Atmospheric parameters of Cepheids from flux ratios with ATHOS[*]

I. The temperature scale

Bertrand Lemasle[1],[**], Michael Hanke[1],[**], Jesper Storm[2], Giuseppe Bono[3],[4], and Eva K. Grebel[1]

[1] Astronomisches Rechen-Institut, Zentrum für Astronomie der Universität Heidelberg, Mönchhofstr. 12-14, D-69120 Heidelberg, Germany
emails: lemasle@uni-heidelberg.de, mhanke@ari.uni-heidelberg.de.
[2] Leibniz-Institut für Astrophysik Potsdam (AIP), An der Sternwarte 16, D-14482 Potsdam, Germany
[3] Department of Physics, Università di Roma Tor Vergata, via della Ricerca Scientifica 1, I-00133 Roma, Italy
[4] INAF-Osservatorio Astronomico di Roma, via Frascati 33, I-00040 Monte Porzio Catone, Italy

ABSTRACT

Context. The effective temperature is a key parameter governing the properties of a star. For stellar chemistry, it has the strongest impact on the accuracy of the abundances derived. Since Cepheids are pulsating stars, determining their effective temperature is more complicated that in the case of non-variable stars.
Aims. We want to provide a new temperature scale for classical Cepheids, with a high precision and full control of the systematics.
Methods. Using a data-driven machine learning technique employing observed spectra, and taking great care to accurately phase single-epoch observations, we have tied flux ratios to (label) temperatures derived using the infrared surface brightness method.
Results. We identified 143 flux ratios that allow us to determine the effective temperature with a precision of a few K and an accuracy better than 150 K, which is in line with the most accurate temperature measures available to date. The method does not require a normalization of the input spectra and provides homogeneous temperatures for low- and high-resolution spectra, even at the lowest signal-to-noise ratios. Due to the lack of a dataset of sufficient sample size for Small Magellanic Cloud Cepheids, the temperature scale does not extend to Cepheids with [Fe/H] < −0.6 dex but nevertheless provides an exquisite, homogeneous means of characterizing Galactic and Large Magellanic Cloud Cepheids.
Conclusions. The temperature scale will be extremely useful in the context of spectroscopic surveys for Milky Way archaeology with the WEAVE and 4MOST spectrographs. It paves the way for highly accurate and precise metallicity estimates, which will allow us to assess the possible metallicity dependence of Cepheids' period-luminosity relations and, in turn, to improve our measurement of the Hubble constant H_0.

Key words. Stars: variables: Cepheids – stars: fundamental parameters – techniques: spectroscopic – Methods: data analysis

1. Introduction

The effective temperature, $T_{\rm eff}$, is one of the basic quantities characterizing a star and, in particular, one of the fundamental parameters shaping its spectrum. It is crucial for the accurate determination of chemical abundances because a biased $T_{\rm eff}$ has a strong impact on the abundances derived, much larger than the surface gravity ($\log g$), for instance. Temperatures are often derived from empirical relations employing photometry, by tying colors to temperature scales (e.g., Bell & Gustafsson 1989; Alonso et al. 1999) previously defined using for example the infrared flux method (IRFM, Blackwell & Shallis 1977). In this technique, the combination of the integrated flux and the infrared flux in a given band enables one to derive the effective temperature and the angular radius of the star.

However, such methods are also sensitive to other parameters (e.g., $\log g$) and can be strongly affected by unusual chemical compositions (Huang et al. 2015, and references therein). Moreover, they require accurate determinations of the interstellar extinction, which is always more difficult in the case of disk stars and especially for Cepheids since they are pulsating and located close to the plane of the Milky Way.

As far as Cepheids are concerned, many studies have tackled the $T_{\rm eff}$ issue. For instance, Pel (1978) did so using Walraven photometry and Kurucz (1975) models, Fernley et al. (1989) using the IRFM method and Kurucz (1979) models, and Kiss & Szatmáry (1998) using Strömgren photometry and Kurucz models. Welch (1994), Fouque & Gieren (1997), and Kervella et al. (2004b) calibrated optical and near-infrared surface-brightness (IRSB) relations anchored to interferometric measurements of the radius of Cepheids. Kervella et al. (2004a) report a remarkable agreement between the previous two Cepheid studies as well as with their own calibration of the surface-brightness relation for dwarf stars (Kervella et al. 2004c). Period-luminosity relations can then be determined

using distances derived from IRSB relations (Storm et al. 2011a,b). The Spectro-Photo-Interferometry of Pulsating Stars code (SPIPS, Mérand et al. 2015) now combines direct interferometric measurements of the Cepheid angular diameter (and its variations along the pulsation phase) with all other observables available into the same Cepheid model (Breitfelder et al. 2016). The method has improved the accuracy and precision on several model parameters.

Below 8000 K, the effective temperature dependence of the hydrogen lines makes them excellent T_{eff} indicators, especially given their independence on $\log g$. In particular, the wings of the Balmer lines provide one of the most reliable T_{eff} indicators for FGK stars (Fuhrmann et al. 1993; Hanke et al. 2018; Amarsi et al. 2018). However, in Cepheids, their distorted profiles (e.g., Nardetto et al. 2008) rules them out so far, in the absence of appropriate atmosphere models for pulsating stars.

The excitation equilibrium of Fe I and Fe II lines is commonly used to derive the effective temperature of FGK stars, including in our own analyses of Cepheids (e.g., Lemasle et al. 2007, 2008, 2017), under the assumptions of hydrostatic and local thermodynamic equilibrium (LTE). The first hypothesis obviously does not hold for pulsating variables and biases the results as recently verified for Cepheids by Vasilyev et al. (2017, 2018). The LTE hypothesis does not hold either (e.g., Fuhrmann 1998; Thévenin & Idiart 1999) and this is also true for Cepheids (e.g., Vasilyev et al. 2019). Alternatively, line depth ratios (LDRs) of pairs of lines with different sensitivity to T_{eff} have been put forward as temperature indicators by, e.g., Gray & Johanson (1991); Gray (1994). The number of ratios specifically calibrated to derive T_{eff} of Cepheids has increased from 32 ratios in Kovtyukh & Gorlova (2000) to 131 in Kovtyukh (2007) and 257 in Proxauf et al. (2018). LDRs pose highly precise observables that are capable of reproducing the relative temperature variations throughout the pulsation phase. Yet, an accurate absolute calibration of the scale is still pending (e.g., Kovtyukh & Gorlova 2000; Mancino et al. 2020) Employing spectroscopic T_{eff} indicators bears the major advantage that they are free of influences from reddening, an important asset in the case of classical Cepheids which are located close to the Galactic plane.

In this paper, we take advantage of the large sample of high-quality spectra that are available for Cepheids to establish a data-driven approach using flux ratios (FRs) as means to predict T_{eff}, in a similar approach as developed for stable stars by Hanke et al. (2018, henceforth H18) for the new code ATHOS (A Tool for HOmogenizing Stellar parameters). Sect. 2 describes the data preparation, in particular how labels, the meaningful information (here, T_{eff}) attached to each spectrum, were derived (Sect. 2.1). The method for selecting FRs (from the bluest wavelengths to the calcium triplet (CaT) region) is outlined in Sect. 3. We report our results in Sect. 4 and discuss them in Sect. 5.

2. Data preparation

2.1. Stellar parameters labels from the IRSB technique

The near-infrared (J,H,K) surface-brightness (IRSB) method (Fouque & Gieren 1997) is a Baade-Wesselink-type method that uses the pulsational variations in radial velocity, color, and luminosity to determine precise distances and radii of pulsating stars. It is based on the geometrical relation between the distance, d, to the star, the radius, R, and angular diameter, θ of the star at different pulsation phases, ϕ, namely:

$$\theta(\phi) = 2R(\phi)/d = 2(R_0 + \Delta R(\phi))/d \qquad (1)$$

With many data points at different phases this linear relation can be trivially solved for the stellar radius and distance.

The radius at a given phase is obtained by integrating the pulsational velocity curve for the star where the pulsational velocity is determined from the observed radial velocity by applying a suitable projection factor, p. The p-factor is largely a geometrical factor but depends also on limb darkening, the exact method used to derive the radial velocity (see Borgniet et al. 2019), and other technicalities. Storm et al. (2011a) calibrated the p-factor empirically and provide a detailed discussion and references on this issue.

The surface-brightness in the V-band, F_V, is defined as:

$$F_V(\phi) = 4.2207 - 0.1V_0(\phi) - 0.5\log\theta(\phi) \qquad (2)$$

where V_0 is the de-reddened visual magnitude. It was shown by Fouque & Gieren (1997) that the surface-brightness is well described by a linear relation in the near-infrared $(V - K)_0$ color. Storm et al. (2011a) adopted the calibration by Kervella et al. (2004a):

$$F_V = -0.1336(V - K)_0 + 3.9530 \qquad (3)$$

This calibration is based on interferometric angular diameters of Cepheids.

To determine $T_{eff}(\phi)$ from the Stefan-Boltzmann equation, in addition to the stellar radii just determined, it is necessary to determine the luminosity at each phase point given the absolute V-band magnitude M_V, and the bolometric correction. M_V is determined from the V-band light curve given the distance from above and the reddening. The bolometric corrections have been computed from the 2012 revised (ODFNEW[1]) ATLAS9 fluxes of Bessell et al. (1998) based on Castelli & Kurucz (2003) models, given the metallicity, $\log g$, and T_{eff}. As T_{eff} is needed to enter the table it was necessary to make a few (usually only one) iterations to make T_{eff} converge.

The surface gravity $\log g$ depends on radius, the adopted mass for the Cepheid, and at any given phase also the additional acceleration from the dynamical atmosphere. A mass of $6M_\odot$ was adopted in Storm et al. (2011a) for all the stars. The exact choice does not make a significant difference on the derived temperature. Had a mass of $8M_\odot$ been adopted then the value of T_{eff} would have changed by less than a degree.

The adopted metallicity also affects the bolometric correction and thus the resulting T_{eff} but the effect is very weak. An error in the adopted metallicity of 0.5 dex would lead to an error in T_{eff} of only 15 K.

Similarly, the adopted p-factor for the conversion between the observed radial velocity and pulsation velocity has no effect on the derived temperature as the resulting change in radius is exactly compensated by the changed distance and thus in the luminosity.

As shown already by Fouque & Gieren (1997), the distance and radius estimates are very robust to errors in the adopted reddenings. On the other hand, M_V depends directly on the reddening so the estimated temperatures are indeed sensitive to the adopted reddening. The effect of artificially lowering the adopted

---

[1] http://wwwuser.oats.inaf.it/castelli/colors/bcp.html

reddening, $E(B-V)$, by 0.05 mag is to reduce the estimated temperature by 150 to 200 K depending on the pulsation period of the Cepheid. This is the only significant systematic uncertainty apart from the adopted surface-brightness relation.

By construction (see Sect. 2.5), our spectroscopic dataset mostly consists of nearby Cepheids for which a wealth of external data is available. In particular, the values of the reddening, E(B-V), are as low as 0.003 mag for ζ Gem and below 0.2 mag for 46% of our sample. This ensures that the values of the reddening are either negligible or well-constrained in order to limit systematic errors on the labels. Our original sample also includes Cepheids in the Large Magellanic Cloud (LMC) and in the Small Magellanic Cloud (SMC) for which the E(B-V) values were determined using the reddening maps of Górski et al. (2020). E(B-V) values range from 0.081 to 0.175 mag for our LMC Cepheids and from 0.058 to 0.103 mag for our SMC Cepheids, with typical uncertainties of 0.005 mag.

The reddening value adopted is crucial for the determination of T_{eff} via the IRSB method. A simple test for the 7.6 day period Cepheid W Sgr showed that the estimated temperature is reduced by 190 K when $E(B-V)$ is artificially lowered by 0.05 mag with respect to the adopted value of 0.108 mag, whereas for the 41 day Cepheid RS Pup such a change of the reddening reduces the estimated temperature by 150 K. Table 3 indicates that the reddening values adopted in our IRSB analysis (used in Storm et al. (2011a) and based on the Fouqué et al. (2007) compilation) match closely those tabulated by Laney & Caldwell (2007). They are on the photometric system of Laney & Caldwell (2007) as described in Fouqué et al. (2007). Laney & Caldwell (2007) showed that externally determined reddenings (for instance, using main-sequence stars that either are companions of the Cepheid or belong to the same open cluster) fall on the same scale as those derived from multicolor photometry, hence providing a homogeneous reddening scale.

