDIC values in Table 2. We find no significant detection of curvature or $A_L > 1$, although our ΔDIC results indicate significant improvements of the fits w.r.t. the flat ΛCDM model, which according to the Jeffreys' scale dominate over the increased complexity of the curved and flat ΛCDM+A_L models. This is confirmed by Planck Collaboration XIII (2015b), which finds that the mean chi-squared $\langle\chi^2\rangle$ of the P15 CMB fit for these two models is lower than for flat ΛCDM. As discussed in detail in Planck Collaboration XIII (2015b, p. 24 and 38), these two models are sensitive to the large angular scale part of the TT P15 CMB spectrum and the power of CMB lensing potential $C_\ell^{\phi\phi}$, as we show in Section 5.4.

4.2 Quickly resampling the *Planck* constraints

The Gaussianization procedure effectively provides an analytic approximation to the P15 CMB likelihood. As we only Gaussianize the constraint on the cosmological parameters, we reconstruct the P15 CMB likelihood marginalized over the nuisance parameters. This is especially useful when using the P15 CMB constraints as priors to be combined with other probes, because it avoids the resampling of the P15 nuisance parameters, significantly reducing the number of parameters involved in this calculation. For example, for flat ΛCDM, the TT+lowTEB likelihood depends on 21 parameters, whereas only 5 are the cosmological parameters we resample. These cosmological parameters are H_0, the present-day physical baryon and CDM densities in units of the critical density, $\Omega_b h^2$ and $\Omega_{cdm} h^2$, where $h = H_0/100$, and the amplitude and spectral index of the primordial scalar fluctuations, $\ln(10^{10} A_s)$ and n_s.[4] The remaining 16 parameters include the optical depth to reionization, τ, and 15 nuisance parameters.

Furthermore, a single call to the analytic likelihood approximation takes less than a milli-second, compared to several seconds for the original *Planck* likelihood. This opens the possibility to quickly resample the P15 CMB constraints and to efficiently combine them with other probes. For further details, see Appendix A. The likelihoods are available at the following URL: https://bitbucket.org/grandiss45/gaussianization/.

The Gaussianization of the samples is not only helpful to approximate and quickly resample the P15 CMB constraints. It is crucial to computing the surprise analytically. This is possible because the relative entropy is invariant under parameter transformation and is analytic for Gaussian constraints. This allows us to compute the expected relative entropy $\langle KL\rangle_{D_2|D_1}$ and a mean fluctuation around this value $\sigma(KL)$ analytically. As these quantities are obtained by averaging over the distribution of data $E(D_2|D_1)$, it would be very difficult to compute them numerically. The same holds true for the calibration of the evidence ratio $\langle\ln R\rangle$. These integrals over the data are analytic if the constraints can be assumed to be Gaussian, as shown explicitly in Appendix B1.

[4] For simplicity, when combining with other data sets, we consider H_0 instead of θ_{MC}, the ratio of the approximate sound horizon to the angular diameter distance at recombination. The impact of this choice is discussed in Section 5.2.

Table 3. Surprise values S, expected fluctuations σ, and significances of tensions S/σ for different data sets added to the P15 TT_lowTEB constraints in the models we considered.

	BAO	CMB len.	TEEE	H_0	SNe	Ly α BAO
Flat ΛCDM						
S	−0.44	0.45	−1.13	1.11	−0.10	0.05
σ	0.68	0.72	0.97	0.23	0.15	0.05
S/σ	−0.65	0.63	−1.16	4.78	−0.67	1.04
Curved ΛCDM						
S	6.33	5.45	−1.19	7.36	2.85	0.77
σ	1.37	1.30	1.09	0.94	0.74	0.44
S/σ	4.63	4.18	−1.10	7.87	3.83	1.76
Flat ΛCDM+A_L						
S	−0.31	3.73	−0.92	0.57	−0.10	0.07
σ	0.77	0.89	1.11	0.37	0.40	0.16
S/σ	−0.40	4.21	−0.83	1.52	−0.25	0.46

4.3 Adding external data to the *Planck* CMB

Here we test the consistency between each of the data sets described in Section 3 and the base P15 CMB constraints, first for the standard flat ΛCDM model and then for the two models that we found in Section 4.1 to be favoured by the base P15 CMB data, i.e. curved ΛCDM and flat ΛCDM+A_L. For the former case, we use the standard set of cosmological parameters listed in Section 4.2, while marginalizing over the other parameters sampled by P15 as they are unconstrained by the additional data. In the curved model we also consider the constraints on Ω_K, whereas in the flat ΛCDM+A_L model we add the parameter A_L.

4.3.1 *Flat ΛCDM*

In flat ΛCDM, the base P15 CMB constraints are very well approximated by a multivariate Gaussian distribution, so no Gaussianization is required for resampling. We approximate the constraints directly as multivariate Gaussians, update them with constraints from external data, and then compute the surprise. We summarize our results in Table 3 and show them in Fig. 2 (blue bars). We find that for flat ΛCDM all external data sets are consistent with the base P15 CMB measurements. However, the H_0 measurement of R16 is in almost 5σ tension with the base P15 CMB data set.[5] Worth mentioning is also the tendency to negative surprises for the BAO and SNe

[5] R16 report that the distance between their mean H_0 value and the mean value obtained from the P15 analysis is 3σ, where $\sigma^2 = \sigma^2_{R16} + \sigma^2_{P15}$ and $\sigma_{P15,\,R16}$ are the measurement uncertainties on H_0 of the two experiments. This result is not in contradiction with our claim, as we instead compute the significance of such a shift. We find that this 3σ shift is significant at almost a 5σ level. This is also confirmed by our calibrated evidence ratio calculation below.

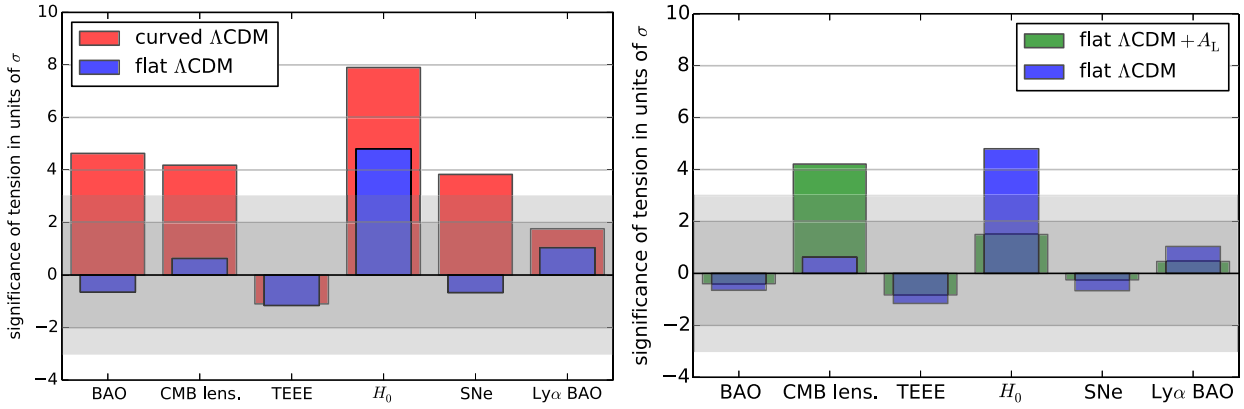

Figure 2. Significances of the surprises in units of σ in the flat ΛCDM (blue), curved ΛCDM (red, left-hand panel), and flat ΛCDM+A_L (green, right-hand panel) models when combining the base P15 CMB constraints with six other probes. The grey regions show the 2σ and 3σ regions. Surprises more significant than 3σ (above the grey regions) indicate tensions of the additional data with the CMB prior. We see that in flat ΛCDM all probes are consistent with the base P15 CMB constraints, except for the distance ladder measurements. In curved ΛCDM, BAO, CMB lensing, H_0 and SNe are in significant tension with the base P15 CMB constraints. In flat ΛCDM+A_L, CMB lensing is in significant tension with the base P15 CMB constraints, whereas the other probes are in agreement.

data and most strongly for the 'TEEE' polarization data. Negative surprises mean that the additional data agree with the prior more than statistically expected. However, these negative surprises are not significant, and can thus be interpreted as statistical fluctuations.

4.3.2 *Curved ΛCDM*

Fig. 3 shows joint constraints in the space of H_0 and Ω_K for the base P15 CMB data set alone (red contours) and in separate combinations with six additional data sets (blue contours). As is clear in this figure and as already presented in equation (7), the base P15 CMB data favour a model with negative curvature at the 2.7σ confidence level. In itself, this is not a detection of curvature. Hence, to improve the constraints, additional data sets can be added. Fig. 3 shows the impact of such combinations and illustrates how the addition of CMB lensing, two flavours of BAO, SNe and H_0 measurements push the P15 CMB constraints noticeably back towards flatness.

By jointly Gaussianizing the prior (the base P15 CMB constraints) and the posterior (combined constraints) for each data set we add, we transform the cosmological parameters into a space where both distributions are well described by Gaussian distributions. In this space, we estimate the surprise values given in Table 3 and shown in Fig. 2 (red bars in the left-hand panel). For an discussion of the accuracy of this method, see Appendix A. As anticipated by the large shifts in the marginalized plane of H_0 and Ω_K, most additional probes are in significant tension with the base P15 CMB constraints: H_0 data at the 8σ level, BAO and CMB lensing data just over 4σ, and SNe slightly less than 4σ. Ly α BAO data also shift the CMB constraints, but this shift is only a bit less than 2σ significant. Finally, also in this model, the TEEE spectrum of the P15 polarization measurements agree with the base P15 constraints more than statistically expected, although not in a statistically significant manner.

