Application of ILIUM to the estimation of the $T_{\rm eff}$ – [Fe/H] pair from BP/RP

prepared by: Coryn A.L. Bailer-Jones
Max Planck Institute for Astronomy, Heidelberg
Email: calj@mpia.de

approved by:
reference: GAIA-C8-TN-MPIA-CBJ-043
issue: 1
revision: 1
date: 2009-02-10
status: Issued

Abstract

A new parameter estimation method ILIUM was introduced in GAIA-C8-TN-MPIA-CBJ-042 where it was demonstrated on the problem of estimating $\mathrm{T_{eff}}$ and $\log g$ from noisy BP/RP spectra. The current implementation is limited to two APs (one "strong" and one "weak"). Here I present results of estimating $\mathrm{T_{eff}}$ and [Fe/H] from the same data. For F,G,K dwarfs ($4000 \leq \mathrm{T_{eff}} \leq 7000$ K) with metallicities ranging from $+1$ to -4 dex, we can estimate [Fe/H] to an accuracy of 0.14 dex and $\mathrm{T_{eff}}$ to 0.4% (mean absolute errors) at G=15, to 0.26 dex and 0.6% respectively at G=18.5 and to 0.82 dex and 1.6% at G=20 (data with an "end-of-mission" SNR corresponding to 72 transits, but non-oversampled spectra). The errors for giants are 50% larger at G=15 but only 10% larger at the fainter magnitudes. Surprisingly, the performance is hardly improved when stars with [Fe/H] < -2.0 are removed from the analysis. If ILIUM is applied to stars with unknown $\log g$ (having been trained on the full range of $\log g$), then the performance at G=18.5 is 0.40 dex in [Fe/H] and 1.2% in $\mathrm{T_{eff}}$. (Given that dwarfs heavily outnumber giants in a magnitude-limited sample, better overall results would be obtained if we trained on dwarfs.) From this I show that all three APs ($\mathrm{T_{eff}}$, [Fe/H] and $\log g$) can be estimated by successively applying the two 2-AP versions of ILIUM.

Contents

1 Introduction

Estimation of the pair $\mathrm{T_{eff}}$ and [Fe/H] is a realistic two parameter problem, because either (a) we can assume that most stars in a magnitude limited survey are dwarfs, or (b) with Gaia we can use the parallaxes (via the derived absolute magnitude) along with a Gaia colour to independently estimate $\log g$. I therefore apply ILIUM to estimate $\mathrm{T_{eff}}$ and [Fe/H] separately for dwarfs and giants. The dwarf sample is defined as having $\log g$ equal to 4.0, 4.5 or 5.0 dex. This comprises 1716 such stars and is randomly split into equal sized train and test sets. (Recall that the training data are used to fit the forward model and do the nearest neighbour initialization.) The giant sample is defined as objects with $\log g$ equal to 1.0, 1.5, 2.0, 2.5 or 3.0 (1882 stars). The AP distribution for the dwarfs is shown in Fig. 1 (the grid for the giants is almost identical). See Fig. 5 of GAIA-C8-MPIA-CBJ-042 for the $\mathrm{T_{eff}}$–$\log g$ distribution. The spectral data are exactly as in CBJ-042 (Sordo & Vallenari 2008), that is, they have a SNR corresponding to a stack of 72 transits, yet with the origial wavelength dispersion, i.e. non-oversampled spectra. Oversampling should improve the spectral resolution which may improve performance above that reported here. On the other hand, the spectral combination and oversampling procedure may introduce additional errors not yet accounted for in the simulations (although GOG does currently include some additional error sources beyond the usual triad of source, background and CCD readout; Zaldua et al. 2008). To get an idea of the quality of the spectra, Fig. 2 plots the median and 10% and 90% quartiles of the SNR at each wavelength at G=18.5 and G=20.0. (Compared to the G=18.5 curve, the SNR at G=15 is 8–6 times larger between 400 and 650 nm and 7–12 times larger between 650 and 1000 nm.) ILIUM is used in its default mode with the internal parameters exactly as shown in Table 2 of CBJ-042. That is, the same values of