Fouqué et al. (2007) attempted to combine Cepheids' reddenings obtained from the 1970's on by different groups onto the system of Laney & Caldwell (2007) using linear transformations. From their Table 1, the scatter around the different transformations reaches 0.04 mag, while for the largest sample in that study (Fernie et al. 1995), it is only 0.03 mag. From the same table, it appears that the Dean Warren and Cousins system (Dean et al. 1978) is quite close to the Laney & Caldwell (2007) system, while most of the other studies[2] seem to cluster around similar values with a systematic offset of ~0.04 mag. Assuming two types of systems, one could claim systematic uncertainties of half of the offset or 0.02 mag. This is probably a too optimistic view and we adopt instead a more conservative value of 0.03 mag. It is worth mentioning that Turner (2016) using space reddenings reported that the values from Storm et al. (2011a) should be transformed linearly with an offset of +0.05 mag for a typical $E(B-V) = 0.5$ mag.

Numerous studies have focused on the reddening scale in the Magellanic Clouds. We used the maps obtained by Górski et al. (2020), that have a systematic uncertainty of the average reddening value of 0.013 mag. A comparison with previous studies in the LMC, for instance, indicates a mean difference of 0.061 mag, with a standard deviation of 0.012 mag between the reddening values from Górski et al. (2020) and those from Haschke et al. (2011). The same comparison between the results of Górski et al. (2020) and those of Inno et al. (2016) leads to an average offset of 0.004 mag with a standard deviation of 0.087 mag.

Reddenings are by far the main source of systematic uncertainties in our study, and improving the accuracy of Cepheids' reddenings will provide an exquisite accuracy for our flux ratio method.

2.2. Spectroscopic data

For training and testing the method, we collected a large number of high-resolution spectra at high signal-to-noise ratios (S/N). We limited ourselves to a small number of spectrographs, namely HARPS (Mayor et al. 2003) at the La Silla 3.6m telescope, HARPS-N (Cosentino et al. 2012) at the Telescopio Nazionale Galileo (TNG), UVES (Dekker et al. 2000) at the Very Large Telescope (VLT), FEROS (Kaufer et al. 1999) at the La Silla 2.2m telescope, and STELLA (Strassmeier et al. 2004; Weber et al. 2012) at the 1.2m telescope at Izana Observatory. We considered only spectra with S/N in excess of 50 pixel^{-1} for Galactic Cepheids, but had to weaken this requirement to a S/N of 30 pixel^{-1} for Magellanic Cepheids in order to be able to incorporate them in our sample[3]. Table 1 lists the properties of the spectrographs and the number of spectra gathered for each of them. Table 3 lists the Cepheids in the sample and the number of spectra gathered for each instrument.

Table 1. Characteristics of the spectra used in this study.

Spectrograph	Wavelength coverage [Å]	Resolving power	Number of spectra
FEROS	3530-9220	48 000	1
HARPS	3780-6910	115 000	274
HARPS-N	3780-6910	115 000	104
STELLA	3870-8700	55 000	123
UVES[(a)]	3730-9464	>34 500	822

Notes. [(a)] For UVES, we employed spectra from a variety of setups with different dichroics. Hence, we only show the extrema in terms of wavelength coverage and the minimum resolving power. Note that UVES dichroic observations are counted here as one spectrum, while the blue and red arm spectra appear in two different files in the ESO archive.

Spectra that were taken with FEROS, HARPS, and UVES were retrieved from the ESO (European Southern Observatory) archive. The spectra taken with HARPS-N were taken from the TNG archive. HARPS and HARPS-N products were used in the 1d spectral data format where individual spectral orders are merged. The STELLA spectra were collected by J. Storm. Since co-addition of the spectral orders is not included in the STELLA pipeline, we performed this operation using a S/N-weighted sum in the spectral regions where orders overlap.

One of the strengths of the our FR-based approach is that it does not rely on normalized spectra by restricting itself to small wavelength ranges for which the assumption of constant continua holds. For preparing the data, we thus only had to place the spectra in the stellar rest frame: we corrected the spectra for line-of-sight motions using a cross-correlation method with three different, synthetic template spectra that are representative

---

[2] The study by Eggen (1996) is an outlier with a slope significantly different from unity

[3] The S/N for HARPS spectra is either well above (nearby stars), or well below (spectra dedicated to radial velocity determination) this threshold.

of Cepheids and that were selected according to the temperature of the star at the time of the spectroscopic observation. The templates' characteristics are listed in Table 2. Finally, we performed a simple cosmic ray rejection by interpolating over features showing strong variances on scales much smaller than the width of the instrumental profile.

Table 2. Characteristics of the synthetic template spectra used for deriving the radial velocity using a cross-correlation method.

T_{eff} range	T_{eff} [K]	$\log g$ [dex]	[Fe/H] [dex]	v_t [km s^{-1}]
$T_{eff} < 5250$ K	4800	0.50	+0.1	2.0
5250 K $< T_{eff} < 6000$ K	5800	1.00	+0.1	2.0
$T_{eff} > 6000$ K	6600	1.30	+0.1	2.0

2.3. Training at a resolving power of $R = 5000$: pros and cons

All the spectra were then rebinned to a common resolving power of $R = 5000$. Training at such a low resolution limits the number of available relations when compared to training at higher resolution ($R \gg 10\,000$). This is not a problem since the accuracy remains driven by the intrinsic accuracy of the labels, and the precision remains better than any other method available.

Training at $R = 5000$ comes instead with two indisputable advantages. Firstly, the already high S/N of our spectra reaches extremely high (>1000) values when down-sampled to $R \sim 5000$, which makes the training insensitive to photon shot noise (this also applies to unseen data, since high-resolution spectra (outside of the training sample) would need to be degraded to the resolving power $R = 5000$ in order to derive T_{eff} using the method developed in this study). Secondly, in the perspective of large spectroscopic surveys with, e.g., WEAVE[4] (Dalton 2016), 4MOST[5] (de Jong et al. 2019), or MOSAIC[6] (Evans et al. 2015), it provides the same temperature scale for samples studied at low (of the order of $R = 5000$) or medium resolving power (of the order of $R = 20\,000$).

However, training at lower resolving power further comes with the caveat of being in all generality more metallicity sensitive because of blending features. This is not problematic within the range of parameters covered by the training sample, as it is intrinsically accounted for, but limits the applicability of the method to this exact same range (which is anyway always the case when using machine-learning techniques). In case of varying [X/Fe] among targets, this very circumstance poses an additional potential bias. It should be noted that in Cepheids, despite star-to-star variations, Genovali et al. (2015) reported that [α/Fe] shows a flat distribution across the entire Galactic thin disk (as one would expect for a young population), and no trend with $\log P$. As far as neutron-capture elements in Cepheids are concerned, da Silva et al. (2016) found a positive slope for [La/Fe], [Ce/Fe], [Nd/Fe], and [Eu/Fe] but no slope for [Y/Fe]. The same authors also reported no trend (except maybe for Ce) as a function of $\log P$. Including a large number of Magellanic Cepheids

---

in the sample is a strong requirement for improving the universality of the method.

At the beginning of Sect. 2.3, we have exposed our reasons for training at $R = 5000$. Determining T_{eff} for an unseen Cepheid spectrum at any resolving power R>5000 then requires that this spectrum is first degraded to the training resolving power (as mentioned above, no continuum normalization is required). It goes without saying that the method can therefore be applied only to spectra with $R \geq 5000$. By construction, it does not loose generality with varying resolution or T_{eff}. This could instead be the case for LDRs (see Mancino et al. 2020) because some line could disappear or become blended at a given T_{eff} which may itself depend on the resolution of the spectrum.

We set a higher priority on our main goal (providing a homogeneous T_{eff} scale for low- and high-resolution Cepheid spectra in upcoming surveys) than on potentially optimizing accuracy and precision via a different choice for the training resolution. Moreover, in the perspective of a canonical chemical analysis, we are fully satisfied with the accuracy and precision reached, as will be detailed in Sect. 4. A detailed analysis of the performances of flux ratios as a function of resolution will therefore not be undertaken in this paper. We anticipate that training at higher resolution will increase the number of calibrating relations and therefore the precision.

2.4. The need for an accurate phasing of the spectra

We quickly realized during the training process that even small errors in the phasing (of the order of a few hundredths of the period) of the training spectra could lead to relatively large biases (up to a few 100 K in the labels, see Fig. 1), and that these biases would be the parameter limiting the accuracy of our temperatures. Here we want to stress that the phase of the observations is only used to attribute T_{eff} labels to the spectra employing values derived via the IRSB method, and is not used further in the process.

The phase of the spectroscopic observations was initially determined using the Modified Julian Date (MJD) at which the spectra were taken, together with the exact same values of reference epoch and pulsation period that were used in the IRSB analysis. An extremely large (and constantly updated) collection of radial velocity measurements (spanning three decades) has been collected to apply the IRSB method (see, for instance, Storm et al. 2011a). To ensure that our high-resolution spectra (spanning 15 years) are placed on the exact same phase scale as the IRSB analysis, we checked our own radial velocity measurements against the IRSB radial velocity curves and phase shifts were applied to ensure a consistent phasing between both datasets. It is no surprise that they concerned primarily:

– stars for which the high-resolution spectra were taken at very different epochs from the spectra used for the IRSB analysis;
– Cepheids with known period changes. These period changes can be related to secular evolution of the Cepheids across the instability strip (e.g., Berdnikov et al. 2019) or to cycle-to-cycle variations that affect, in particular, long-period Cepheids (Anderson et al. 2014, 2016; Anderson 2016).

An alternative approach would have been to take into account period drifts. However, data are often too scarce to derive them accurately, and it is usually difficult to reach a consensus on the value of even a simple, linear period drift (see Breitfelder et al. 2016, and references therein). To our knowledge, RS Pup is the only Cepheid for which the modulation of the period over several decades has been derived accurately, using a

Star	$\log P$	$E(B-V)_{\mathrm{IRSB}}$ [mag]	$E(B-V)_{\mathrm{L07}}$ [mag]	F	H	N	S	U
					Spectrograph			
RT Aur	0.571489	0.059	0.079	⋯	3	⋯	3	89
AD Gem	0.578408	0.173	0.382	⋯	⋯	⋯	⋯	1
BF Oph	0.609329	0.235	0.223	⋯	2	⋯	3	5
T Vul	0.646934	0.060	0.052	⋯	3	⋯	26	⋯
FF Aql	0.65039	0.196	0.188	⋯	⋯	⋯	27	⋯
δ Cep	0.729678	0.075	0.073	⋯	⋯	104	18	⋯
BB Sgr	0.821971	0.281	0.302	⋯	2	⋯	3	5
U Sgr	0.828997	0.402	0.408	⋯	2	⋯	3	60
W Sgr	0.880529	0.108	0.450	⋯	3	⋯	⋯	⋯
S Sge	0.923352	0.099	0.084	⋯	⋯	⋯	19	⋯
β Dor	0.993131	0.051	0.042	1	46	⋯	⋯	78
ζ Gem	1.006497	0.014	0.003	⋯	47	⋯	3	85
Z Sct	1.110645	0.492	0.450	⋯	⋯	⋯	⋯	9
TT Aql	1.138459	0.435	0.417	⋯	2	⋯	3	50
Y Oph	1.233609	0.647	0.450	⋯	7	⋯	⋯	⋯
RU Sct	1.29448	0.911	0.924	⋯	⋯	⋯	2	43
RY Sco	1.307927	0.718	0.746	⋯	3	⋯	3	9
RZ Vel	1.309564	0.300	0.294	⋯	7	⋯	⋯	48
WZ Sgr	1.339443	0.435	0.450	⋯	1	⋯	3	9
VZ Pup	1.364945	0.455	0.450	⋯	⋯	⋯	1	60
T Mon	1.431915	0.179	0.165	⋯	10	⋯	3	55
l Car	1.550816	0.146	0.138	⋯	111	⋯	⋯	80
U Car	1.58897	0.263	0.260	⋯	6	⋯	⋯	50
RS Pup	1.61742	0.457	0.454	⋯	19	⋯	3	54
HV 6093	0.679881	0.081	⋯	⋯	⋯	⋯	⋯	4
HV 2405	0.840333	0.114	⋯	⋯	⋯	⋯	⋯	2
HV 12452	0.941454	0.175	⋯	⋯	⋯	⋯	⋯	2
HV 1335	1.157800	0.093	⋯	⋯	⋯	⋯	⋯	3
HV 1328	1.199692	0.058	⋯	⋯	⋯	⋯	⋯	3
HV 1333	1.212084	0.100	⋯	⋯	⋯	⋯	⋯	3
HV 822	1.223807	0.086	⋯	⋯	⋯	⋯	⋯	3
HV 1023	1.424235	0.121	⋯	⋯	⋯	⋯	⋯	2
HV 879	1.566167	0.146	⋯	⋯	⋯	⋯	⋯	2
HV 837	1.630509	0.076	⋯	⋯	⋯	⋯	⋯	2
HV 877	1.655215	0.128	⋯	⋯	⋯	⋯	⋯	2
HV 2369	1.684646	0.142	⋯	⋯	⋯	⋯	⋯	2
HV 2827	1.896354	0.12	⋯	⋯	⋯	⋯	⋯	2

Table 3. Properties of the Cepheids in the dataset: logarithm of the pulsation period (P), $E(B-V)$ from Storm et al. (2011a) used for the IRSB analysis, $E(B-V)$ tabulated by Laney & Caldwell (2007). The table also details the spectroscopic data available for these stars (F: FEROS, H: HARPS, N: HARPS-N, S: STELLA, U:UVES). The upper panel lists Galactic Cepheids, the lower panel Magellanic Cepheids. Note that SMC stars (identified in italics) have not been used for training or testing, but only for assessing that the current relations do not hold anymore at the metallicities of SMC Cepheids (see Sect. 5.1).

fifth-degree polynomial (Kervella et al. 2017). Given the current insufficient accuracy of period drifts, we adopted instead the method described above. We cannot exclude that a fraction of the remaining residuals discussed in Sect. 4.3 originates from a still inaccurate phasing of some spectra.