The large surprises in four of the external data sets when combined with the P15 CMB constraints is evidence of significant tensions among the data sets. Thus, our analysis emphasizes that while the combined constraints (P15 CMB + external data set) prefer flatness more than the P15 CMB data set alone, this comes at the cost of combining data sets that in four cases are significantly in tension with one another.

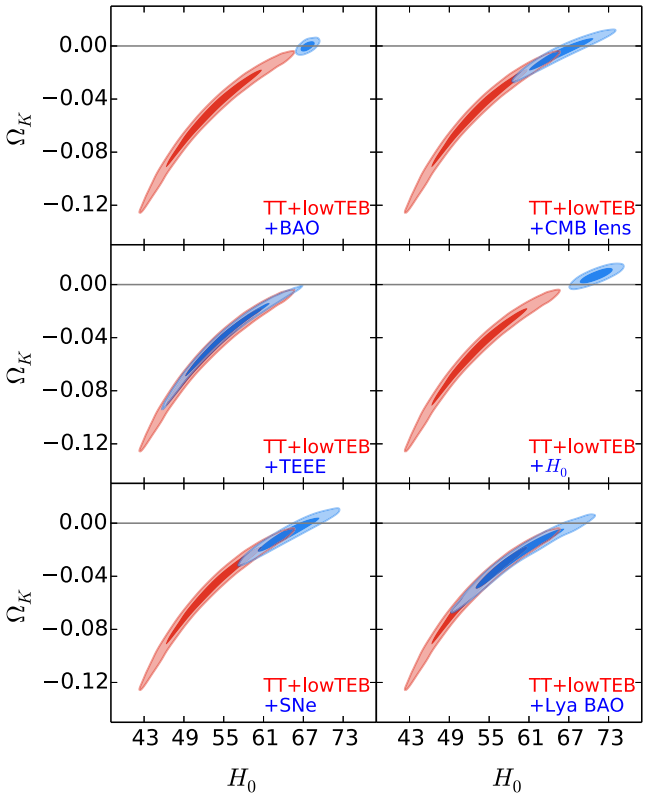

Figure 3. Marginal constraints on H_0 and Ω_K from the base P15 CMB data set (red contours) and the addition of different data sets to the latter (blue). Adding the P15 small-scale polarization data (TEEE) results produces no significant shift of the constraints. However, all external data sets shift the constraints back to flatness, at the cost of increasing tension with the base CMB measurements.

4.3.3 *Flat ΛCDM+*A_L

Considering the highly significant tensions we find in the curved ΛCDM model, we also investigate the consistency of the different data sets with the base P15 CMB data in the flat ΛCDM+A_L model. We show our results in Fig. 2 (green bars in the right-hand panel)

Table 4. Evidence ratio results for some of the data sets. $\ln \hat{R}$ denote the numerical and $\ln R$ the analytic estimates respectively. $\langle \ln R \rangle$ is the calibration of the evidence ratio and $\ln R - \langle \ln R \rangle$ the calibrated evidence ratio. 'Sig' stands for the significance $(\ln R - \langle \ln R \rangle)/\sigma(\ln R)$, where in one dimension $\sigma(\ln R) = 1/\sqrt{2}$. Note that contrary to the surprise values, in the case of evidence ratios negative values indicate tension and positive values indicate agreement.

Flat	$\ln \hat{R}$	$\ln R$	$\langle \ln R \rangle$	$\ln R - \langle \ln R \rangle$	Sig
H_0 R16	-5.6 ± 1.9	-5.59	-2.13	-3.46	-4.89
H_0 E14	-2.6 ± 0.3	-2.61	-2.65	0.04	0.06
SNe, flat	2.2 ± 0.2	2.28	1.89	0.39	0.54
Curved	$\ln \hat{R}$	$\ln R$	$\langle \ln R \rangle$	$\ln R - \langle \ln R \rangle$	Sig
H_0 R16	-9.2 ± 3.7	-9.39	-3.01	-6.30	-8.90
H_0 E14	-6.6 ± 1.5	-6.85	-3.22	-3.63	-5.13

and Table 3. Contrary to the curved ΛCDM model, we find that all distance measures are in good agreement with the base P15 CMB constraints. In the case of the H_0 measurement, we find that the significance of tension is reduced from 4.8σ in flat ΛCDM to 1.5σ in the flat ΛCDM+A_L model. This is to some extent unsurprising, as these data sets do not directly constrain the additional parameter A_L. But it is worth noting that leaving the A_L parameter free in the CMB fit, does not change the constraints on the other parameters in a way that is inconsistent with the various distance measure data sets. Actually, it allows for higher values of H_0, reducing the tension with the distance ladder measurements.

However, CMB lensing measurements are sensitive to the lensing of the CMB by construction. This data set shows a tension of 4σ with the base P15 CMB data. This tension is driven by the constraints on the lensing amplitude. As shown by the Planck Collaboration XIII (2015b, p. 24) the constraints from the base CMB ($A_L = 1.22 \pm 0.10$) are shifted strongly when the CMB lensing data are added ($A_L = 1.04 \pm 0.06$). The latter is an indication that two data sets which are inconsistent with each other have been combined. We will discuss the underlying physical description of these constraints in Section 5.4.

4.4 Another independent measurement of tension

As a consistency check for our results, we also employ evidence ratios. We compute the evidence ratios analytically (see Appendix B1) for those data sets and models where the likelihood of the data could be assumed to be a simple Gaussian. We use special care in calibrating the analytic evidence ratio $\ln R - \langle \ln R \rangle$, as discussed in Appendix B1. We also validate our analytic computations with numerical estimates, $\ln \hat{R}$ (see Appendix B2), which allow us to relax the assumption of Gaussianity for the base P15 CMB likelihood. We summarize our findings in Table 4.

We find that the numerical evidence $\ln \hat{R}$ and the analytic evidence $\ln R$ agree. Importantly, in most of the cases we find that the expected evidence ratio $\langle \ln R \rangle$ is very different from zero. Not accounting for the correct calibration can therefore lead to a serious mis-estimation of the degree of tension, as can be seen in the case of H_0 E14 and SNe for flat ΛCDM. Both agree with the base P15 CMB data, as seen with both the surprise S and the calibrated evidence ratio $\ln R - \langle \ln R \rangle$. However, just considering the evidence ratio $\ln R$ would have biased our conclusion, leading to an overestimation of the agreement in the case of SNe and an underestimation of the agreement in the case of H_0 E14. We conclude from this simple example, that uncalibrated evidence ratios can be significantly biased, as discussed further in Appendix B1.

Considering the calibrated evidence ratios $\ln R - \langle \ln R \rangle$ in Table 4 we detect the same tensions as with the surprise (see Tables 3 and 6). Furthermore, the calibrated evidence ratio, which scatters with $\sigma(\ln R) = 1/\sqrt{2}$, have significances comparable to the significances of the surprise. We conclude that in these examples the two measures of tension give very similar results, despite the fact that they detect tensions in different ways, as discussed in Appendix B1. This is reassuring for our primary results with the surprise estimated after a Gaussianization process, and for the validity of the calibrated evidence ratio, introduced here for the first time.

5 DISCUSSION

In this section we consider three possible origins for the significant tensions we detect between various data sets and the base P15 CMB constraints. First, we discuss the fact that data sets could be affected by systematic effects biasing their constraints; second, we explore the impact on the base P15 CMB constraints of using a flat prior on θ_{MC} instead of on H_0 for the curved ΛCDM model; and finally, we investigate the physical processes underlying the tensions measured in parameter space.

5.1 Impact of systematics

Each of the data sets we consider might be affected by residual systematic uncertainties large enough to lead to tensions with others. As shown elsewhere (Seehars et al. 2014, 2016), unresolved systematic uncertainties in the *Planck* half mission CMB data (Planck Collaboration I 2014) resulted in highly significant tensions with the CMB constraints from *WMAP* (Bennett et al. 2013; Hinshaw et al. 2013), whereas the base P15 CMB constraints are in a far better agreement with *WMAP*, which holds true also in a series of extended models (see Grandis et al. 2016).

To check whether any systematic effect accounted for by the Planck Collaboration might play a role in the 2.7σ deviation from flatness, and the 2.3σ deviation from $A_L = 1$, in the base P15 CMB data, we show in Fig. 4 the constraints on the nuisance parameters sampled by the Planck Collaboration with the largest variations between the flat ΛCDM model (blue lines) and either the curved ΛCDM (red) or the flat ΛCDM+A_L (green) models. We find no major shifts in the nuisance parameter constraints. Thus, treatment of systematic effects in the base P15 CMB data appears stable under these extensions and not responsible for the tensions reported here. However, this does not exclude the possibility that there are unresolved residual systematics in the P15 data.

Interestingly, the minor shifts induced by the curved ΛCDM and flat ΛCDM+A_L models are very similar. This hints at a similarity in the way these two models impact the P15 CMB constraints, as discussed in detail in Planck Collaboration I (2014) and Planck Collaboration XIII (2015b, p. 29).