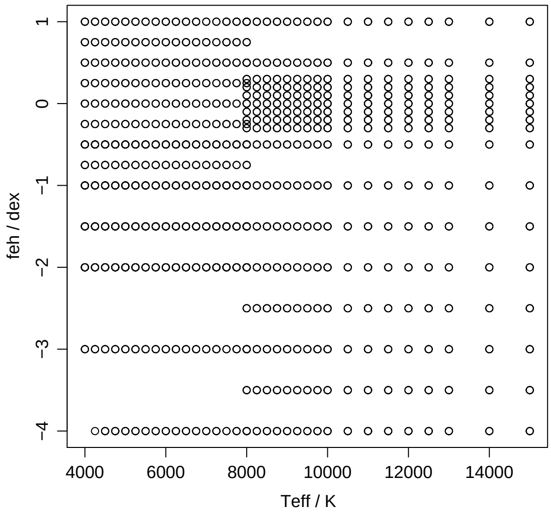

FIGURE 1: The AP grid for the dwarfs ($\log g \in \{4.0, 4.5, 5.0\}$)

parameters used there for the (standardized) $\log g$ measures are used here for the (standardized) [Fe/H] measures. Performance is reported using statistics of the AP residuals: the RMS, σ_ϕ; the mean absolute residual, $\overline{|\delta\phi|}$; the mean residual, $\overline{\delta\phi}$ (a measure of the systematic error).

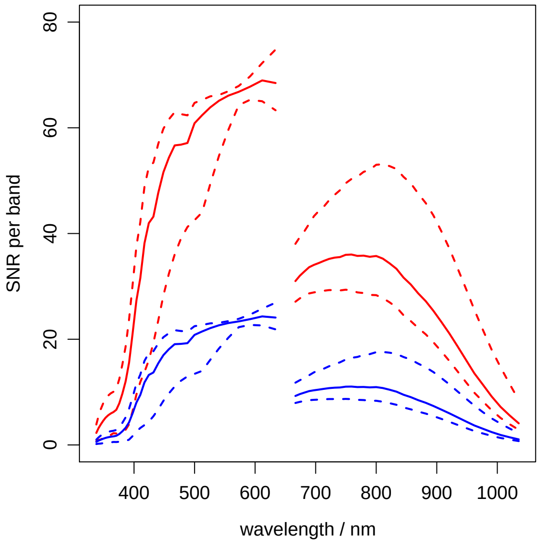

FIGURE 2: The median SNR (solid line) and 0.1 and 0.9 quartiles (dashed lines) across the dwarf sample for G=18.5 (red) and G=20.0 (blue)

2 Application to dwarfs

The forward model fits are shown in Figs. 3 and 4. The scatter in the plot against [Fe/H] is due to the $\log g$ variations. The fits are as good as we could expect.

2.1 G=15

The general pattern of the iterative updates is similar to those seen in Fig. 8 of CBJ-042, so is not shown. The spectra of AP updates are interesting because they allow us to see which spectral bands contribute to the APs for which stars (and how these evolve over the iterations), but they are too numerous to include here.

The residuals are shown in Fig. 5. The summary statistics are

	[Fe/H]	$\log{(T_{\rm eff})}$
$\overline{\delta\phi}$	-0.15	$5.8e^{-4}$
$\overline{\lvert\delta\phi\rvert}$	0.68	0.0058
σ_ϕ	1.25	0.0087

ILIUM, dwarfs, G=15, full AP range

The overall metallicity performance is poor, because the sample includes many hot stars, and it