For a few stars, an accurate rephasing based on the radial velocity curve proved impossible. It is the case of the exceptional Cepheid X Sgr, where the signature of strong shock waves (emission, line doubling) has been reported by Mathias et al. (2006). In a few other stars (S Mus, EV Sct, U Aql, V Car, V Cen, and V350 Sgr), a strong orbital motion was detected, which limited the accuracy of the phasing, and these stars were therefore not further considered despite a good number of spectra available. Finally, a few other spectra were rejected because the radial velocity curves of the stars were of lower quality (possibly due to orbital motion as well) and/or with incomplete phase coverage (KQ Sco, RZ Vel: spectra between phases 0.9 and 0.1, S Nor, T Vel, UZ Sct, WZ Car, XX Sgr, Y Sgr).

2.5. The dataset

A good dataset must have the following characteristics:
- the labels are accurate;
- they cover the entire parameter space and are homogeneously distributed, not only in label space (T_{eff}), but also in other parameters that may affect an FR (e.g., [Fe/H] and $\log g$);
- a large number of spectra is available.

However, as with many if not all data-driven methods, it is close to impossible to generate such a dataset. Here we investigate whether the conditions listed above are satisfied.

We retained 24 Galactic Cepheids for our sample. They are listed in Table 3, together with the number of spectra used for each of them. We considered only stars for which a good number (>10) of spectra were available and for which accurate phases could be determined (see Sect. 2.4). We note in particular that a few stars are represented by a considerable number of
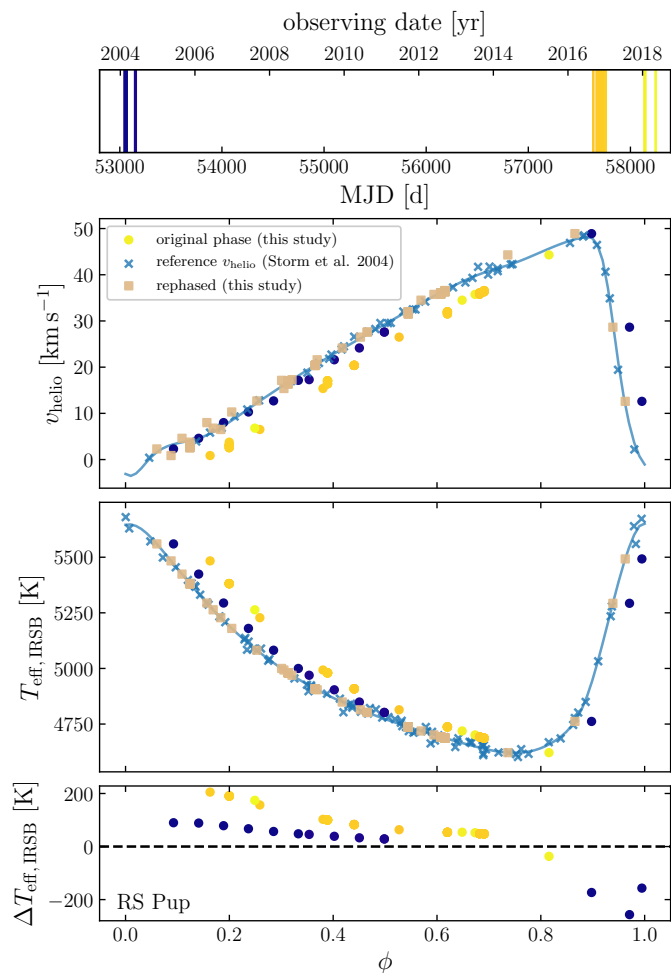
Fig. 1. Impact of erroneous phasing of individual spectra on the T_{eff} labels for the Cepheid RS Pup: the *top panel* displays the MJD of the observation for individual spectra. The *second panel* shows our individual radial velocity measurements (circles that are color-coded by MJD, cf., *upper panel*) versus the radial velocity data used for the IRSB analysis (blue crosses, Storm et al. 2004). Spectra rephased to match the radial velocity curve are shown as pale brown squares. Using the same color coding, the *third* and *fourth panels* display the T_{eff} variation, and the T_{eff} offset between original labels and their corrected value after rephasing, respectively.

spectra, namely *l* Car (191), ζ Gem (135), β Dor (125), δ Cep (122), and RT Aur (95). We end up with a total of 1324 spectra.

Although in the Magellanic Clouds only few spectra per star are available, 32 spectra of 13 Magellanic Cepheids were also added to the dataset. Including a large number of spectra of LMC and especially SMC stars in the sample would present the advantage that metallicity-dependent relations would naturally be excluded[7]. However, for these stars we only have few spectra, which in addition do not cover the entire wavelength range. This means that temperatures derived from flux ratios located

---

[7] In practice, the same number of spectra for metal-rich (typically Galactic) and metal-poor (typically SMC) Cepheids would be required. Such a sample of SMC Cepheids does not currently exist. An alternative would be to attribute very high weights to the SMC spectra, which comes with the caveat that the labels need to be extremely well constrained and the spectra need to have a high S/N and further be free of artefacts such as cosmic-ray hits. Unfortunately, none of the latter is satisfied for the spectra at hand.
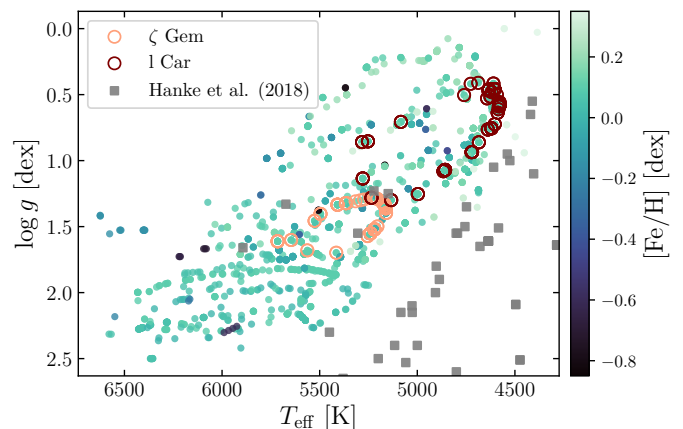
Fig. 2. Parameter space (T_{eff}, $\log g$, [Fe/H]) covered by the compiled sample of spectra. Highlighted by salmon and red circles are the Cepheids ζ Gem and *l* Car, for which we have the most spectra at hand (135 and 191, respectively). Gray squares indicate a small portion of the giants used in the ATHOS dataset for stable stars (H18).

in different spectral domains would be differently affected by possible metallicity dependences. This does not seem to be a satisfying outcome. We decided instead to train at high metallicity only and to check for T_{eff} consistency of individual relations at low metallicity (see Sect. 5.1).

The first condition we set for our sample is met: we selected bright, nearby Cepheids that were thoroughly investigated over the years. The radial velocity curves and the light curves are well sampled. The spectra were phased with great care. As a result, accurate labels could be derived following the method described in Sect. 2.1.

Our sample covers a broad range of periods (3.7 days $\leq$ $P \leq$ 78.8 days, see Table 3), which ensures a good coverage of the T_{eff}-$\log g$ space. Indeed, the coolest stars in our sample, *l* Car and RS Pup, have temperatures ranging from as low as 4578 K to 5282 K (respectively from 4621 K to 5560 K). The hottest Cepheid in our sample, RT Aur, has temperatures ranging from 5573 K up to 6592 K. For these stars, $\log g$ varies between 0.41 dex and 1.30 dex (*l* Car), between 0.07 dex and 1.39 dex (RS Pup), and between 2.16 dex and 2.50 dex (RT Aur). The other Cepheids have intermediate values for T_{eff} and $\log g$.

However, the condition is not met as far as [Fe/H] is concerned. The best-studied stars are naturally among the closest Cepheids (which also allowed to collect large amounts of spectra in a reasonable amount of observing time). Given the presence of abundance gradients in the Milky Way (e.g., Luck et al. 2011; Yong et al. 2012; Genovali et al. 2013), those Cepheids have metallicities close to Solar. Cepheids in the Milky Way span almost 1.0 dex in [Fe/H], from $\approx$+0.5 dex in the inner disk (e.g., Pedicelli et al. 2010; Andrievsky et al. 2016) to $\approx$–0.5 dex in the outer disk (e.g., Lemasle et al. 2008; Luck & Lambert 2011; Genovali et al. 2014). Due to their distances (>3 kpc) and their location close to the plane, less spectra are available for the stars located at the edges of the metallicity range, and the IRSB data is also scarce. To extend the metallicity coverage, a good alternative is to turn to the Magellanic Clouds, where a few tens of Cepheids have been studied spectroscopically at high resolution (Romaniello et al. 2008; Lemasle et al. 2017) and for which IRSB analyses are available (Storm et al. 2011b; Gieren et al. 2018). [Fe/H] roughly ranges from –0.2 to –0.6 dex in the

LMC and from –0.6 to –0.9 dex in the SMC. Unfortunately, very few medium- or high-resolution spectra are available, and even less when requiring S/N>30 pixel^{-1}. Moreover, all stars are represented by a unique spectrum taken at random phase, which means that we do not have at our disposal spectra covering a large range of T_{eff}, $\log g$. and periods for low-metallicity Cepheids.

The third condition is largely met: we have gathered a large number of spectra for each of the training Cepheids (see Table 3), namely a grand total of 1324 spectra, which provide a full phase coverage for each of the selected stars. The number of non-pulsating benchmark stars for covering the entire Hertzsprung-Russell diagram is considerably lower (Jofré et al. 2014, *Gaia* benchmark stars), Hanke et al. (2020, HD 20).

3. Method

The approach pursued throughout this work closely follows the one described in H18. To this extent, in order to enable the computationally efficient handling of the data, all spectra were binned to a common log-linear dispersion scale, where the *n*th element is expressed as

$$\log_{10} \lambda_n = \log_{10} \lambda_0 + n \cdot \delta \log_{10} \lambda, \text{ for } n \in \{0, ..., N\}, \quad (4)$$

with λ_0 being the starting wavelength of 3800 Å, $\delta \log_{10} \lambda$ the dispersion spacing, $2.7 \cdot 10^{-5}$, and N the total number of pixels, here taken to be 13 333. This way, our maximum considered wavelength is ~ 8705 Å. Employing a log-linear λ scale bears the main advantage that profile widths remain approximately constant in pixel space[8] over the extensive range used here.

3.1. Feature selection

We recall here the definition from H18 of a spectral FR between two wavelength windows that are centered at λ_i and λ_j, respectively:

$$\mathrm{FR}_{\lambda_i, \lambda_j} = \frac{\langle F_{\lambda_i} \rangle}{\langle F_{\lambda_j} \rangle}, \quad (5)$$

where

$$\langle F_{\lambda_i} \rangle = \frac{1}{5} \sum_n F(\lambda_n), \quad n \in \{i - 2, ..., i + 2\} \quad (6)$$

is the mean flux of five pixels.