The resulting tensions could also come from the other probes. We discuss the impact of different H_0 measurements in Section 5.3. For exhaustive discussions of the treatment of systematics in the data sets employed here, we refer the reader to the literature referenced in Section 3.

5.2 Effect of a prior choice

Another effect which could contribute to the preference for non-flat models is the weight assigned to different regions of parameter space by the priors used to sample the base P15 CMB constraints in the curved ΛCDM model. The Planck Collaboration assumed flat

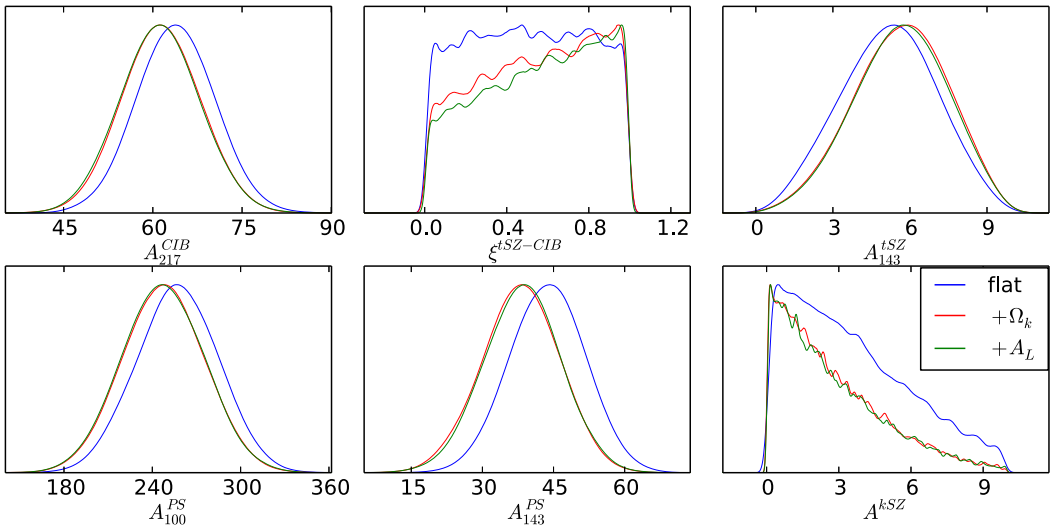

Figure 4. Constraints on nuisance parameters of the base P15 CMB data in flat ΛCDM, curved ΛCDM, and flat ΛCDM+$A_{\rm L}$. For simplicity, we include only the marginalized constraints on the parameters that display the largest changes with respect to the flat ΛCDM model. However, no major shifts are present in these nuisance parameters. Interestingly, both extensions, free $\Omega_{\rm K}$ and $A_{\rm L}$, shift these constraints in a very similar manner.

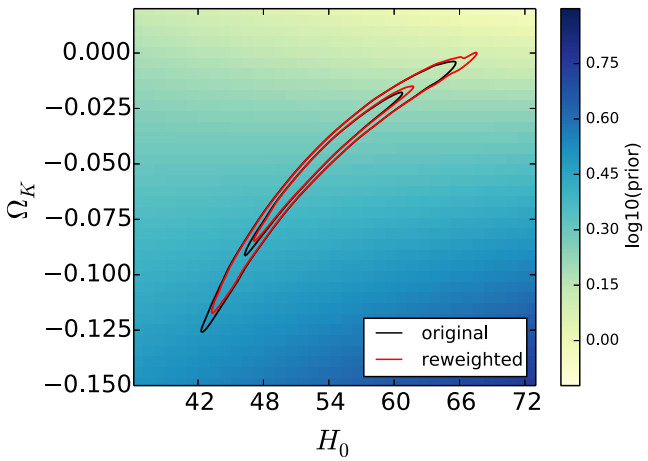

Figure 5. In black, the marginalized contours of the base P15 CMB constraints in curved ΛCDM over plotting a colour-coded background of the $\log_{10}$ of the prior weights, derived from flat priors on both $\Omega_{\rm K}$ and $\theta_{\rm MC}$. Clearly, the prior puts more weight (up to $10^{0.5} \sim 3$) on the low H_0, negative $\Omega_{\rm K}$ tail of the degeneracy. The red contours are obtained by crudely reweighting the sample (see equation 9), to make it correspond to flat priors on $\Omega_{\rm K}$ and H_0 instead.

priors on $\Omega_{\rm K}$ and $\theta_{\rm MC}$. In Fig. 5 we show the marginalized contours of the base P15 CMB constraints on the H_0, $\Omega_{\rm K}$ plane. To crudely estimate the weight of the prior, we fix the other cosmological parameters to their best-fitting values and compute $\theta_{\rm MC}$ on a grid as a function of H_0 and $\Omega_{\rm K}$ using CAMB. We then numerically compute the flat prior

$$p(H_0,\ \Omega_{\rm K}) = \left|\frac{\partial(\theta_{\rm MC},\ \Omega_{\rm K})}{\partial(H_0,\ \Omega_{\rm K})}\right| p(\theta_{\rm MC},\ \Omega_{\rm K}) \propto \left|\frac{\partial\theta_{\rm MC}}{\partial H_0}\right|, \qquad (9)$$

where $p(\theta_{\rm MC}, \Omega_{\rm K})$ is the prior on $\Omega_{\rm K}$ and $\theta_{\rm MC}$, which can be assumed $\propto 1$, and $|\partial(\theta_{\rm MC},\ \Omega_{\rm K})/\partial(H_0,\ \Omega_{\rm K})|$ stands for the determinant of the Jacobian of the transformation $(\theta_{\rm MC}, \Omega_{\rm K}) \mapsto (H_0, \Omega_{\rm K})$, which can be simplified to $|\partial\theta_{\rm MC}/\partial H_0|$, the absolute value of the partial derivative

Table 5. Surprise values S and expected fluctuation σ for different data sets added to the P15 `TT_lowTEB` constraints in curved ΛCDM after accounting for the reweighing due to the change between using a flat prior on H_0 instead of on $\theta_{\rm MC}$.

	BAO	CMB len.	TEEE	H_0	SNe	Ly α BAO
S	5.34	4.93	−1.23	6.57	2.49	0.73
σ	1.33	1.29	1.08	0.94	0.74	0.45
S/σ	4.02	3.82	−1.14	6.98	3.36	1.62

of $\theta_{\rm MC}$ with respect to H_0, evaluated at the relevant position in parameter space.

We find that the original priors give more weight to regions away from $\Omega_{\rm K} = 0$, with up to a factor of ∼3 at the low end of the degeneracy, as shown in Fig. 5. We also show there the marginalized contours of the original (in black) and the reweighted (in red) sample obtained from the former using equation (9). As an effect of the reweighting, the deviation from flatness is reduced from 2.7σ to 2.5σ. We also calculate numerically the impact of the reweighting on the ΔDIC, finding that it is insignificant and that the clear preference for curved ΛCDM is maintained.

In Table 5, we show the entropy results after reweighting. The significances of the tensions are slightly lower than before reweighting. This comes from the fact that the reweighting pushes the base CMB constraints towards flatness and therefore to better agreement with the other data sets. Nevertheless, as before, with the exception of Ly α BAO, all additional probes maintain more than 3σ tension. Thus, we conclude that this change in the prior does not resolve the tensions we find in curved ΛCDM because it reduces the significances of the tensions and deviations only by ∼10 per cent. However, it is worth noting that any choice of prior (even flat) in parameter space can indeed introduce unintended preferences for certain regions in this space.

5.3 CMB and distance ladder

The consistency between the Hubble rate inferred from the CMB and distance ladder measurements is a popular and important topic

Table 6. Surprises S and expected fluctuation σ for different H_0 measurements when added to the P15 TT_lowTEB constraints in flat, curved and flat ΛCDM+$A_{\rm L}$.

	Flat ΛCDM		Curved ΛCDM		Flat ΛCDM+$A_{\rm L}$	
	R16	E14	R16	E14	R16	E14
S	1.11	0.01	7.36	3.74	0.57	−0.17
σ	0.23	0.08	0.94	0.75	0.37	0.17
S/σ	4.78	0.09	7.87	4.97	1.52	−1.00

in the recent literature (see for instance Verde et al. 2013, 2014; Bennett et al. 2014; Riess et al. 2016). Since Planck Collaboration XIII (2015b) adopted $H_0^{\rm E14} = 70.6 \pm 3.3\,{\rm km\,s^{-1}\,Mpc^{-1}}$ (E14), we repeat our analysis with this measurement and obtain the results shown in Table 6 for the surprise. E14 agrees better with the base P15 CMB constraints than R16 in all models we considered. For flat ΛCDM, E14 is consistent with the CMB constraints, as also found by Planck Collaboration XIII (2015b). Compared to previous results from Riess et al. (2011) and Riess (2014), the tighter measurements on H_0 from R16 are however in significant tension with the P15 CMB constraints even for this simple model.

Interestingly, when we consider the curved ΛCDM model, all distance ladder measurements show significant tensions with the base P15 CMB constraints. The presence of the tension between the P15 CMB constraints and the distance ladder measurements in the curved ΛCDM model is thus independent of the specific H_0 measurement we choose, although its significance varies (4.97 for E14 and 7.87 for R16).