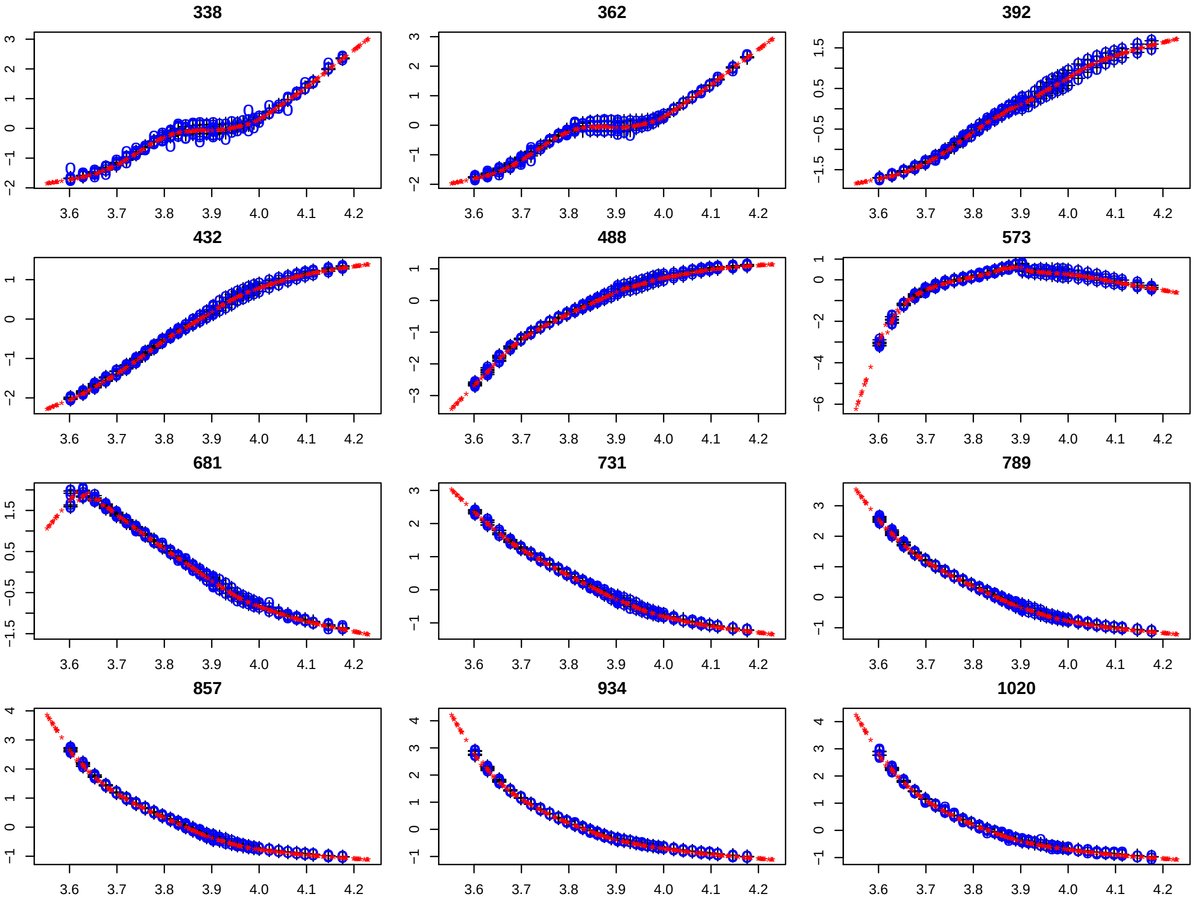

FIGURE 3: Predictions of the full forward model for the dwarfs as a function of $\log(\mathrm{T_{eff}})$ at constant [Fe/H]=-1.0 in 12 different bands (with wavelength in nm at the top of each panel). The black crosses (barely visible) are the (noise-free) grid points, the red stars are the forward model predictions (at randomly selected AP values) and the blue circles the noisy (G=15) grid points. The flux plotted on the ordinate is in standardized units.

is well known that [Fe/H] cannot be accurately estimated for hotter stars. This is also responsible for the systematic overestimation of [Fe/H] at low metallicity (the trend in the middle left panel). There is also a strong dependence of both the [Fe/H] and $\mathrm{T_{eff}}$ accuracy on $\mathrm{T_{eff}}$, with both being worse for hotter stars (middle and bottom right panels). This is due to the metallicity spread, because we saw no such trend for constant metallicity. (Indeed, we see the opposite effect, namely lower $\mathrm{T_{eff}}$ error for hot stars: see Fig. 11 of CBJ-042). This strong dependence of the results on AP renders the above summary statistics (which averages over a more or less uniform sampling in $\log(\mathrm{T_{eff}})$ up to 14 000 K) rather meaningless.

For this reason we replot the residuals and recalculate[1] the statistics removing stars with true

[1] In this and all following examples we only remove objects from the analysis. ILIUM is not retrained, so it can