Each FR potentially qualifies as feature that encodes information about T_{eff}, which in principle leaves us with $O(10^{10})$ explanatory variables. This enormous number calls for an efficient and robust feature-selection process prior to the actual training and validation of feasible relations. Our first – and arguably most important – pre-selection criterion considered only those FRs for which the two flux windows are not separated by more than 30 pixels. While reducing the number of features by two orders of magnitude, this limit makes the FRs essentially scale-free, that is, normalization-independent, since it ensures that the continuum variation between the two closely-spaced

regions of interest can be safely neglected (see Sect. 4.2). An additional applied criterion was that only those FRs were allowed as features that showed a maximum range of values in excess of 0.15. This was enforced to ensure that noise has a limited impact on the predicted labels from each FR (see also Sect. 4.1).

As a next step, we computed Pearson correlation coefficients, $\rho_{\mathrm{FR}, T_{\mathrm{eff}}}$, between all remaining FRs and the T_{eff} labels associated to the underlying spectra. By requiring strong (anti-)correlations (i.e., $|\rho_{\mathrm{FR}, T_{\mathrm{eff}}}| > 0.93$) between the features and the labels, we further restricted the feature space to 33 291 candidates. We emphasize at this point that by proceeding like this we potentially rejected a large number of tight feature-label relations, in particular those that are strongly non-linear and/or non-monotonic. This is owed to using computationally cheap correlation coefficients that test for linearity and could only be circumvented by investing substantially larger computational resources.
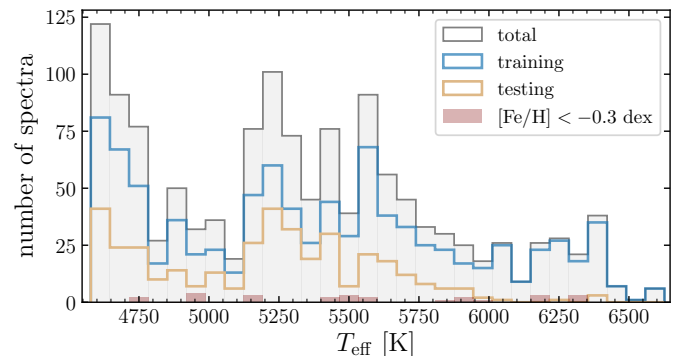
3.2. Training and testing

Fig. 3. Temperature distribution of the spectra in our dataset, shown as a gray histogram. The temperature distributions of the training and testing samples are shown in blue and pale brown, respectively. We also indicated in red the temperature distribution of the spectra of stars with |Fe/H| < −0.3 dex that did not enter the training.

For fitting, the dataset was split into a training set (70% of the sample) and a holdout (or testing) dataset (30%), which was used for unbiased estimates of the model fitness. In principle, it would be desirable to have at the same time both a homogeneous coverage in label space and the training and testing sets obeying identical probability distributions. In practice, however, since our T_{eff} distribution is skewed towards lower temperatures (see Fig. 3), we decided to foster T_{eff}-homogeneity for the training sample by randomly drawing with the inverse temperature density as weight. As a consequence, the holdout set follows a slightly different probability distribution, which is indicated in Fig. 3. The only alternative approach would be to disregard a substantial number of spectra in the intermediate- and low-temperature regime from the analysis to establish equal numbers throughout the whole T_{eff} range. Yet, we favor our choice for two reasons: First, some spectra have a discontinuous wavelength coverage due to chip gaps and others are limited to the visual range and do not extend to the near-infrared CaT region (8498–8662 Å). Hence, not all FRs can be measured in all spectra, which adds an additional layer of complication that is best mitigated by having a higher number of spectra

---

[8] Amounting to $(5000 \cdot (10^{\delta \log_{10} \lambda} - 1))^{-1} \approx 3.22$ pixels from the instrumental resolution alone. We note that the effective line width may be larger owing to intrinsic line broadening mechanisms, such as rotation and/or macroturbulent motions in the stellar atmosphere.
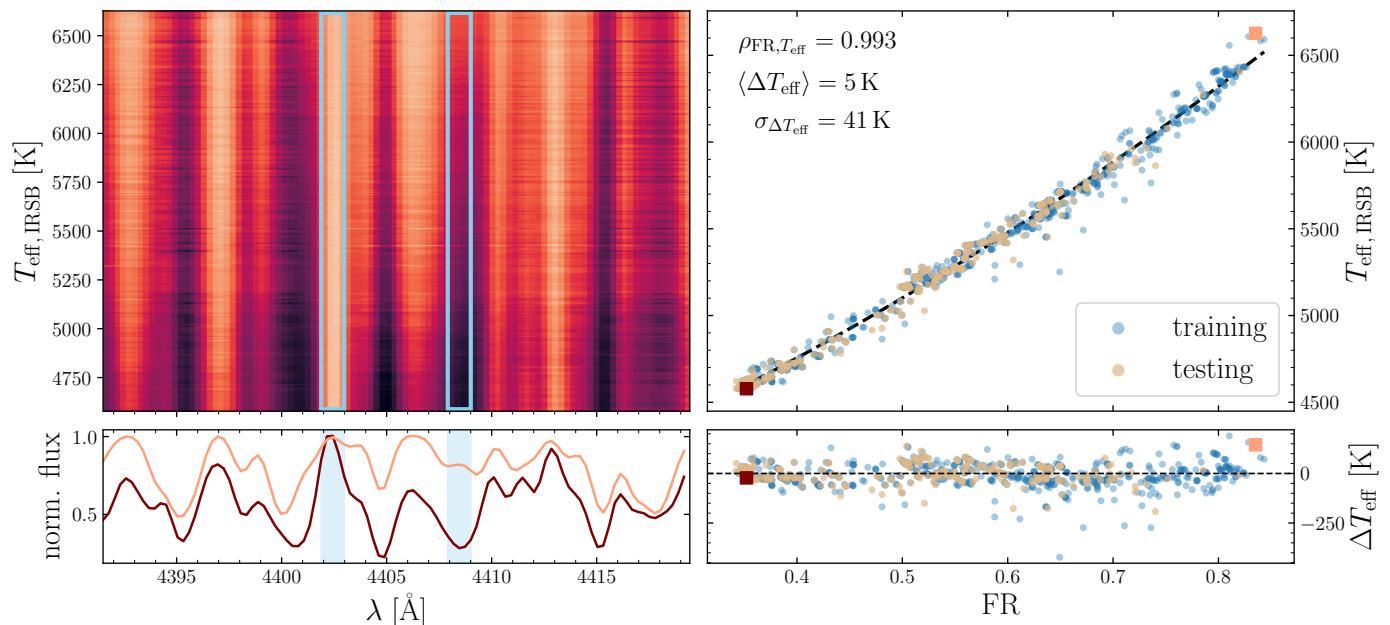
Fig. 4. Properties of an individual relation using flux ratios around 4405 Å. The *top left panel* shows the evolution of the neighboring spectral domain with T_{eff}, where darker colors indicate deeper features. In the *bottom left panel*, the coolest (dark brown) and hottest (orange) spectra at our disposal are displayed. For illustration purposes, the spectra were pseudo-normalized by the 99th percentile of the flux in the presented range. In both panels, the regions considered for computing the flux ratios are marked in clear blue. The *top right panel* shows the analytical relation (dashed black curve) together with the T_{eff} labels for the training (blue) and testing (pale brown) samples. Using the same color coding, the *bottom right panel* shows the difference between the temperatures computed using the individual analytical relation and the input labels. In both panels, the coolest and hottest spectra are again depicted in dark brown and orange, respectively.

from different spectrographs. The second reason for keeping all observed spectra is that it enables a better coverage in fundamental parameters other than T_{eff}, such as [Fe/H], $\log g$, or period, which are key to spotting degeneracies with these higher-order parameters.

Even though we used correlation coefficients for feature selection and therefore technically pre-filtered our features for close-to linear relations, we allowed our model to be more flexible by adding the possibility of curvature through second-order polynomials. It is noteworthy at this point that having such a simple model renders a validation set for tuning hyper-parameters superfluous. One such case is presented in Fig. 4, where we, on the one hand, exemplarily demonstrate how our most well-constrained spectral region (see next section for the selection criteria) around ~ 4405 Å behaves with temperature, and, on the other hand, we illustrate how the identified FR constrains T_{eff} both in the training and testing sets.

The actual training was performed on each of the 33 291 FRs that survived the feature selection. While the model was fit using the training set, its fitness in terms of accuracy (mean deviation) and precision (standard deviation, σ) was independently determined from the testing set. To lessen the impact of artefacts on individual FRs induced by, for example, erroneous phasing, wavelength calibration errors, and cosmics, we allowed a clipping of up to five of the most extremely deviating spectra both from the training and testing sets. The distribution of the respective findings for the individual features is depicted in Fig. 5.

We found that there is a slight accuracy bias that tends towards negative values with decreasing precision. This could be owed to the circumstance that for any set of two physically identical spectra in the training and testing sets, respectively, the test-
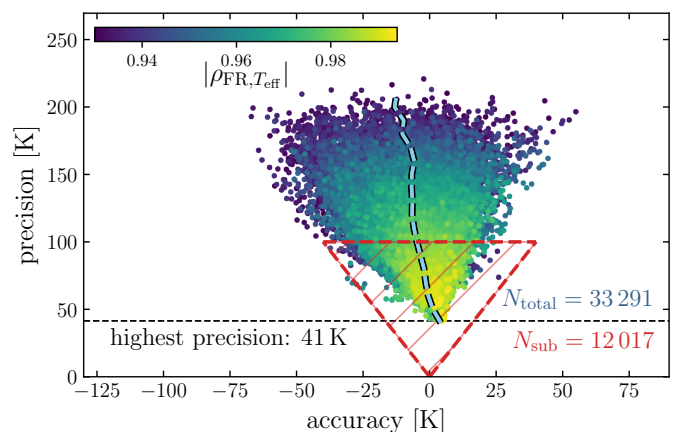

Fig. 5. Precision and accuracy (in K) for all the relations retrieved by the analysis, color-coded by their correlation coefficient. Only relations for which the precision is better than 100 K and less than 3 × the accuracy are further considered (red triangle). The dashed line shows a slight accuracy bias (see text) as a function of precision, which was computed using a boxcar median filter.

ing spectrum shows an offset T_{eff} label. Such an effect could occur if these spectra are associated with different Cepheids, where inter-target systematics (e.g., reddening, projection factor, etc.) could hamper the labeling in the IRSB analysis. Nevertheless, given that the systematic bias remains well below 25 K, we do not investigate this behavior further (but see Sects. 4.3 and 5.2).