To reconcile the constraints on H_0 from the CMB and the local distance measures, a variety of mechanisms have been proposed, including an increased $N_{\rm eff}$ (Archidiacono et al. 2013; Di Valentino et al. 2016b), phantom dark energy (Planck Collaboration I 2014; Di Valentino, Melchiorri & Silk 2016a), or interacting dark energy (Salvatelli et al. 2013; Costa et al. 2014). It is worth stressing that here we find consistency between the H_0 measurements from both E14 and R16 with the base P15 CMB data in the flat ΛCDM+$A_{\rm L}$ model. Thus, contrary to models with free $N_{\rm eff}$ or $w \neq -1$, a model with $A_{\rm L} > 1$ is not only preferred by the CMB data alone but also provides consistency between these data and all the local distance measures (see additional discussion in Section 4.1).

5.4 Physical effects involved in the tensions

To investigate the physical effect causing the deviation from flatness and $A_{\rm L} = 1$ in the base P15 CMB constraints, we compare the theoretical predictions of the TT spectrum in flat, curved, and flat ΛCDM+$A_{\rm L}$ models. To do so, we draw random points from the base P15 CMB samples in these models and compute the theoretical expectation of the angular power spectrum of the temperature anisotropies, C_ℓ, using CAMB. In Fig. 6 we show the fractional differences with respect to the best-fitting values of C_ℓ in flat ΛCDM. We find that the 1σ uncertainty on the flat ΛCDM prediction (blue region) ranges from 4 per cent at low ℓ to less than 1 per cent at high ℓ, underlining the impressive constraining power of the P15 CMB measurements. For the distribution of the C_ℓ in the curved ΛCDM model (in red), we find that above $\ell \sim 50$ the TT spectra predicted by both models are consistent with each other at the 1σ level and within a 2 per cent fractional difference. However, at low $\ell < 30$ the curved model is able to predict noticeably less power than the flat model. For the lowest ℓ, the dipole term, the preferred curved model predicts almost 20 per cent less power than the flat model. As discussed elsewhere (Planck Collaboration XVI 2015f; Schwarz et al. 2015, and references therein), the lack of power on large scales is one of the anomalies observed in all CMB surveys, P15 included. The 2.7σ deviation from flatness seems to be driven by these anomalies and due to the ability of the curved model to predict less power on large scales. Similarly, also the C_ℓ's predicted in the model with free $A_{\rm L}$ are in excellent agreement with the flat ΛCDM prediction above $\ell \sim 30$. But also in this model, we find a lack of power on large angular scales, although in a less pronounced way than in the curved model. At low redshift, this can be achieved through the Integrated Sachs–Wolfe (ISW) effect (see Sachs & Wolfe 1967; Kofman & Starobinskij 1985; Planck Collaboration XIV 2015g).

However, as discussed by the Planck Collaboration XIII (2015b, p. 38), the constraints on curvature can also come from an increase of the lensing potential, which directly manifests itself as a deviation of its amplitude $A_{\rm L} > 1$ (see Section 4.1, fig. 1 and p. 24 in Planck Collaboration XIII 2015b). To investigate this possibility in

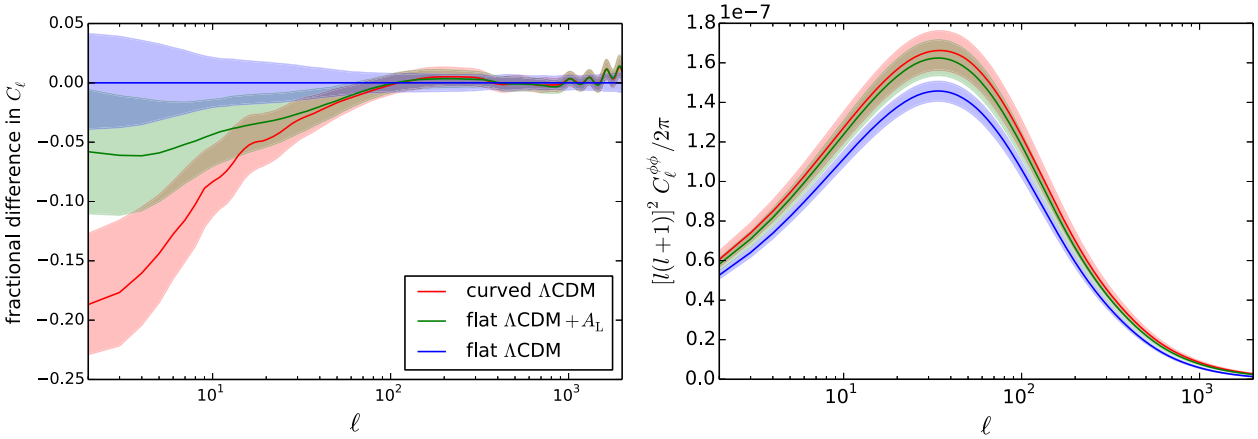

Figure 6. Left-hand panel: fractional differences between the flat ΛCDM best-fitting value of the TT power spectrum and those predicted by the constraints obtained in flat ΛCDM (blue), curved ΛCDM (red), and flat ΛCDM+$A_{\rm L}$ (green). For multipole moments $\ell < 30$, the P15 temperature anisotropy measurements prefer less power than that predicted by flat ΛCDM. This lack of power is stronger in the curved model than in the model with free $A_{\rm L}$. Right-hand panel: CMB lensing power spectrum predictions from the base CMB constraints obtained in flat ΛCDM (blue), curved ΛCDM (red), and flat ΛCDM+$A_{\rm L}$ (green). Remarkably, the curved and the flat ΛCDM+$A_{\rm L}$ models predict very similar lensing power spectra, both larger than the prediction from flat ΛCDM.

further detail, we compute the CMB lensing potential power spectrum $C_\ell^{\phi\phi}$ predicted by the base CMB constraints in flat ΛCDM, curved ΛCDM and flat ΛCDM+$A_{\rm L}$. The results, together with the 1σ uncertainties, are shown in Fig. 6, where we show the predictions of the CMB lensing power spectra for the flat (blue), curved (red) and flat ΛCDM+$A_{\rm L}$ (green) models. Remarkably, the curved and the flat ΛCDM+$A_{\rm L}$ models predict very similar $C_\ell^{\phi\phi}$s, which are about 2σ larger than those predicted by flat ΛCDM. From this we conclude that both the deviation from flatness and the deviation from $A_{\rm L}$ might be sourced by the same anomaly in the CMB lensing potentials. This might also be supported by the fact that the constraints on the nuisance parameters sampled by P15 are very similar in these two models, as already noted in Section 5.1 (see Fig. 4).

Although the constraints on the CMB lensing potentials are very similar for the curved and the flat ΛCDM+$A_{\rm L}$ models and show similar trends in the predicted temperature power spectrum, this is not true for the predicted background evolutions. This manifests itself in our tests of the curved model, where different distance measurements are in significant tension with the CMB. We show that considering the flat ΛCDM+$A_{\rm L}$ model, the tensions between the base P15 CMB and H_0, SNe and BAO are considerably alleviated, both compared to flat and curved ΛCDM. Thus, the consistency of the CMB with distance measures in the flat ΛCDM+$A_{\rm L}$ model seems to suggest that a modification of the CMB lensing potential is preferred to deviations from flatness. However, such modifications to the CMB lensing potential should not only fit the CMB spectra better, they should also be consistent with the CMB lensing measurements, which we find to be in tension with the base CMB data both in the curved and in the flat ΛCDM+$A_{\rm L}$ models.

As shown in Acquaviva & Baccigalupi (2006, see also e.g. Carbone et al. 2013), $A_{\rm L} > 1$ is naturally related to theories of modified gravity. Furthermore, the Planck Collaboration XIV (2015g) reported that the base P15 CMB constraints on some classes of modified gravity models deviate more than 2σ from General Relativity. Such models are found to fit the CMB data better than flat ΛCDM. It would be interesting to see whether such models can reconcile the CMB lensing measurements with the constraints from the base P15 CMB data.

6 CONCLUSIONS

In this work we first investigate which model is preferred by the CMB temperature and large scale polarization anisotropy measurements of the Planck Collaboration (base P15 CMB; Planck Collaboration XIII 2015b). Applying the *deviance information criterion* on the posterior samples made publicly available by the Planck Collaboration XIII (2015b), we find that the base P15 CMB constraints present a strong preference for a ΛCDM model with free curvature, $\Omega_{\rm K}$, over the flat ΛCDM paradigm. This strong preference comes from the fact that the curved model fits the CMB data at low multipoles ($\ell < 30$) better that the flat model, as reported by the Planck Collaboration XIII (2015b, p. 38). We also find that the constraints on $\Omega_{\rm K}$ deviate at a 2.7σ level from flatness ($\Omega_{\rm K} = 0$). Furthermore, we find that the base P15 CMB data prefer a model with a CMB lensing potential amplitude $A_{\rm L} \neq 1$. In this model, the constraints on the additional parameter $A_{\rm L}$ are found to deviate from the flat ΛCDM expectation ($A_{\rm L} = 1$) by 2.3σ. If this result is not due to residual systematics in the data, our model selection analysis (see Section 4.1) indicates that it represents a challenge to the standard flat ΛCDM model.