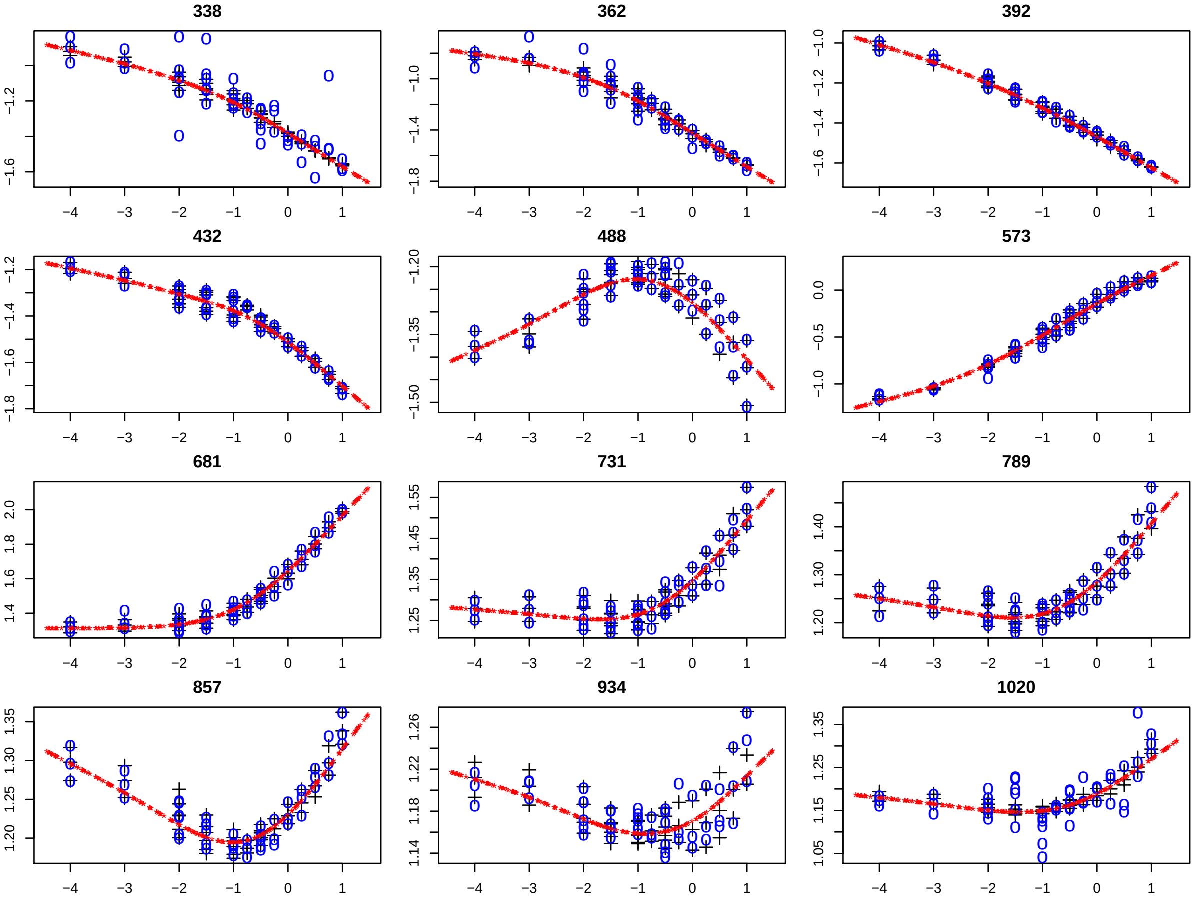

FIGURE 4: As Fig. 3, but now showing predictions of the full forward model as a function of [Fe/H] at constant $\mathrm{T_{eff}}$=5000 K.

$\mathrm{T_{eff}} > 7000$ K

	[Fe/H]	$\log(\mathrm{T_{eff}})$
$\overline{\delta\phi}$	5.6e^{-3}	5.2e^{-5}
$\overline{\lvert\delta\phi\rvert}$	0.14	0.0017
σ_ϕ	0.24	0.002

ILIUM, dwarfs, G=15, $\mathrm{T_{eff}} \leq 7000$ K

The residuals are plotted in Fig. 6. The results are now dramatically different: the residuals for [Fe/H] have dropped by a factor of 5 and those for $\log(\mathrm{T_{eff}})$ by a factor of 3. (Note that the error in $\log(\mathrm{T_{eff}})$ of 0.0017 corresponds to an error in $\mathrm{T_{eff}}$ of 0.4% – multiply by 2.3.) We can still estimate [Fe/H] to an accuracy of better than 0.5 dex even at [Fe/H]=−4.0. (However, from the middle left panel we do see a slight tendency to overestimate the metallicity for [Fe/H]=−3.0 and −4.0.) The systematic in [Fe/H] residual at low [Fe/H] has now vanished, confirming that it is a problem only for the hot stars.

still produce APs spanning the whole grid plus/minus the permitted 10% extrapolation.