Not surprisingly, the precision correlates with the correlation coefficient, although – due to allowing for curvature in the model – we deem the precision to be the more desirable quantifier for the goodness of the model fit.
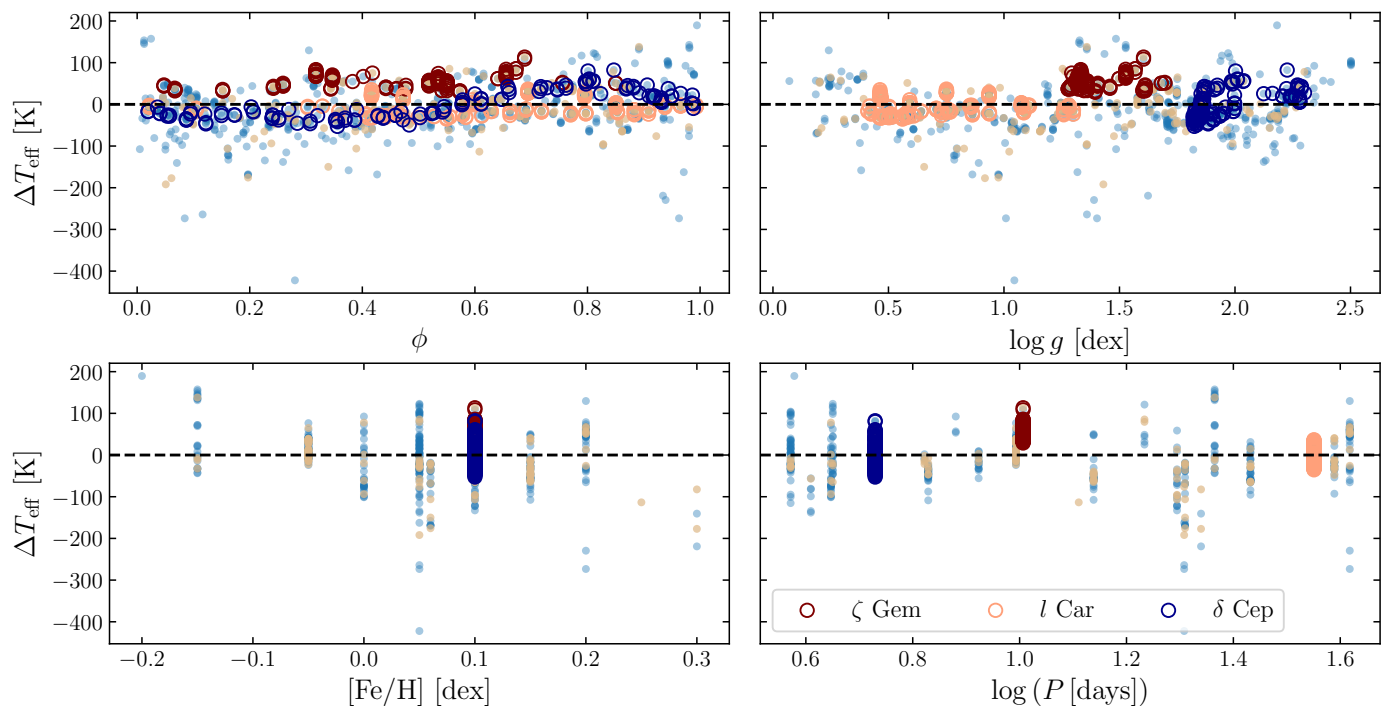
Fig. 6. For the same relation as in Fig. 4, the residuals $T_{\text{eff, IRSB}} - T_{\text{eff, FR}}$ are shown as a function of phase (ϕ, *top left*), $\log g$ (*top right*), [Fe/H] (*bottom left*), and $\log P$ (*bottom right*). Colored open circles highlight spectra for the targets ζ Gem, l Car, and δ Cep.

3.3. Final selection of flux ratios

After having fit all candidates for useful features, we imposed further quality cuts, in a sense that only such features were kept, for which the model assessment yielded precision/accuracy > 3 and the precision remained below 100 K. These criteria are superimposed in Fig. 5 and resulted in 12 017 FRs for further analyses.

Next, we had to make sure that the final selection of features is linearly independent. Only then, the predicted labels from all models and unseen data can be averaged in a meaningful way, that is, without having to deal with inevitable (unknown) covariances[9]. We achieved this by ranking the features by their associated model precision and subsequently exclude those that have pixel overlaps with a feature of higher precision. This procedure further reduced the number of leftover features to 247, which indicates that the vast majority are clustered around a limited number of T_{eff}-sensitive line profiles and are only shifted by incremental λ steps with respect to each other.

Finally, the remaining 247 relations were visually checked for erratic behaviors in the residuals $T_{\text{eff, IRSB}} - T_{\text{eff, FR}}$ to exclude (strong) influences on the FRs from parameters in excess of T_{eff}. Another considered factor was substructure that is indicative of an insufficient description by our quadratic model (e.g., discontinuities or non-flat residuals as a function of FR). For our best relation, the residuals are presented with respect to ϕ, $\log g$, [Fe/H], and $\log P$ in Fig. 6. There, we also highlight the Cepheids ζ Gem, l Car, and δ Cep as representatives with good phase and $\log g$ coverage as well as very different pulsation periods. It is

evident that there is a persistent, statistically significant bias of ~ 59 K between ζ Gem and l Car ($\langle \Delta T_{\text{eff}} \rangle = 56 \pm 16$ K and -3 ± 19 K, respectively, both only computed from the testing set) irrespective of phase and $\log g$. We would like to stress at this point that this shows that what we call precision of a relation (i.e., the scatter measured from test spectra of many different Cepheids) is largely governed by star-to-star systematics in the labels and therefore encompasses external errors. In fact, the internal precision of a single of our FR relations as deduced from one target alone (in the example above ζ Gem and l Car) resides below 20 K. The reader is referred to Sect. 4.3 where we discuss in detail the implications of labeling systematics on the full sample of relations. After our visual sanity check, we are left with 143 relations.

4. Results

After carefully selecting and phasing the dataset, we trained our relations using IRSB T_{eff} labels and obtained 143 calibration relations that can be used to determine the effective temperature of Cepheids. Before describing the properties of the identified relations, we want to test the stability of individual relations with respect to the signal-to-noise ratio and continuum variations, and briefly discuss the remaining systematics on the final temperature.

4.1. Signal-to-noise ratio stability

Given a high-S/N (e.g., > 100 Å^{-1}) spectrum, even a single of our discovered features is capable of predicting T_{eff} at a precision that is more than sufficient for subsequent abundance

---

[9] We note, in particular, that some lines are used simultaneously in several LDR relations (Proxauf et al. 2018), their depth being employed sometimes in the numerator and sometimes in the denominator of the ratio. For a detailed comparison of the IRSB labels used in this paper and the LDR method, we refer the reader to paper II in this series.

[10] 4MIDABLE-LR: Chiappini et al. (2019)
[11] 1001MC: Cioni et al. (2019)
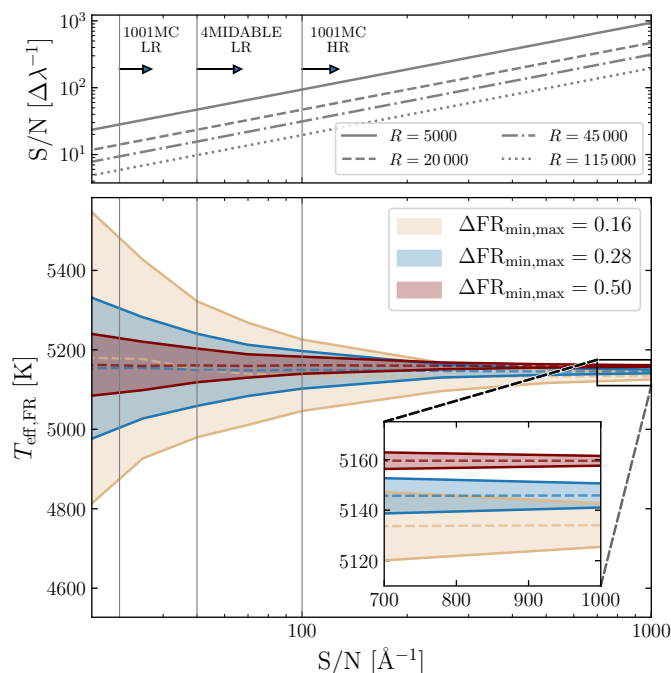[12] 4MIDABLE-HR: Bensby et al. (2019)

Fig. 7. Stability against S/N per Å for three relations of different maximum FR excursion (cf., color coding and legend in the *lower panel*). The corresponding S/N per resolution element for four different resolving powers (see legend) are given in the *top panel*. Shown are the measured mean values (colored dashed lines) and 1σ standard deviations (solid lines enclosing filled areas) from an MC simulation with artificially injected noise (see main text for details). The zoom-in view emphasizes the (marginal) relation-to-relation inaccuracies (see middle panel of Fig. 10). Vertical lines indicate the minimum S/N requirements for different 4MOST consortium surveys: S/N = 50 Å^{-1} in the magnesium or calcium triplet regions (4MIDABLE-LR[10]); S/N = 30 Å^{-1} in the magnesium or calcium triplet regions (1001MC[11], LR); S/N = 100 Å^{-1} at 6200 Å (1001MC, HR; 4MIDABLE-HR[12]).

analyses (i.e., $\sigma_{T_{\rm eff,FR}} \ll 100$ K). In practice, for larger samples this is only feasible for nearby and therefore bright Cepheids for which spectra can be obtained within reasonable exposure times and/or 8+ m-class telescopes. In light of the fact that upcoming spectroscopic campaigns such as WEAVE and 4MOST at 4 m-class facilities will only allow for a more restrictive spectrum quality (S/N $\gtrsim$ 30 Å^{-1}), our extended number of relations with a wide wavelength coverage is paramount. In order to optimize the averaging strategy for predicted labels from multiple features, knowledge about the S/N stability is desirable.

The parameter affecting the noise sensitivity most is $\Delta{\rm FR}_{\rm min,max}$, which denotes the excursion an FR covers over the full range of $T_{\rm eff}$. To test this, we have conducted a Monte Carlo (MC) analysis using a single-epoch spectrum of ζ Gem with attributed parameters ($\phi = 0.48$, $T_{\rm eff} = 5206$ K, $\log g = 1.29$ dex, [Fe/H] = +0.10 dex) that are roughly consistent with the midpoint of the parameter coverage of our dataset (cf., Fig. 2). The spectrum was obtained with $R \approx 115\,000$ and S/N = 276 pixel^{-1} using HARPS. At 0.01 Å per pixel, this latter value corresponds to S/N $\approx$ 2760 Å^{-1}, which leads us to the conclusion that for tests with artificial noise satisfying S/N $\leq$ 1000 Å^{-1} contributions from the real observed noise level can be safely neglected. To estimate both mean values and standard deviations at various S/N values between 25 and 1000 Å^{-1}, per S/N realization,

we generated 500 spectra with injected random noise, degraded each to $R = 5000$, measured FRs, and predicted the $T_{\rm eff}$ labels.

The result can be seen in Fig. 7 where we indicate the course of $\sigma_{T_{\rm eff,FR}}$ as a function of S/N for three representative features with different realizations of $\Delta{\rm FR}_{\rm min,max}$; the lowest allowed spacing of 0.15, an intermediate value (0.28), and a large excursion (0.50). For these three relations, the noise-induced standard deviation ranges from (366 K, 177 K, 77 K) at S/N = 25 Å^{-1} through (90 K, 47 K, 22 K) at S/N = 100 Å^{-1} to (9 K, 5 K, 2 K) at S/N = 1000 Å^{-1}. Hence, even at S/N = 25 Å^{-1}, from a mathematical point of view – i.e., neglecting artefacts that real spectra inevitably suffer from – averaging only 16 of our relations with the strongest (worst) S/N sensitivity already yields internal temperature precisions that are sufficient for meaningful abundance analyses (366 K/$\sqrt{16}$ = 92 K).

An important – though not unexpected – observation is that, considering the scatter, the mean $T_{\rm eff,FR}$ is perfectly stable with S/N. This highlights the superiority of our method over many already existing analysis tools that show systematically increasing over- or underestimations of the continuum level with varying S/N. As a consequence, further down the line such tools introduce potentially strong systematic scatter when comparing results (e.g., [Fe/H] or other chemical abundances) obtained from spectra of substantially varying quality (see for instance Fig. 2 in Reichert et al. 2020.

4.2. Stability against continuum variations

Another investigated source of error is the stability of $T_{\rm eff}$ predictions from our models against extreme variations of the continuum level between two involved wavelength ranges due to, for instance, a mistreated removal of the instrument's blaze function. As opposed to noise-induced errors – which can be reduced by averaging a number of relations – potential errors from this are of systematic origin and ultimately affect our accuracy. In Fig. 8 we illustrate how the effect differs for two FRs at the extreme ends of the distribution of wavelength spacings, $\Delta\lambda$. We employed the same spectrum of ζ Gem introduced in the previous section and multiplied linear continua with slopes as drastic as $\pm 0.1\,\%\,$Å^{-1}. Not even the steepest slopes exceed a ± 25 K effect for the closely-spaced case, whereas a maximum excursion of ± 50 K was found for the largest $\Delta\lambda$. One might desire to drop or downweight the ratios with largest $\Delta\lambda$ when analyzing high-dispersion spectra with narrow orders, since they are potentially affected by somewhat stronger continuum variations than other devices.

4.3. Size and origin of remaining systematics

In the two previous sections, we investigated individual relations. From now on, we are dealing with the average $T_{\rm eff}$ derived from a large number of relations, and not anymore from a temperature derived from a single relation.