To investigate whether there is concordance between different measurements in these models, we consider the addition of external data sets to the base P15 CMB constraints. We utilize the joint constraints published by the Planck Collaboration XIII (2015b) from measurements of the base P15 CMB together with CMB lensing, CMB small-scale polarization, BAO, SNe, distance ladder or Ly α forest BAO. To analyse these data sets, we simultaneously *Gaussianize* the constraints from the base P15 CMB data and the combined data sets, and obtain an analytic approximation to their likelihood that enables the calculation of the entropy based measure *surprise* (Seehars et al. 2014, 2016; Grandis et al. 2016) and a *calibrated* evidence ratio, as well as a more efficient evaluation of the likelihood.

In the flat ΛCDM model, we find that all external data sets agree with the base P15 CMB, except for the distance ladder measurement performed by R16, which we find to be in 4.8σ tension. In the curved ΛCDM model, which is clearly preferred by the base P15 CMB data, we find significant tensions between the CMB and distance ladder (7.9σ), BAO (4.6σ), CMB lensing (4.2σ) and SNe (3.8σ) measurements. The curved model is thus unable to describe these observations adequately. Given these high levels of tension, these data sets should not currently be added to the base P15 CMB constraints in the curved model until these inconsistencies can be resolved. Considering instead a model with a free CMB lensing potential amplitude $A_{\rm L}$, the base P15 CMB constraints are consistent with the different distance measures, even resolving the tensions between the CMB and distance ladder measurements. However, in this model the CMB lensing measurements are still in about 4σ tension with the base P15 CMB data.

Using a simple example, we also show the importance of accurately calibrating the evidence ratio to have an unbiased assessment of the consistency between two data sets. To validate our primary measure of tension, we introduce the *calibrated evidence ratio* and calculate its expected fluctuation. Applying this measure to some of the data sets gives us significances of the tensions that are in good agreement with those from the surprise.

We also discuss the possible effects driving the deviation from flatness in the base P15 CMB constraints and therefore the tensions of these data with different external data sets. Our examination uncovers no evidence that these are due to systematics currently accounted for in the CMB analysis; however, we cannot exclude that these are due to unresolved, residual systematics. Also, the choice of using a flat prior on $\theta_{\rm MC}$ instead of H_0 for the CMB analysis introduces only a 10 per cent bias on the reported significances of the deviations and tensions, and is thus insufficient to explain them.

We also compute the TT spectra predicted by the base CMB constraints in the flat model and in the preferred models with free curvature and lensing amplitude. When comparing them to flat ΛCDM, we find a lack of power on large scales of almost 20 per cent for the curved, and 5 per cent for the $+A_{\rm L}$ model, respectively. Large scale lack of power has been consistently found in all CMB all-sky surveys, and might source the deviation we find here. This anomaly partially manifests itself as an increment of the CMB lensing potential. Remarkably, both the curved and the flat ΛCDM+$A_{\rm L}$ models predict larger CMB lensing potentials than the flat ΛCDM model. However, the curved model increases the lensing potentials at the cost of altering the cosmological background in a way that is incompatible with external distance measurements. On the other hand, a model that impacts the CMB lensing potentials without significantly changing the background expansion would allow consistency between the base P15 CMB data and external distance measurements. Such an alternative model should also be able to reconcile the direct CMB lensing measurements with the constraints coming from the temperature anisotropy power spectrum, which is

not the case with the flat ΛCDM+A_L model, as we have shown here. The important ongoing efforts in measuring the cosmic large scale structure in large survey projects such as, for example, DES[6] (DES Collaboration 2005), eROSITA[7] (Merloni et al. 2012), EUCLID[8] (Laureijs et al. 2011) and LSST [9] (LSST Science Collaboration 2009) will provide us with additional consistency checks among data sets while yielding tighter constraints that enable further systematic tests of alternative models to flat ΛCDM.

ACKNOWLEDGEMENTS

SG thanks Alexander Refregier, Adam Amara and Sebastian Seehars for fruitful discussions on various aspects of this work. We also thank the anonymous referee for useful comments. We acknowledge the support by the DFG Cluster of Excellence 'Origin and Structure of the Universe', the Transregio program TR33 'The Dark Universe' and the Ludwig-Maximilians-Universität. DR is currently supported by a NASA Postdoctoral Program Senior Fellowship at the NASA Ames Research Center, administered by the Universities Space Research Association under contract with NASA. We acknowledge use of the Planck Legacy Archive. *Planck* (http://www.esa.int/Planck) is an ESA science mission with instruments and contributions directly funded by ESA Member States, NASA, and Canada.

REFERENCES

Acquaviva V., Baccigalupi C., 2006, Phys. Rev. D, 74, 103510
Akeret J., Seehars S., Amara A., Refregier A., Csillaghy A., 2012, preprint (arXiv:1212.1721)
Allen S. W., Rapetti D. A., Schmidt R. W., Ebeling H., Morris R. G., Fabian A. C., 2008, MNRAS, 383, 879
Amendola L., Marra V., Quartin M., 2013, MNRAS, 430, 1867
Anderson L. et al., 2014, MNRAS, 441, 24
Archidiacono M., Giusarma E., Hannestad S., Mena O., 2013, preprint (arXiv:1307.0637)
Aubourg É. et al., 2015, Phys. Rev. D, 92, 123516
Battye R. A., Moss A., 2014, Phys. Rev. Lett., 112, 051303
Bennett C. L. et al., 2013, ApJS, 208, 20
Bennett C. L., Larson D., Weiland J. L., Hinshaw G., 2014, ApJ, 794, 135
Benson B. A. et al., 2013, ApJ, 763, 147
Betoule M. et al., 2013, A&A, 552, A124
Beutler F. et al., 2011, MNRAS, 416, 3017
Bocquet S. et al., 2015, ApJ, 799, 214
Bond J. R., Efstathiou G., Tegmark M., 1997, MNRAS, 291, L33
Carbone C., Baldi M., Pettorino V., Baccigalupi C., 2013, J. Cosmol. Astropart. Phys., 9, 004
Conley A. et al., 2011, ApJS, 192, 1
Costa A. A., Xu X.-D., Wang B., Ferreira E. G. M., Abdalla E., 2014, Phys. Rev. D, 89, 103531
Das S. et al., 2011, Phys. Rev. Lett., 107, 021301
de Haan T. et al., 2016, preprint (arXiv:1603.06522)
Delubac T. et al., 2015, A&A, 574, A59
DES Collaboration, 2005, preprint (arXiv:astro-ph/0510346)
DES Collaboration, 2015, preprint (arXiv:1507.05552)
Di Valentino E., Melchiorri A., Silk J., 2016a, preprint (arXiv:1606.00634)
Di Valentino E., Giusarma E., Mena O., Melchiorri A., Silk J., 2016b, Phys. Rev. D, 93, 083527
Efstathiou G., 2014, MNRAS, 440, 1138 (E14)
Grandis S., Seehars S., Refregier A., Amara A., Nicola A., 2016, J. Cosmol. Astropart. Phys., 5, 034
Hamann J., Hasenkamp J., 2013, J. Cosmol. Astropart. Phys., 10, 044
Hasselfield M. et al., 2013, J. Cosmol. Astropart. Phys., 7, 8
Heneka C., Marra V., Amendola L., 2014, MNRAS, 439, 1855
Hinshaw G. et al., 2013, ApJS, 208, 19
Jeffreys H., 1961, The Theory of Probability. Oxford Univ. Press, Oxford
Joudaki S. et al., 2016, preprint (arXiv:1601.05786)
Karpenka N. V., Feroz F., Hobson M. P., 2014, in Heavens A., Starck J.-L., Krone-Martins A., eds, Proc. IAU Symp. Vol. 306, Statistical Challenges in 21st Century Cosmology. Cambridge Univ. Press, Cambridge, p. 322
Kilbinger M. et al., 2013, MNRAS, 430, 2200
Kofman L., Starobinskij A. A., 1985, Pisma v Astron. Zh., 11, 643
Kullback S., Leibler R. A., 1951, Ann. Math. Stat., 22, 79
Laureijs R. et al., 2011, preprint (arXiv:1110.3193)
Lewis A., Challinor A., Lasenby A., 2008, ApJ, 538, 473
Liddle A. R., 2007, MNRAS, 377, L74
LSST Science Collaboration, 2009, preprint (arXiv:0912.0201)
Mandelbaum R., Slosar A., Baldauf T., Seljak U., Hirata C. M., Nakajima R., Reyes R., Smith R. E., 2013, MNRAS, 432, 1544
Mantz A., Allen S. W., Ebeling H., Rapetti D., 2008, MNRAS, 387, 1179
Mantz A., Allen S. W., Ebeling H., Rapetti D., Drlica-Wagner A., 2010, MNRAS, 406, 1773
Mantz A. B., Allen S. W., Morris R. G., Rapetti D. A., Applegate D. E., Kelly P. L., von der Linden A., Schmidt R. W., 2014, MNRAS, 440, 2077
Mantz A. B. et al., 2015, MNRAS, 446, 2205
Marshall P., Rajguru N., Slosar A., 2006, Phys. Rev. D, 73, 067302
Martin J., Ringeval C., Trotta R., Vennin V., 2014, Phys. Rev. D, 90, 063501
Merloni A. et al., 2012, preprint (arXiv:1209.3114)
Planck Collaboration I, 2014, A&A, 571, A1
Planck Collaboration I, 2015a, preprint (arXiv:1502.01582)
Planck Collaboration XIII, 2015b, preprint (arXiv:1502.01589)
Planck Collaboration XI, 2015c, preprint (arXiv:1507.02704)
Planck Collaboration XV, 2015d, preprint (arXiv:1502.01591)
Planck Collaboration XXIV, 2015e, preprint (arXiv:1502.01597)
Planck Collaboration XVI, 2015f, preprint (arXiv:1506.07135)
Planck Collaboration XIV, 2015g, preprint (arXiv:1502.01590)
Porciani C., Norberg P., 2006, MNRAS, 371, 1824
Raveri M., 2015, preprint (arXiv:1510.00688)
Riess A. G., 2014, The Local Measurement of the Hubble Constant, Presented in the 2014 Cosmic Distance Scale Workshop. Available at http://realserver4v.stsci.edu/t/data/2014/03/3951/AdamRiess033114.mp4
Riess A. G. et al., 2011, ApJ, 730, 119
Riess A. G. et al., 2016, ApJ, 826, 56
Ross A. J., Samushia L., Howlett C., Percival W. J., Burden A., Manera M., 2015, MNRAS, 449, 835
Sachs R. K., Wolfe A. M., 1967, ApJ, 147, 73
Salvatelli V., Marchini A., Lopez-Honorez L., Mena O., 2013, Phys. Rev. D, 88, 023531
Schuhmann R. L., Joachimi B., Peiris H. V., 2016, MNRAS, 459, 1916
Schwarz D. J., Copi C. J., Huterer D., Starkman G. D., 2015, preprint (arXiv:1510.07929)
Seehars S., Amara A., Refregier A., Paranjape A., Akeret J., 2014, Phys. Rev. D, 90, 023533
Seehars S., Grandis S., Amara A., Refregier A., 2016, Phys. Rev. D, 93, 103507
Sellentin E., Heavens A. F., 2016, MNRAS, 456, L132
Spiegelhalter D. J., Best N. G., Bradley P. C., van der Linde A., 2002, J. Roy. Stat. Soc., 64, 583
van Engelen A. et al., 2012, ApJ, 756, 142
Verde L., Protopapas P., Jimenez R., 2013, Phys. Dark Universe, 2, 166
Verde L., Protopapas P., Jimenez R., 2014, Phys. Dark Universe, 5, 307
Vikhlinin A. et al., 2009, ApJ, 692, 1060
Zaldarriaga M., Spergel D. N., Seljak U., 1997, ApJ, 488, 1