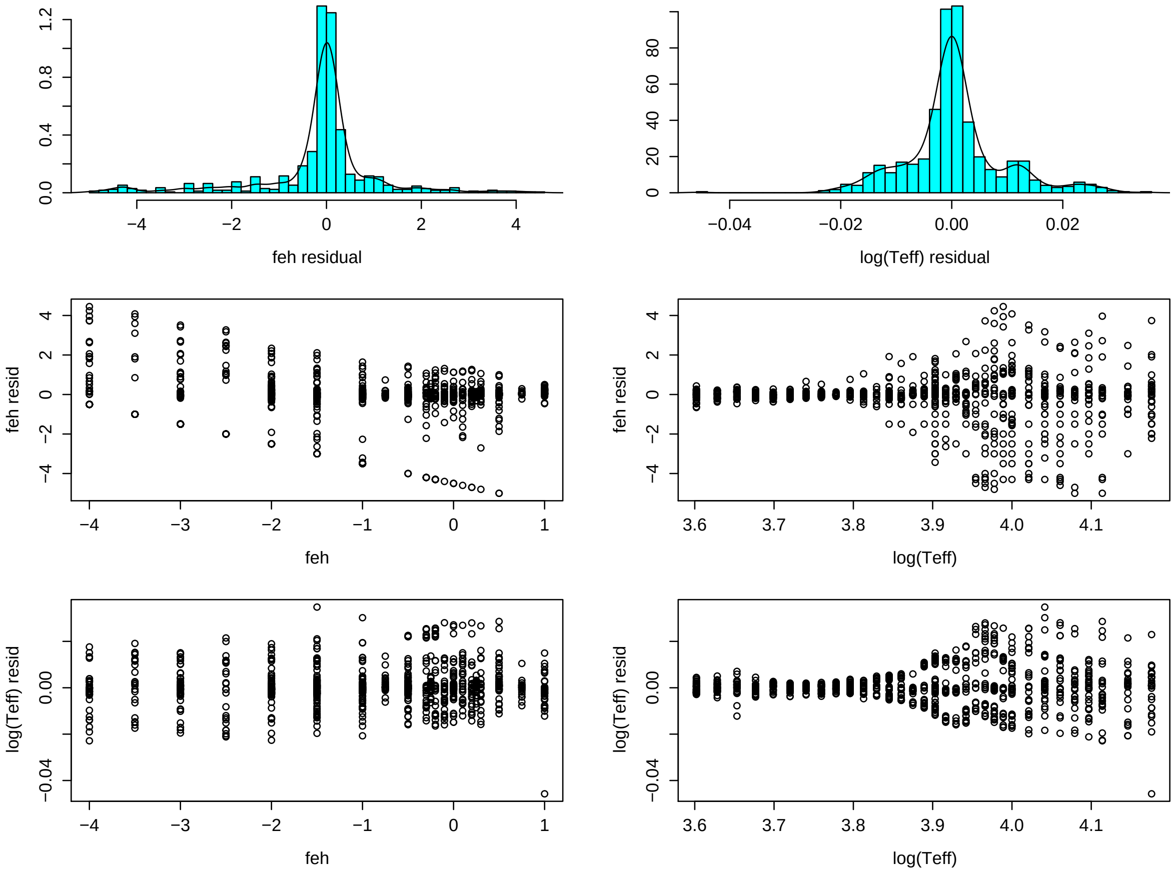

FIGURE 5: AP residuals for the dwarfs at G=15, plotted as a function of the true APs, for the full range of $T_{\rm eff}$ and [Fe/H] shown in Fig. 1