Thanks to our efforts to minimize systematics, we are left with small-scale residuals (a few tens K) that show different patterns, displayed in Fig. 9. The first pattern is a phase-dependent offset that seems independent from the MJD of the observation, at least for the relatively small range of MJDs considered here (case of, e.g., δ Cep). The residuals are likely related to the IRSB method. We checked, however, that there is no phase shift between the optical and near-infrared photometry, nor with the radial velocity data (see Storm et al. 2011a,b, and references therein). Since the peak of the deviation happens at maximum
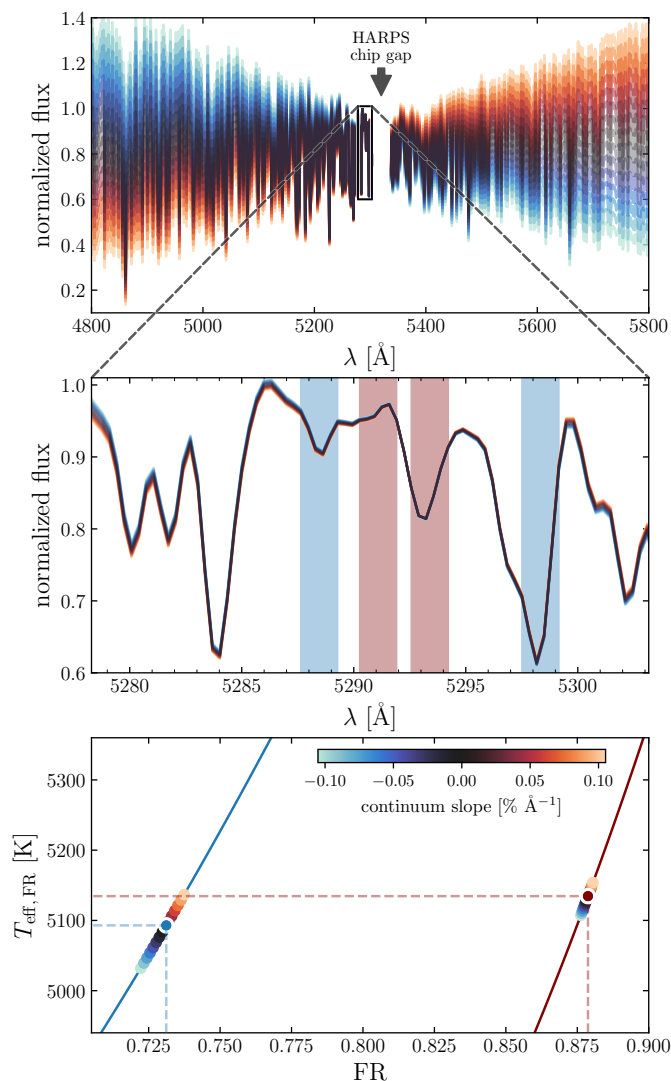
Fig. 8. Impact of continuum variations on FRs from close- (red) and wide-spaced (blue) wavelength components. *Upper panel*: 15 times the same ζ Gem spectrum as was used in Sect. 4.1 with varying continuum slopes (see color bar in *lower panel*). The missing part in the presented spectrum corresponds to HARPS' chip gap. *Middle panel*: zoomed-in view of the region around $\sim$ 5295 Å where the λ regions are highlighted by vertical spans. *Lower panel*: deduced $T_{\mathrm{eff,FR}}$ values as a function of the measured FRs at different continuum slopes. The models are represented by red and blue solid curves while dashed red and blue lines and filled circles of the same colors indicate the finding from the original spectrum without any modification of the continuum.

and minimum radii, we also investigated whether the radius amplitude could be influenced by the value selected for the p-factor. Using p=1.3 instead of the nominal value (p=1.41), we found that the effect on $T_{\mathrm{eff,IRSB}}$ is less than 2 K over the entire phase range, thus excluding the p-factor from the possible source for the residuals. The second pattern is a similar phase-dependent offset, this time MJD-dependent (case of, e.g., l Car). Although the residuals are again likely related to the IRSB analysis, their variation probably originates from an inaccurate rephasing or from cycle-to-cycle variation of the temperature curve. The third pattern resembles a constant offset between the retrieved temperature, $T_{\mathrm{eff,FR}}$, and the labels (case of, e.g., ζ Gem), possibly due to an erroneous value for a global, phase-independent parameter (reddening?) in the IRSB analysis (see Sect. 2.1).

4.4. Properties of the identified relations

It is not a surprise that the majority of the flux ratios involves features in the blue region of the spectra (Fig. 10, top panel), given the wealth of potentially temperature-sensitive metallic lines in this spectral domain. However there is still a good number of relations in the [5000–6000] Å range, and more specifically around the magnesium triplet (MgT) and the Hα line. A handful of ratios will certainly prove very useful in the CaT region.

There are only three relations overlapping the spectral range of the Gaia Radial Velocity Spectrometer (RVS, Sartoretti et al. 2018). As shown in the second and third panels of Fig. 10, their precision and accuracy are among the lowest within our set of relations, probably because only a small fraction of the spectra in our sample cover this wavelength range. Such a paucity would also prevent a proper outlier rejection in case of undesirable artifacts in the analyzed spectra. At this stage, we would therefore caution against using these three lines solely, for instance in order to analyze Gaia RVS spectra. However, training at the resolution of the RVS ($R = 11\,500$) might provide more calibration relations with better accuracy and precision. This will be investigated in a forthcoming paper.

In Fig. 11, we display a few interesting spectral regions harboring flux ratios:

(a) The first feature we show is located in the blue spectral region, in the vicinity of Hβ. FRs involving the wings of hydrogen lines are typical T_{eff} indicators as already found by H18. A T_{eff}-sensitive feature (the blend of a Ni I and a Fe I line) is normalized by another spectral domain in the blue wing of Hβ, which varies much less with T_{eff}.

(b) The second example is located in the MgT region (5167-5184 Å). One of the spectral domains is located directly in the bluemost of the Mg I lines of the Mgb triplet and varies notably with T_{eff} while the other one is only slightly temperature-sensitive.

(c) Another example involves metallic lines, where the blend of a Fe I line and a Ti I line shows a strong dependence on T_{eff}, which seems to be compensated by the wing of a Fe I line whose sensitivity on T_{eff} is much less pronounced and might also correct for the effect of another parameter like $\log g$ or [Fe/H].

(d) Finally, we show a flux ratio with two components located close to a H I line in the near-infrared. It does not have a typical behavior, as both components have an opposite sensitivity to T_{eff}: the first one is located in the wing of the H I line, which becomes stronger when T_{eff} increases, while the other one lies in the wing of a Fe I line, which becomes more prominent when T_{eff} decreases.

Before examining the second panel of Fig. 10, we wish to recall the argument already made at the end of Sect. 3.3: what we call precision is the scatter measured from test spectra of many different Cepheids. It combines two effects, of which star-to-star systematics is the dominant one, while the internal precision of an individual relation as deduced from a single target falls below 20 K. Keeping this in mind, we see that the precision on individual relations, capped at 100 K by construction, is actually much better (Fig. 10, second panel). Indeed roughly half of the relations have an individual precision better than 80 K, and the best relations have precisions ranging from 40 to 60 K; they cover, however, a more restricted wavelength domain.

When combining all relations, we are able to trace T_{eff} variations of a few Kelvins. For instance, the middle bottom panel of Fig. 9 indicates the outcome of our analysis when investigat-
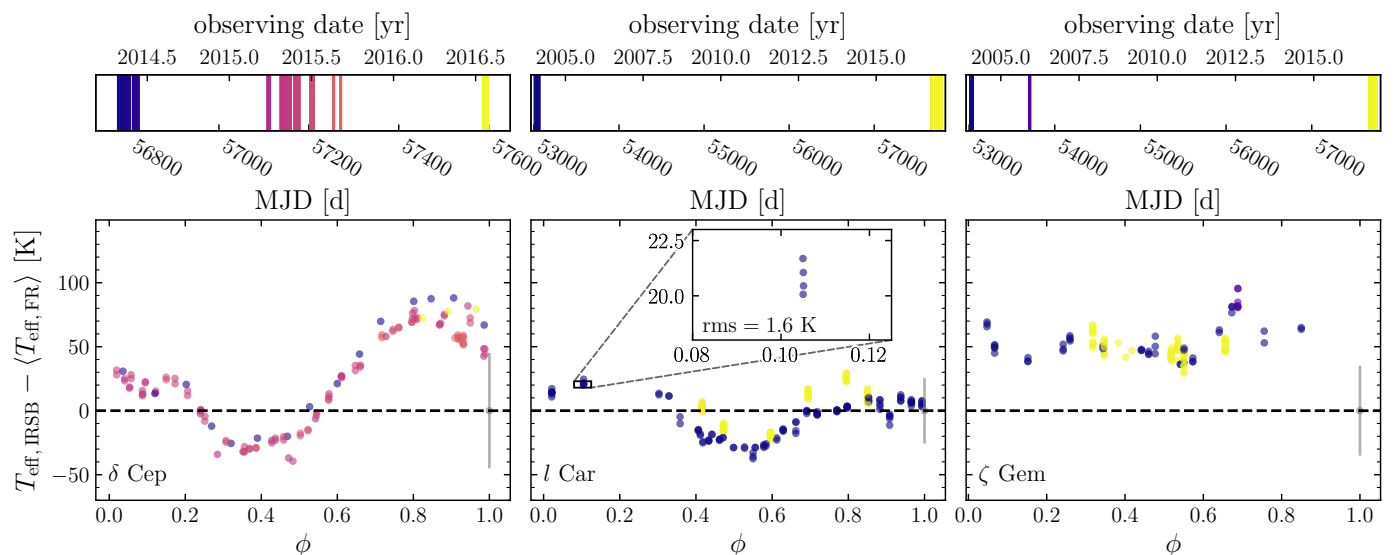
Fig. 9. Representative examples of the three distinct residual patterns as discussed in the text: δ Cep (IRSB issues seemingly constant with MJD, *left panel*); *l* Car (IRSB issues with MJD-dependence, *middle panel*); ζ Gem (constant offset possibly due to erroneous reddening, *right panel*). Shown are the median residuals of the best 100 trained relations that were measured in individual spectra with respect to the input labels. The abscissa indicates the pulsation phase, ϕ. As in Fig. 1, the color coding resembles the time of observation of the individual spectra (cf., *upper panels*). Typical values for the 1σ scatter are depicted by gray error bars. The inlay in the *middle panel* highlights five spectra of *l* Car that were taken with HARPS back to back within less than 30 min. The rms scatter of 1.6 K underlines the internal precision of our method.

ing consecutive short exposure spectra of *l* Car (P~35.5d) taken with HARPS within less than 30 min. We achieved a rms scatter of 1.6 K (using only the 100 best relations), which underlines the internal precision of our method.

Although our criterion was slightly looser (±40 K), the accuracy of our relations mostly falls within ±10 K (Fig. 10, third panel) and rarely exceeds 20 K, thus ensuring that we reproduce almost exactly the $T_{\rm eff}$ scale of the input labels. From a practical point of view, we would recommend to make use of only a selected number of the best relations (since their individual accuracy and precision are known) rather than sigma-clipping temperatures derived from individual relations, if the goal of the operation is to improve the precision of the final $T_{\rm eff}$. Note that this will probably come at the cost of the wavelength coverage, as shown in the second and third panel of Fig. 10. An alternative (possibly better) approach would be to perform a weighted median using the (known) inverse variance (i.e., squared precision) of individual relation as weight.