[6] http://www.darkenergysurvey.org
[7] http://www.mpe.mpg.de/eROSITA
[8] http://sci.esa.int/euclid/
[9] http://www.lsst.org

APPENDIX A: GAUSSIANIZATION PROCEDURE

Following Schuhmann et al. (2016), we compute a suite of optimized transformations to Gaussianize a distribution. We first apply a linear transformation $\mathbf{M}$, mainly to decorrelate strongly degenerate parameters. Thereafter, we apply a BoxCox transformation to each dimension individually. A BoxCox transformation is defined by

$$\mathrm{BC}_{(a,\lambda)}(x) = \begin{cases} \frac{1}{\lambda}(x+a)^{\lambda} - 1 & \text{if } \lambda \neq 0 \\ \log(x+a) & \text{if } \lambda = 0. \end{cases} \tag{A1}$$

The optimal transformation parameters are found by maximizing the probability that the transformed sample is Gaussian. This transformation is only defined for $x + a > 0$, so given an optimal a, the transformation is not defined for all x. However, we always choose $a > \max(-x_i)$ for a sample x_i such that the transformation is defined for every point of the sample, but not in every point of parameter space. For a sufficiently large sample, however, we can assume that the value of the probability density distribution is arbitrarily close to zero in regions without sample points.

After the first BoxCox transformation, we apply a principal component analysis (PCA), re-centring the sample by its mean $\boldsymbol{\mu}_{\mathrm{PCA}}$ and applying a linear transformation $\mathbf{L}^{-1}$ such that after the transformation the sample is standardized. The linear transformation can be obtained from a Cholesky decomposition of the covariance matrix $\mathbf{C}_{\mathrm{PCA}} = \mathbf{L}\,\mathbf{L}^{T}$.

After the PCA, we perform another family of transformations. Inspired by Schuhmann et al. (2016), we apply an Arsinh transformation defined by

$$\mathrm{Arsinh}_{(b,t)}(x) = \begin{cases} \frac{1}{t}\sinh(t(x-b)) & \text{if } t > 0 \\ x - b & \text{if } t = 0 \\ \frac{1}{t}\mathrm{arsinh}(t(x-b)) & \text{if } t > 0. \end{cases} \tag{A2}$$

The transformation is applied again to each dimension individually. The optimal transformation parameters are determined by maximizing the probability that the transformed sample is Gaussian, as done by Schuhmann et al. (2016). The Arsinh transformation is helpful, because it can transform away some excess kurtosis.

As the last transformation step, we apply again a BoxCox transformation. At this point, for our cases the samples we consider are well approximated by a Gaussian. Thus, we estimate the final mean $\boldsymbol{\mu}_{\mathrm{final}}$ and the final covariance $\mathbf{C}_{\mathrm{final}}$. Table A1 summarizes the transformations and the transformation parameters necessary in every point.

The Gaussianization procedure gives an analytic approximation to the distribution from which the original sample has been drawn. Any point in cosmological parameter space $\boldsymbol{\theta}$ needs to be transformed by the transformations shown in Table A1, yielding $\boldsymbol{\psi} = \mathrm{trans}(\boldsymbol{\theta})$. Then its likelihood can be approximated by using the expression derived by Sellentin & Heavens (2016), accounting for the scatter introduced by estimating the covariance of the sample. Using this method, we obtain analytic approximations for the P15 CMB likelihood for the models we consider. We make various of these products publicly available on https://bitbucket.org/grandiss45/gaussianization/.

Optimising the above given suite of transformations to optimally Gaussianize two samples allows one to jointly Gaussianize two distributions. A joint Gaussianization is theoretically not possible in general, but for prior and posterior distributions, a joint Gaussianization is feasible, because the posterior is generally better behaved than the prior. This allows us to estimate the surprise and its significance analytically using equation (3).

Table A1. Summary of the transformations (trans.) employed to Gaussianize a generic sample. We also specify the transformation parameters (params.) for each transformation. The index i runs from 1 to n_{dim}, which is the number of dimensions.

	Trans.	Params.
1st	linear	$\mathbf{M}$
2nd	BoxCox	$(a_i^{(1)}, \lambda_i^{(1)})$
3rd	PCA	$\boldsymbol{\mu}_{\mathrm{PCA}}, \mathbf{C}_{\mathrm{PCA}}$
4th	Arsinh	(b_i, t_i)
5th	BoxCox	$(a_i^{(2)}, \lambda_i^{(2)})$

A1 Test case

To provide an example of the Gaussianization procedure and compute the accuracy with which we can estimate the significance of tensions after Gaussianization, we construct the following test case. We start from two Gaussian distributions, shown in the upper left panel of Fig. A1, for which we can compute the significance of the tension analytically. Applying a non-linear transformation different from those in Table A1, we transform these samples to the degenerate constraints shown in the upper middle panel of the figure. This transformation is defined as follows. Given the two components of the original Gaussian samples, $x_{1,2}$, we transform them into $y_1 = x_1^{\beta}(x_2^{\alpha} + C)$ and $y_2 = x_1^{\beta}/(x_2^{\alpha} + C)$, where we choose $\alpha = 1.1$, $\beta = 0.4$, and $C = 4$. This transformation cannot be obtained analytically from the Gaussianizing transformations and therefore defines an interesting test case for the Gaussianization procedure. Furthermore, this simple case has a clear similarity to the P15 CMB constraints in the curved model.

As expected, the significance of the tension estimated from these samples is highly inaccurate, because the estimation of the surprise and its variance assumes that the distributions are Gaussian. In fact, all entropy derived quantities are invariant under arbitrary invertible parameter transformations, but our estimation assumes that the samples are Gaussian. Besides statistical uncertainty due to the finiteness of the samples, any systematic uncertainty in the estimation of the surprise derives from the fact that the underlying samples are not accurately described by Gaussian distributions.

By construction, after every Gaussianizing transformation, the samples are more accurately described by Gaussians and consequently the estimation of the significance is more accurate (see Fig. A1). After applying all the transformations we choose to perform (i.e. those in Table A1 with the exception of the first), the fractional accuracy on the significance is 1.5 per cent. Note that the final Gaussianized parameter space need not be equal to that of the initial Gaussian distributions.

A2 Accuracy of the Gaussianization

To estimate the accuracy of the significances of the underlying tensions of the cosmological constraints analysed in this work we propose the following scheme.