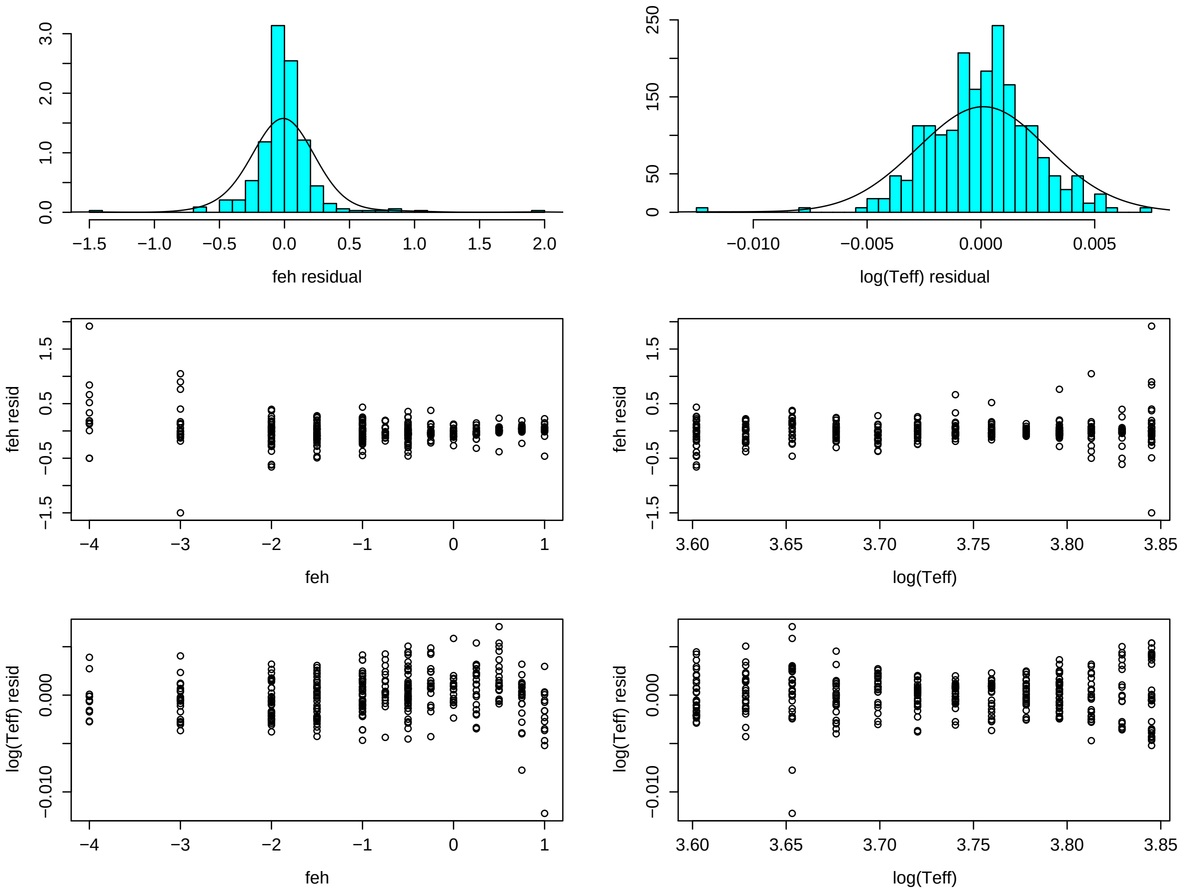

FIGURE 6: AP residuals for the dwarfs at G=15 shown just for stars with true $T_{eff} \leq 7000$ K, plotted as a function of the true APs