The last two panels of Fig. 10 display the excursion $\Delta\mathrm{FR}_{\rm min,\,max}$ covered by flux ratios, and the wavelength spacing $\Delta\lambda$ of the two components of the flux ratios, respectively. They both can be interpreted by the larger occurrence of temperature-sensitive features in the blue regions of Cepheids' spectra.
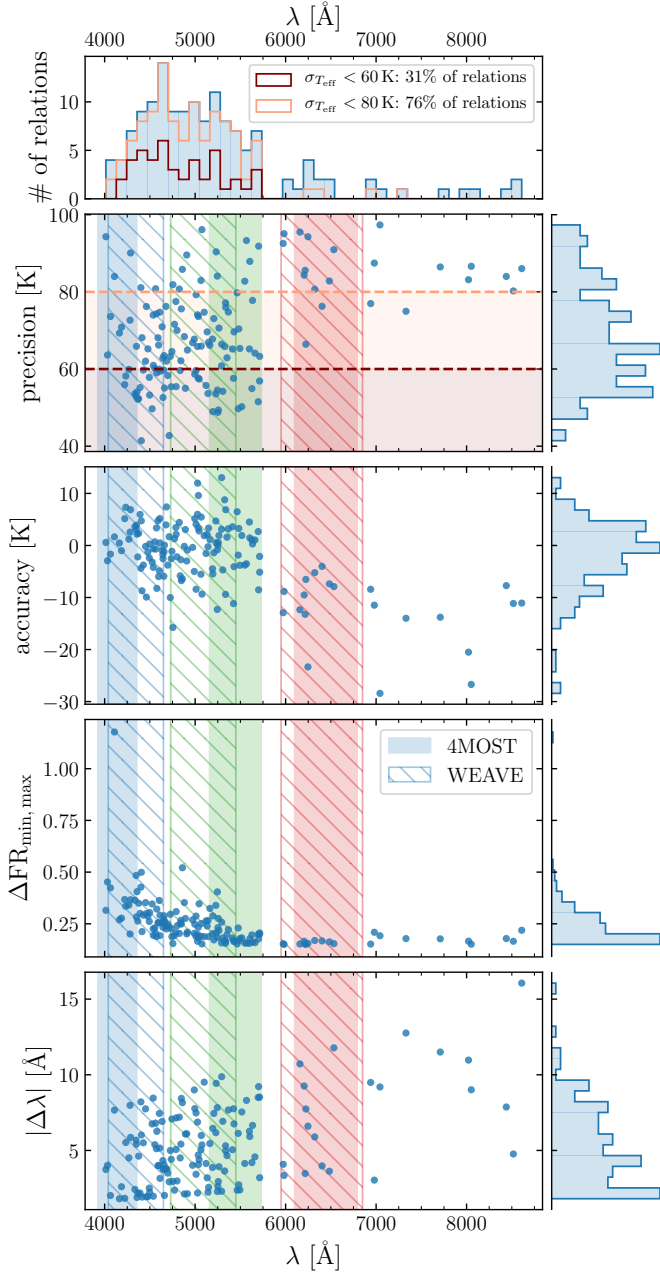
Fig. 10. Distribution of the properties of FRs as a function of wavelength. Blue, green, and red vertical spans depict the three spectral windows of 4MOST (filled) and WEAVE (hatched) high-resolution spectrographs. *Upper panel*: Number of independent relations to derive $T_{\rm eff}$ from flux ratios per wavelength range. *2nd panel*: Precision of individual relations to derive $T_{\rm eff}$ from flux ratios as a function of wavelength. *3rd panel*: Accuracy of individual relations to derive $T_{\rm eff}$ from flux ratios as a function of wavelength. *4th panel*: Flux ratio excursion, $\Delta {\rm FR}_{\rm min, max}$, as a function of wavelength. *Bottom panel*: Wavelength spacing $\Delta\lambda$ of the two components of a ratio as a function of wavelength.
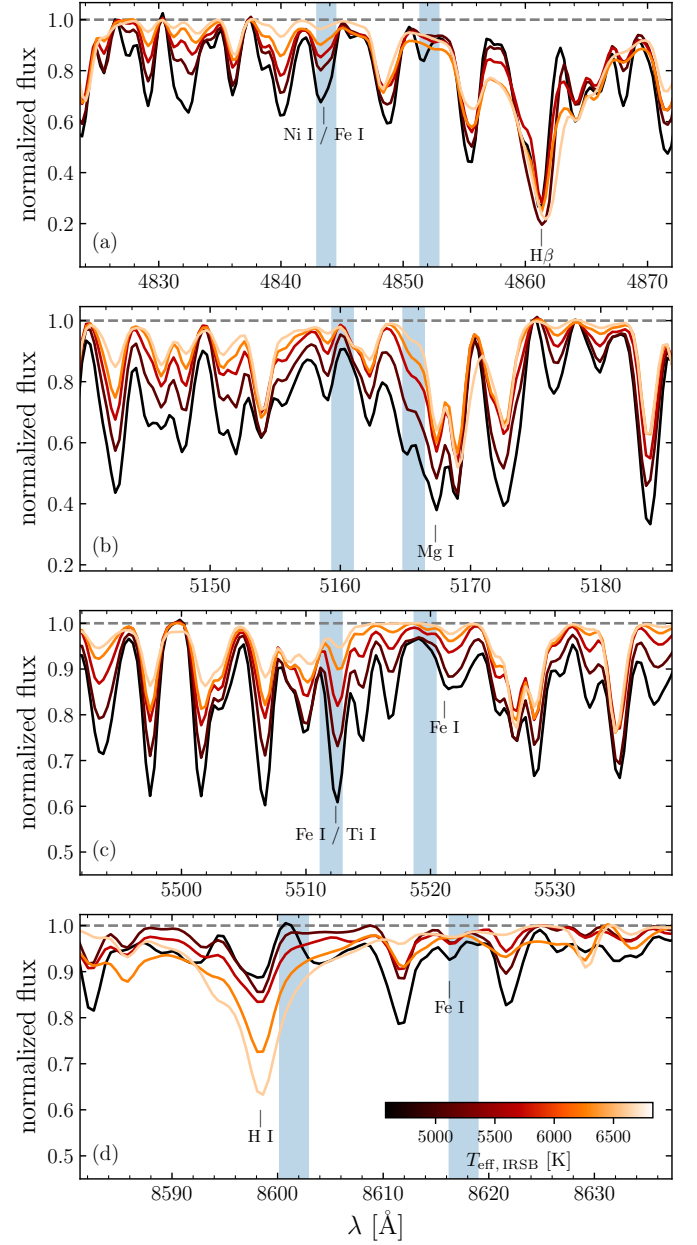
Fig. 11. Illustration of a few spectral domains where $T_{\rm eff}$-sensitive flux ratios have been recovered: a) blue wing of Hβ, b) MgT region, c) metallic lines around 5510 Å, d) H I line in the near-infrared. See Sect. 4.4 for details. Spectra are color-coded by temperature, where darker colors indicate cooler temperatures. For illustration purposes, the spectra were pseudo-normalized by the 99th percentile of the flux in the presented range.

5. Discussion

5.1. Applicability at low metallicity

We used the small number of spectra of SMC Cepheids at our disposal to test the applicability of our method at low metallicity. Since the spectra do not cover the entire wavelength domain, not all the relations could be investigated. For those that could be tested, we see, however, a clear pattern exemplified in Fig. 12: the relations hold down to [Fe/H]~-0.6 dex and the residuals quickly increase below this value, reaching ≈800 K at [Fe/H]~-0.8 dex.

This is not a surprise, as the training sample does not contain any low-metallicity SMC Cepheids (see Sect. 2.5). In its current stage, the method is then applicable to Galactic Cepheids in the outer disk and to a large fraction of the LMC Cepheids, but not to the bulk of the SMC ones. Here we want to stress that the current values of [Fe/H] have been obtained using a $T_{\rm eff}$ scale based on line depth ratios, and could therefore be modified when adopting temperatures derived from our flux ratios.

Finally, we note that the lack of SMC Cepheid spectra not only deprives us from a training sample at low metallicity, but also decreases the quality of the spectral phasing: with only a few consecutive spectra at hands, we cannot investigate possible period changes or cycle-to-cycle variations, which are to be expected since the current sample is biased towards brighter, longer-period Cepheids.

Extending our $T_{\rm eff}$ scale to low-metallicity Cepheids would require a large number of spectra of SMC Cepheids. Including such spectra in the training will certainly lead to a different (presumably smaller) set of (more) universal relations. A larger number of relations may however be recovered by using higher-order/more-dimensional models.

5.2. A new (spectroscopic) method for determining the reddening of Cepheids ?

Different parameters can limit the efficiency of the various flavors of IRSB methods in deriving accurate distances for Cepheids, namely the limb darkening correction (see Sabbey et al. 1995; Neilson & Lester 2013), the possible dependence of the projection factor on period (see Groenewegen 2013 and Kervella et al. 2017 for instance) or metallicity (discarded by e.g., Nardetto et al. 2011), the presence of a circumstellar envelope (e.g., Kervella et al. 2006; Hocdé et al. 2020), or the reddening.

If one assumes that reddening is the dominant uncertainty in the determination of $T_{\rm eff}$ as a by-product of IRSB methods, those Cepheids with minimal reddening (which we used in our training/testing sample) provide the absolute zero-point of the temperature scale. The other Cepheids should then align on the same scale. As a consequence, any systematic offset between the temperature derived (with unprecedented precision) using flux ratios and the temperature determined using the IRSB method could be attributed to uncertainties on the reddening. Conversely, it is possible to compute such a temperature offset and, in turn, to find out how the reddening would be modified for both $T_{\rm eff}$ scales to match. For the sake of the argument, we conducted such an exercise for the Cepheid δ Cep and managed to compensate for a systematic shift of 23 K between $T_{\rm eff, IRSB}$ and $T_{\rm eff, FR}$ by lowering $E(B - V)$ by −0.006 mag (0.069 instead of the initial 0.075 mag). However, modifying $E(B - V)$ does not result in a simple, global shift of $T_{\rm eff, IRSB}$: Fig. 13 shows the outcome of our δ Cep test case. It indicates that the difference in the IRSB

temperatures computed either with $E(B - V) = 0.075$ mag or with $E(B - V) = 0.069$ mag is phase-dependent. Incidentally, since this phase dependence is different from the one seen in Fig. 9, and since its magnitude is different by a factor of ≈20, it cannot smooth out the current residuals. It is beyond the scope of this paper to investigate this behavior further. These findings may lead to a better understanding of the current scatter of IRSB-based PL relations and/or the *p*-factor controversy.
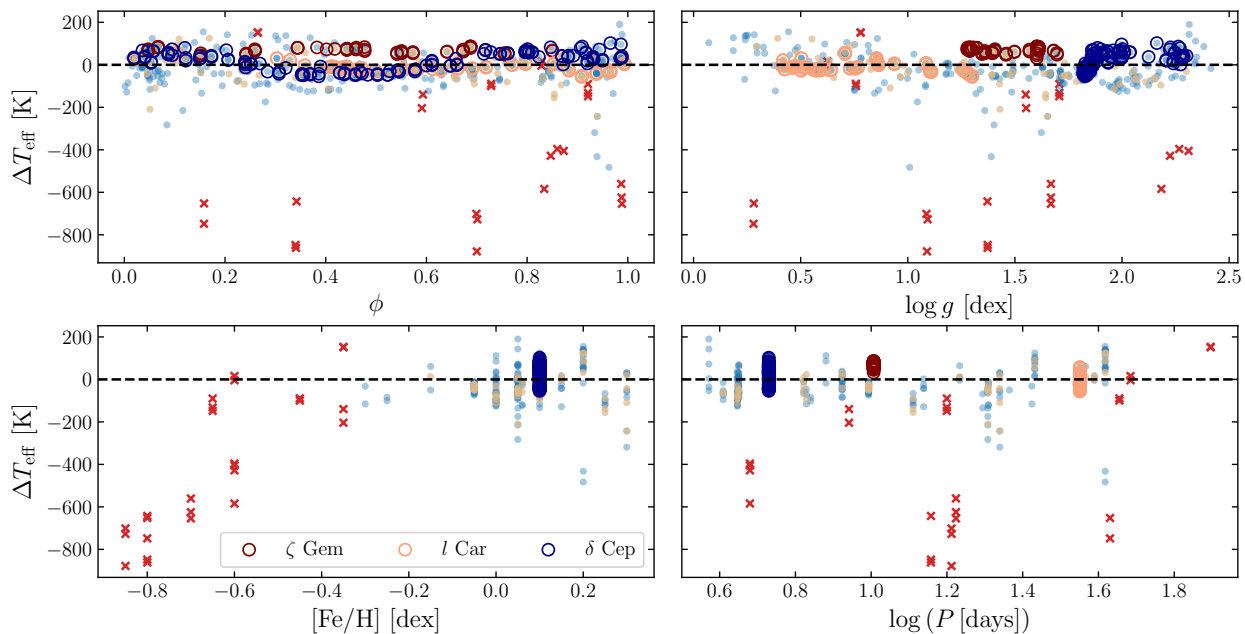
Fig. 12. Same as Fig. 6, where $T_{\rm eff}$ derived for low-metallicity SMC Cepheids are shown by red crosses.
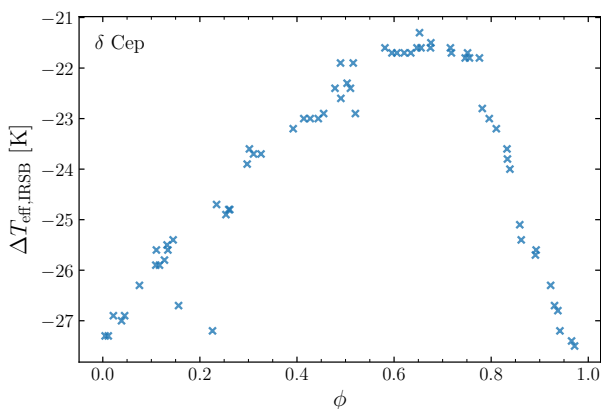

Fig. 13. Phase-dependent shift (modified − adopted) of the IRSB temperature of δ Cep when computed either with the modified $E(B-V) = 0.069$ mag or with $E(B-V) = 0.075$ mag as adopted throughout this work.