As described above, the accuracy of the significance depends solely on the degree to which the samples are well described by Gaussians. A natural measure of how well a distribution $\hat{p}$ describes

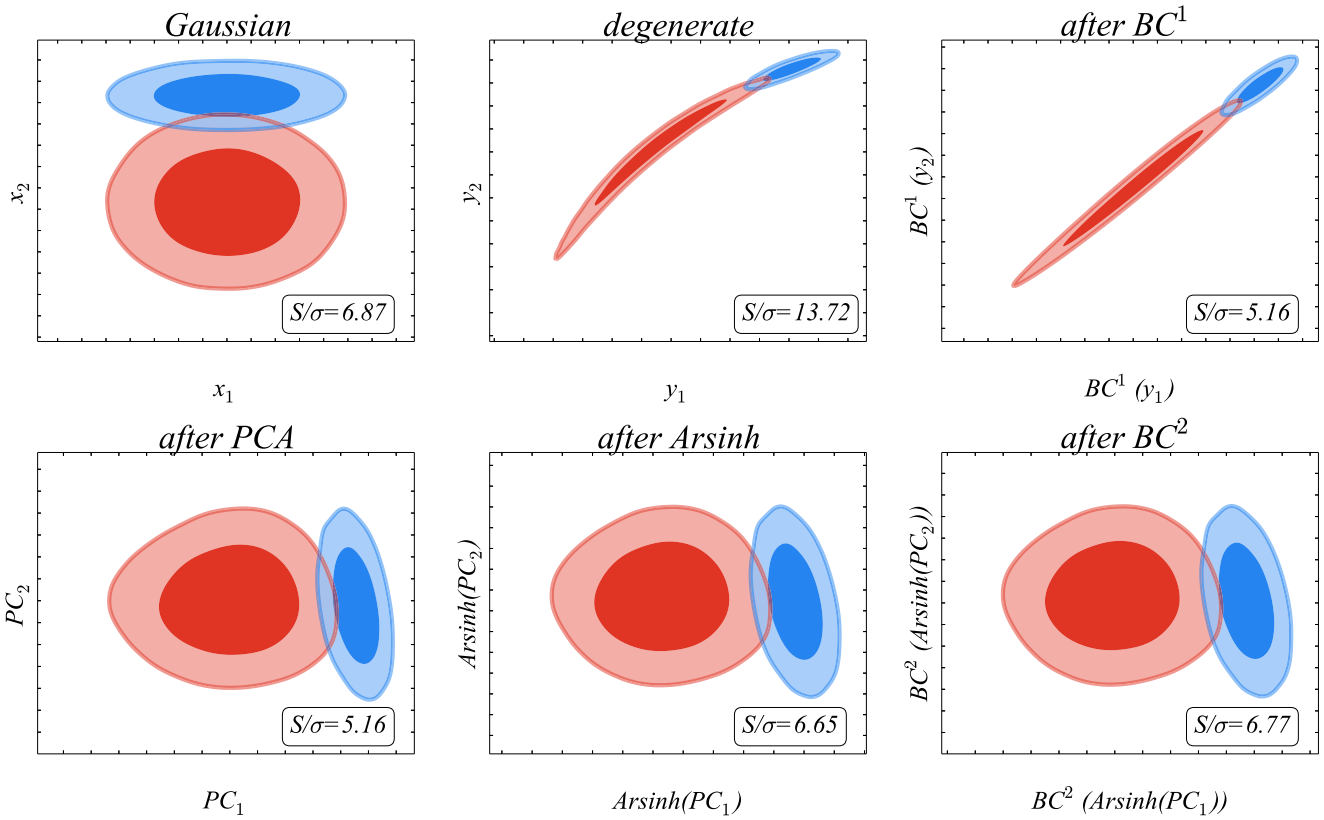

Figure A1. Test case, illustrating the Gaussianization procedure. The upper left panel shows the original, Gaussian samples, for which the significance of the tension S/σ can be computed analytically. By applying a non-linear mapping, these samples can be transformed into the degenerate constraints shown in the upper middle panel. The subsequent panels (upper right, lower left, lower middle and lower right) show the samples after applying the Gaussianizing transformations presented in Table A1 (except for the first). The final significance is very close to the correct, initial value, indicating that the samples are well approximated by Gaussian distributions.

a sample $\mathcal{X} = \{x_i\}$, with $i = 1, ..., N$, where N is the length of the sample, is given by the *logarithmic score*

$$H_{\mathcal{X}}^{\hat{p}} = \frac{1}{N}\sum_{i=1}^{N} \ln \hat{p}(x_i). \tag{A3}$$

If the sample points are drawn independently, the logarithmic score can be interpreted as the average log-likelihood that the sample $\mathcal{X}$ has been drawn from $\hat{p}$.[10] Consequently, a higher logarithmic score indicates a better fit.

For the case of the P15 CMB constraints in curved ΛCDM, labelled hereinafter *pr* for prior, and the joint constraints of P15 CMB and CMB lensing data, labelled *po* for posterior, we define $\hat{q}$ and $\hat{p}$, the approximations to the prior and posterior in the space of parameters after the Gaussianization process, as Gaussian likelihoods with means and covariances estimated from the transformed samples. We then draw 2000 samples $\mathcal{X}^{pr}$ and $\mathcal{X}^{po}$ from $\hat{q}$ and $\hat{p}$, respectively, and compute the logarithmic scores $H^{\hat{q}}$ and $H^{\hat{p}}$. We also evaluate the logarithmic scores of the original samples $H_{pr}^{\hat{q}}$ and $H_{po}^{\hat{p}}$. We find that

$$H_{pr}^{\hat{q}} = -3.000 \text{ and } \langle H^{\hat{q}} \rangle = -2.998 \pm 0.007, \tag{A4}$$

for the prior, and

$$H_{po}^{\hat{p}} = -3.000 \text{ and } \langle H^{\hat{p}} \rangle = -2.994 \pm 0.006, \tag{A5}$$

for the posterior. We note that for both, the prior and the posterior, the Gaussian samples $\mathcal{X}^{pr,po}$ on average fit better than the original samples *pr* and *po*. The logarithmic scores, however, are consistent with those of these Gaussian samples within the statistical uncertainties of the sampling process. It is thus safe to assume that the original samples are fitted by $\hat{q}$ and $\hat{p}$ to an accuracy consistent with the statistical noise of samples of their size.

To evaluate how large the impact of this statistical sampling noise is on the errors of estimating the significance, for each of the 2000 cases we estimate the significance S/σ of the tension between the sample drawn from $\hat{q}$ and the sample drawn from $\hat{p}$. We find that the average $\langle S/\sigma \rangle = 4.23 \pm 0.03$. For the case of P15 CMB versus P15 CMB plus CMB lensing in curved ΛCDM we have from Table 3 that $S/\sigma = 4.18$. This implies an average absolute error of 0.05, which corresponds to a fractional error of 1.1 per cent. Note that this should be interpreted as a systematic error. For the other data combinations and models, since this case is in no way special, we expect similar results.

In summary, the Gaussianization process is successful within the statistical uncertainties of the samples, and introduces systematic errors of the order of only 1 per cent, thus allowing a robust inference of the significance of an underlying tension.

[10] Given the likelihood $L(x_i|\hat{p}) = \hat{p}(x_i)$ that the sample points x_i are drawn independently from $\hat{p}$, the likelihood that the sample $\mathcal{X}$ is drawn from $\hat{p}$ is $L(\mathcal{X}|\hat{p}) = \prod_i \hat{p}(x_i)$. Thus, $\ln L(\mathcal{X}|\hat{p}) = \sum_i \ln \hat{p}(x_i) = N\, H_{\mathcal{X}}^{\hat{p}}$.

APPENDIX B: EVIDENCE RATIOS

It is common practice in cosmology to use $\ln R$, as derived from equation (4), to assess the agreement between two data sets D_1 and D_2, and $\ln R = 0$ is used as a reference point for such assessments. However, $\ln R$ depends on D_1 and D_2, which themselves are random variables, making also $\ln R$ a random variable. Consequently, in this section we follow the reasoning of Seehars et al. (2016) and for a class of likelihood models we propose a statistically well motivated reference point. We also analyse the statistical scatter of the measure $\ln R$.

B1 Statistics in one dimension

We first consider a simple one dimensional model with a flat prior $p(\theta)$ and likelihood

$$L(D_i|\,\theta) = \frac{1}{\sqrt{2\pi s_i^2}} \exp\left(-\frac{1}{2}\left(\frac{\theta - D_i}{s_i}\right)^2\right) \text{ for } i = 1,\ 2, \quad \text{(B1)}$$

where s_i are the uncertainties of the data sets. These likelihoods are normalized in a way that $E(D_i) = 1$. This model accurately describes the constraints on H_0 from the CMB and distance ladder measurements used here both in flat and curved ΛCDM, and the constraints of SNe and CMB on Ω_M in flat ΛCDM.

In this setting, the joint distribution of the parameter θ and the data sets D_1, D_2 is given by

$$p(\theta,\ D_1,\ D_2) = \frac{1}{2\pi\sqrt{s_1^2 s_2^2}} \exp\left(-\frac{1}{2}\frac{(D_1 - D_2)^2}{s_1^2 + s_2^2}\right) \times \exp\left(-\frac{1}{2}\frac{(s_1^2 + s_2^2)(\theta - \mu)^2}{s_1^2 s_2^2}\right), \quad \text{(B2)}$$

with $\mu = (s_2^2\, d_1 + s_1^2\, d_2)/(s_1^2 + s_2^2)$. Marginalizing the expression (B2) over the parameter θ with the flat prior gives the joint evidence of D_1, D_2 in the form

$$E(D_1,\, D_2) = \frac{1}{\sqrt{2\pi(s_1^2 + s_2^2)}} \exp\left(-\frac{1}{2}\frac{(D_1 - D_2)^2}{s_1^2 + s_2^2}\right), \quad \text{(B3)}$$

which illustratively is a Gaussian distribution of the difference between the data sets $\Delta D = D_1 - D_2$, with variance given by the sum of the variances of the single data sets. Note also that dividing equation (B2) by equation (B3) gives the posterior distribution $p(\theta|\, D_1,\ D_2)$, which consistently has expected value $\mathbb{E}[\theta|\, D_1, D_2] = \mu = (s_2^2\, d_1 + s_1^2\, d_2)/(s_1^2 + s_2^2)$ and variance $\mathrm{Var}[\theta|\, D_1, D_2] = s_2^2 s_1^2/(s_1^2 + s_2^2)$.

Using equation (B3) and equation (4) we can compute $\ln R$ analytically

$$\ln R = -\frac{1}{2}\frac{\Delta D^2}{s_1^2 + s_2^2} - \frac{1}{2}\ln\left(s_1^2 + s_2^2\right) - \frac{1}{2}\ln(2\pi)\,. \quad \text{(B4)}$$

From this expression, it becomes clear that perfectly agreeing data sets ($\Delta D = 0$) will have $\ln R < 0$ to a degree depending mainly on the measurement uncertainties. For example, one could obtain $\ln R = -6$, when comparing the two measurements $D_1 = D_2 = 0 \pm 114$. Using Jeffreys' scale for the natural logarithm, we would describe these results as the data sets being in 'strong disagreement', but in fact the data could not agree better! This example should clarify the importance of calibrating $\ln R$ correctly. In the same spirit as that used to calibrate the relative entropy, we propose $\langle \ln R \rangle_{D_1, D_2}$, the expected evidence ratio, as the reference point from which to assess the agreement between two data sets. In our simple model this quantity can be computed analytically as follows

$$\langle \ln R \rangle_{D_1, D_2} = \int \mathrm{d}D_1\, \mathrm{d}D_2\, E\,(D_1,\, D_2)\, \ln R = = -\frac{1}{2}\ln\left(s_1^2 + s_2^2\right) - \frac{1}{2}\ln(2\pi) - \frac{1}{2}. \quad \text{(B5)}$$

Combining equations (B4) and (B5), we find that the calibrated evidence ratio is given by

$$\ln R - \langle \ln R \rangle_{D_1, D_2} = -\frac{1}{2}\frac{\Delta D^2}{s_1^2 + s_2^2} + \frac{1}{2}, \quad \text{(B6)}$$

which effectively cancels the second term of equation (B4), which depends on the data set uncertainties. Applying this calibrated evidence ratio to the previous example we find $\ln R - \langle \ln R \rangle_{D_1, D_2} = 1/2$, so a better agreement than statistically expected.

Equation (B6) also allows a direct comparison of the calibrated evidence ratio and the surprise, because both are normalized and have scatter around 0. There is, however, a subtle difference in the way the surprise and the calibrated evidence ratio spot tensions between two data sets D_1, D_2. The calibrated evidence ratio is a symmetric measure of the consistency between the two data sets in data space. It considers directly the square difference between the data sets compared to the sum of their variances. The surprise is not symmetric and acts in parameter space, as can be seen in equation (3). Instead, it considers the agreement between $p(\boldsymbol{\theta}|\, D_1, M)$ and $p(\boldsymbol{\theta}|\, D_2, D_1, M)$, and assesses how probable the difference between $p(\boldsymbol{\theta}|\, D_1, M)$ and $p(\boldsymbol{\theta}|\, D_2, D_1, M)$ is. It goes after the question: given D_1, how probable is it that D_2 shifts the mean values of $p(\boldsymbol{\theta}|\, D_1, M)$ to the mean value of $p(\boldsymbol{\theta}|\, D_2, D_1, M)$? Consequently, it is suited to test whether D_2 should be added to the constraints of D_1, which is in general different from the question of adding D_1 to D_2.

As with the surprise, we can also derive an expected fluctuation of the calibrated evidence ratio $\sigma(\ln R)$

$$\sigma^2(\ln R) = \left\langle (\ln R - \langle \ln R \rangle)^2 \right\rangle_{D_1, D_2} = = \left\langle \left(-\frac{1}{2}\frac{\Delta D^2}{s_1^2 + s_2^2} + \frac{1}{2}\right)^2 \right\rangle_{D_1, D_2} = \frac{1}{2}. \quad \text{(B7)}$$

Thus, in the previous example, the calibrated evidence ratio has a significance 0.7σ. Calibrating and calculating the scatter of the $\ln R$ for more general likelihoods and priors, however, might require costly numerical computations. For this reason, we prefer the surprise as a measure of tension in the current analysis.

B2 Estimation for Gaussian likelihoods

For a Gaussian likelihood such as that in equation (B1), which approximates the distance ladder measurements of H_0 in flat and curved ΛCDM and the SNe constraints on Ω_M in flat ΛCDM, we have $E(D_1) = 1$. If we want to compute the evidence ratio between these and the base P15 CMB data set, D_2, we can use the fact that

$$R = \frac{E(D_1,\, D_2)}{E(D_1)\, E(D_2)} = E(D_1|\, D_2) = \int \mathrm{d}\theta\, L(D_1|\,\theta)\, p(\theta|\, D_2), \quad \text{(B8)}$$

where $p(\theta|\, D_2)$ is the posterior derived from D_2 (for a proof see Seehars et al. 2016). Given a sample of $p(\theta|\, D_2)$, and an analytic expression for $L(D_1|\,\theta)$, equation (B8) can be estimated with Monte Carlo Integration.

B3 Calibrating evidence ratios in n dimensions

For completeness, we give here the n-dimensional generalization of the equations given in Appendix B1. Assume a linear likelihood model for the data sets D_i, $i = 1, 2$ given by

$$L(D_i|\,\boldsymbol{\theta}) = \frac{1}{\sqrt{(2\pi)^n \det \boldsymbol{\Sigma}_i}} \exp\left(-\frac{1}{2}(\boldsymbol{\theta}-\boldsymbol{\mu}_i)^T \boldsymbol{\Sigma}_i^{-1}(\boldsymbol{\theta}-\boldsymbol{\mu}_i)\right), \tag{B9}$$

where $\boldsymbol{\theta}$ is the n-dimensional model parameter vector, $\boldsymbol{\mu}_i$ a n-dimensional vector depending linearly on the data set D_i, and $\boldsymbol{\Sigma}_i$ are symmetric $n \times n$ matrices, independent of the data set D_i and the model $\boldsymbol{\theta}$.

Integrating equation (B9) over a flat prior $p(\boldsymbol{\theta}) = 1$, we find the evidence $E(D_i) = 1$. Applying Bayes Theorem, we obtain the posterior distributions $p(\boldsymbol{\theta}|\,D_i) = L(D_i|\,\boldsymbol{\theta})$. Thus, $\boldsymbol{\mu}_i$ is the mean of the posterior $p(\boldsymbol{\theta}|\,D_i)$, and $\boldsymbol{\Sigma}_i$ its covariance.

Performing the same calculations as in the one dimensional case, we find the joint evidence

$$E(D_1, D_2) = \frac{1}{\sqrt{(2\pi)^n \det(\boldsymbol{\Sigma}_1 + \boldsymbol{\Sigma}_2)}} \times \exp\left(-\frac{1}{2}\boldsymbol{\Delta\mu}^T(\boldsymbol{\Sigma}_1 + \boldsymbol{\Sigma}_2)^{-1}\boldsymbol{\Delta\mu}\right). \tag{B10}$$

where $\boldsymbol{\Delta\mu} = \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2$ is the difference in means of the posterior distributions $p(\boldsymbol{\theta}|\,D_{1,2})$. This form is a manifest generalization of equation (B3). In the same way as described above, we can derive the evidence ratio

$$\ln R = -\frac{1}{2}\boldsymbol{\Delta\mu}^T(\boldsymbol{\Sigma}_1 + \boldsymbol{\Sigma}_2)^{-1}\boldsymbol{\Delta\mu} - \frac{n}{2}\ln 2\pi - \frac{1}{2}\ln\det(\boldsymbol{\Sigma}_1 + \boldsymbol{\Sigma}_2). \tag{B11}$$

We can thus confirm that also the n-dimensional evidence ratio scatters around a term that depends on the covariance. To find the correct zero-point, we need to calibrate it by subtracting its expected value. This gives the n-dimensional calibrated evidence ratio

$$\ln R - \langle \ln R \rangle = -\frac{1}{2}\boldsymbol{\Delta\mu}^T(\boldsymbol{\Sigma}_1 + \boldsymbol{\Sigma}_2)^{-1}\boldsymbol{\Delta\mu} + \frac{n}{2}, \tag{B12}$$

with a variance $\mathrm{Var}[\ln R] = n/2$.

Since the evidence is invariant under parameter transformations, these quantities could be easily estimated after a joint Gaussianization of the two independent posteriors $p(\boldsymbol{\theta}|\,D_1)$ and $p(\boldsymbol{\theta}|\,D_2)$. Here we did not use this method because we had a simpler access to the joint posteriors $p(\boldsymbol{\theta}|\,D_1, D_2)$, which are in general better behaved and thus easier to Gaussianize.

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