6. Conclusions

Using an innovative, data-driven method, we inferred a $T_{\rm eff}$ scale for classical Cepheids tying flux ratios to temperatures based on IRSB labels. For the first time in such studies, special emphasis was placed on accurately phasing the spectra when matching them to labels.

The next step to consider is to extend the $T_{\rm eff}$ scale towards lower, SMC-like, metallicities, which could not be done here simply because of the lack of spectra to build a training/testing sample. We also want to recover other atmospheric parameters like $\log g$ or [Fe/H] using a similar technique. On a more technical side, one could envision a more refined weighting scheme, involving all four quantities: precision, accuracy, ΔFR, $\Delta\lambda$.

Since it is purely spectroscopic, the method is independent from reddening to the same extent as the IRSB labels are, and could actually offer a new path to estimate reddening values for Cepheids. The derived temperatures are precise down to a few K, and we have full control of the accuracy, We estimate that the

accuracy remains better than 150 K at all phases, a value which will be greatly reduced once the uncertainties on the reddening scale, hence the $T_{\rm eff}/E(B-V)$ degeneracy will be lifted. The uncertainties on $E(B-V)$ obtained via the SPIPS algorithm are of the order of 0.025 mag (Breitfelder et al. 2016). Their use may then provide a different set of labels with reduced systematic uncertainties. On the other hand, since the number of Cepheids currently within reach of interferometric measurements is very limited, this would come at the expense of number statistics, and in particular period coverage. We desire not to train an inhomogeneous sample with different sources for the labels. Rescaling relations trained on a given set of labels would be the safer way to pursue, in case there is a constant or linear shift between IRSB and interferometric labels.

Our method to derive the temperature of classical Cepheids is very flexible as it works even at low S/N, and without requiring a normalization stage. Since it is always possible to downsample high-resolution spectra to $R = 5000$ (at which the method is trained), we ensure homogeneous temperatures for both low- and high-resolution spectra, a very interesting feature with respect to large spectroscopic surveys. For instance, the stellar, circumstellar, interstellar component (SCIP) survey with WEAVE, the Milky Way Disc and Bulge Low-Resolution (4MIDABLE-LR, Chiappini et al. 2019), the Milky Way Disc and Bulge High-Resolution (4MIDABLE-HR, Bensby et al. 2019) and the One Thousand and One Magellanic Fields (1001MC, Cioni et al. 2019) 4MOST surveys will all observe classical Cepheids at $R = 6500$ and/or $R = 20\,000$.

Our novel method paves the way for highly accurate and precise metallicity estimates, which will allow us to investigate the possible metallicity dependence of period-luminosity relations (e.g., Romaniello et al. 2008; Bono et al. 2010; Storm et al. 2011b) and ultimately solidify our measurement of the Hubble constant H_0.

References

Alonso, A., Arribas, S., & Martínez-Roger, C. 1999, A&AS, 140, 261
Amarsi, A. M., Nordlander, T., Barklem, P. S., et al. 2018, A&A, 615, A139
Anderson, R. I. 2016, MNRAS, 463, 1707
Anderson, R. I., Ekström, S., Georgy, C., et al. 2014, A&A, 564, A100
Anderson, R. I., Mérand, A., Kervella, P., et al. 2016, MNRAS, 455, 4231
Andrievsky, S. M., Martin, R. P., Kovtyukh, V. V., Korotin, S. A., & Lépine, J. R. D. 2016, MNRAS, 461, 4256
Bell, R. A. & Gustafsson, B. 1989, MNRAS, 236, 653
Bensby, T., Bergemann, M., Rybizki, J., et al. 2019, The Messenger, 175, 35
Berdnikov, L. N., Pastukhova, E. N., Kovtyukh, V. V., et al. 2019, Astronomy Letters, 45, 227
Bessell, M. S., Castelli, F., & Plez, B. 1998, A&A, 333, 231
Blackwell, D. E. & Shallis, M. J. 1977, MNRAS, 180, 177
Bono, G., Caputo, F., Marconi, M., & Musella, I. 2010, ApJ, 715, 277
Borgniet, S., Kervella, P., Nardetto, N., et al. 2019, A&A, 631, A37
Breitfelder, J., Mérand, A., Kervella, P., et al. 2016, A&A, 587, A117
Castelli, F. & Kurucz, R. L. 2003, in IAU Symposium, Vol. 210, Modelling of Stellar Atmospheres, ed. N. Piskunov, W. W. Weiss, & D. F. Gray, A20
Chiappini, C., Minchev, I., Starkenburg, E., et al. 2019, The Messenger, 175, 30
Cioni, M. . R. L., Storm, J., Bell, C. P. M., et al. 2019, The Messenger, 175, 54
Cosentino, R., Lovis, C., Pepe, F., et al. 2012, in Ground-based and Airborne Instrumentation for Astronomy IV, ed. I. S. McLean, S. K. Ramsay, & H. Takami (SPIE)
da Silva, R., Lemasle, B., Bono, G., et al. 2016, A&A, 586, A125
Dalton, G. 2016, Astronomical Society of the Pacific Conference Series, Vol. 507, WEAVE: The Next Generation Spectroscopy Facility for the WHT, ed. I. Skillen, M. Balcells, & S. Trager, 97
de Jong, R. S., Agertz, O., Berbel, A. A., et al. 2019, The Messenger, 175, 3
Dean, J. F., Warren, P. R., & Cousins, A. W. J. 1978, MNRAS, 183, 569
Dekker, H., D'Odorico, S., Kaufer, A., Delabre, B., & Kotzlowski, H. 2000, in Proc. SPIE, Vol. 4008, Optical and IR Telescope Instrumentation and Detectors, ed. M. Iye & A. F. Moorwood, 534–545
Eggen, O. J. 1996, AJ, 111, 1313
Evans, C., Puech, M., Afonso, J., et al. 2015, arXiv e-prints, arXiv:1501.04726
Fernie, J. D., Evans, N. R., Beattie, B., & Seager, S. 1995, Information Bulletin on Variable Stars, 4148
Fernley, J. A., Skillen, I., & Jameson, R. F. 1989, MNRAS, 237, 947
Fouqué, P., Arriagada, P., Storm, J., et al. 2007, A&A, 476, 73
Fouque, P. & Gieren, W. P. 1997, A&A, 320, 799
Fuhrmann, K. 1998, A&A, 338, 161
Fuhrmann, K., Axer, M., & Gehren, T. 1993, A&A, 271, 451
Genovali, K., Lemasle, B., Bono, G., et al. 2014, A&A, 566, A37
Genovali, K., Lemasle, B., Bono, G., et al. 2013, A&A, 554, A132
Genovali, K., Lemasle, B., da Silva, R., et al. 2015, A&A, 580, A17
Gieren, W., Storm, J., Konorski, P., et al. 2018, A&A, 620, A99
Górski, M., Zgirski, B., Pietrzyński, G., et al. 2020, ApJ, 889, 179
Gray, D. F. 1994, PASP, 106, 1248
Gray, D. F. & Johanson, H. L. 1991, PASP, 103, 439
Groenewegen, M. A. T. 2013, A&A, 550, A70
Hanke, M., Hansen, C. J., Koch, A., & Grebel, E. K. 2018, A&A, 619, A134
Hanke, M., Hansen, C. J., Ludwig, H.-G., et al. 2020, A&A, 635, A104
Haschke, R., Grebel, E. K., & Duffau, S. 2011, AJ, 141, 158
Hocdé, V., Nardetto, N., Lagadec, E., et al. 2020, A&A, 633, A47
Huang, Y., Liu, X. W., Yuan, H. B., et al. 2015, MNRAS, 454, 2863
Inno, L., Bono, G., Matsunaga, N., et al. 2016, ApJ, 832, 176
Jofré, P., Heiter, U., Soubiran, C., et al. 2014, A&A, 564, A133
Kaufer, A., Stahl, O., Tubbesing, S., et al. 1999, The Messenger, 95, 8
Kervella, P., Bersier, D., Mourard, D., et al. 2004a, A&A, 428, 587
Kervella, P., Fouqué, P., Storm, J., et al. 2004b, ApJ, 604, L113
Kervella, P., Mérand, A., Perrin, G., & Coudé du Foresto, V. 2006, A&A, 448, 623
Kervella, P., Thévenin, F., Folco, E. D., & Ségransan, D. 2004c, A&A, 426, 297
Kervella, P., Trahin, B., Bond, H. E., et al. 2017, A&A, 600, A127
Kiss, L. L. & Szatmáry, K. 1998, MNRAS, 300, 616
Kovtyukh, V. V. 2007, MNRAS, 378, 617
Kovtyukh, V. V. & Gorlova, N. I. 2000, A&A, 358, 587
Kurucz, R. L. 1975, in Multicolor Photometry and the Theoretical HR Diagram, 271
Kurucz, R. L. 1979, ApJS, 40, 1
Laney, C. D. & Caldwell, J. A. R. 2007, MNRAS, 377, 147
Lemasle, B., François, P., Bono, G., et al. 2007, A&A, 467, 283
Lemasle, B., François, P., Piersimoni, A., et al. 2008, A&A, 490, 613
Lemasle, B., Groenewegen, M. A. T., Grebel, E. K., et al. 2017, A&A, 608, A85
Luck, R. E., Andrievsky, S. M., Kovtyukh, V. V., Gieren, W., & Graczyk, D. 2011, AJ, 142, 51
Luck, R. E. & Lambert, D. L. 2011, AJ, 142, 136
Mancino, S., Romaniello, M., Anderson, R. I., & Kudritzki, R.-P. 2020, arXiv e-prints, arXiv:2001.05881
Mathias, P., Gillet, D., Fokin, A. B., et al. 2006, A&A, 457, 575
Mayor, M., Pepe, F., Queloz, D., et al. 2003, The Messenger, 114, 20
Mérand, A., Kervella, P., Breitfelder, J., et al. 2015, A&A, 584, A80
Nardetto, N., Fokin, A., Fouqué, P., et al. 2011, A&A, 534, L16
Nardetto, N., Groh, J. H., Kraus, S., Millour, F., & Gillet, D. 2008, A&A, 489, 1263
Neilson, H. R. & Lester, J. B. 2013, A&A, 554, A98
Pedicelli, S., Lemasle, B., Groenewegen, M., et al. 2010, A&A, 518, A11
Pel, J. W. 1978, A&A, 62, 75
Proxauf, B., da Silva, R., Kovtyukh, V. V., et al. 2018, A&A, 616, A82
Reichert, M., Hansen, C. J., Hanke, M., et al. 2020, arXiv e-prints, arXiv:2004.01195
Romaniello, M., Primas, F., Mottini, M., et al. 2008, A&A, 488, 731
Sabbey, C. N., Sasselov, D. D., Fieldus, M. S., et al. 1995, ApJ, 446, 250
Sartoretti, P., Katz, D., Cropper, M., et al. 2018, A&A, 616, A6
Storm, J., Carney, B. W., Gieren, W. P., et al. 2004, A&A, 415, 531
Storm, J., Gieren, W., Fouqué, P., et al. 2011a, A&A, 534, A94
Storm, J., Gieren, W., Fouqué, P., et al. 2011b, A&A, 534, A95
Strassmeier, K. G., Granzer, T., Weber, M., et al. 2004, Astronomische Nachrichten, 325, 527
Thévenin, F. & Idiart, T. P. 1999, ApJ, 521, 753
Turner, D. G. 2016, Rev. Mexicana Astron. Astrofis., 52, 223
Vasilyev, V., Amarsi, A. M., Ludwig, H.-G., & Lemasle, B. 2019, A&A, 624, A85
Vasilyev, V., Ludwig, H.-G., Freytag, B., Lemasle, B., & Marconi, M. 2017, A&A, 606, A140
Vasilyev, V., Ludwig, H.-G., Freytag, B., Lemasle, B., & Marconi, M. 2018, A&A, 611, A19
Weber, M., Granzer, T., & Strassmeier, K. G. 2012, in Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series, Vol. 8451, Proc. SPIE, 84510K
Welch, D. L. 1994, AJ, 108, 1421
Yong, D., Carney, B. W., & Friel, E. D. 2012, AJ, 144, 95