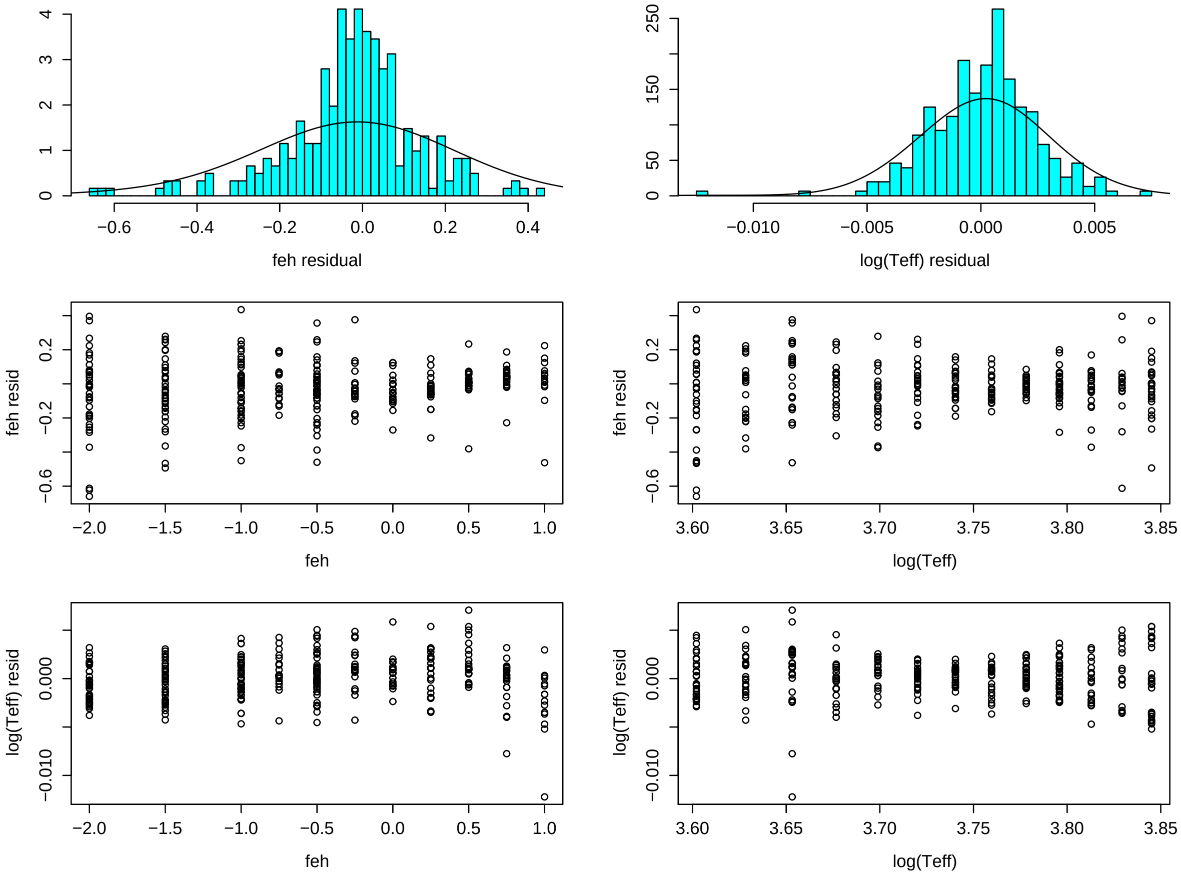

FIGURE 7: AP residuals for the dwarfs at G=15 shown just for stars with true $T_{eff} \leq 7000\,K$ and [Fe/H]≥ -2.0, plotted as a function of the true APs

As [Fe/H] decreases, its signature in the spectra weakens and in principle is harder to detect and estimate. We might therefore expect that removing the most metal poor stars (which are anyway very rare in the Galaxy which Gaia will observe) improves the results. Removing also stars with true [Fe/H]$<\ 2.0$ dex (this removes 20% of the objects from the cool star sample) yields residuals as shown in Fig. 7 and

	[Fe/H]	$\log(\mathrm{T_{eff}})$
$\overline{\delta\phi}$	$5.6\mathrm{e}^{-3}$	$5.2\mathrm{e}^{-5}$
$\overline{\lvert\delta\phi\rvert}$	0.11	0.0018
σ_ϕ	0.16	0.002

ILIUM, dwarfs, G=15, $\mathrm{T_{eff}}\leq 7000$ K and [Fe/H]≥ -2.0 dex

The errors have hardly decreased, which tells us not only that ILIUM can estimate [Fe/H] equally well across the metallicity range (something we could anyway see in Fig. 6), but also that the $\mathrm{T_{eff}}$ accuracy is not affected by metallicity. Curiously, the systematic error in [Fe/H] has increased a bit, but this may not be significant (it is still four time smaller than the mean absolute error). These are now realistic summary statistics for Gaia, as even if we didn't know $\log g$ from the astrometry, the majority of objects are dwarfs which are not very metal poor.

If we remove only the low metallicity stars but retain the hot stars, then the results hardly improve with respect to the original (all APs) case. This confirms that it is the removal of hot stars which is crucial for estimating metallicity, and this because they retain hardly any metallicity signature in their BP/RP spectra. Metal poor stars, in contrast, retain a metallicity signature which we can still detect to an accuracy of 0.5 dex or better down to [Fe/H]=-4.0

I remind the reader that ILIUM is allowed to estimate AP values which extend beyond the training grid, by 10% of the AP range in each direction (the default setting). Even though all the test data have true APs within the grid limits, we see that ILIUM does assign a few values beyond this, e.g. the two -4.5 dex stars we can identify in Fig. 6.

So far we have made cuts on the true APs in order to predict performance on populations of certain types of stars. In the real application, we would need to make cuts based on the estimated APs. If we do this and recalculate the statistics for the cool star sample we get

	[Fe/H]	$\log(\mathrm{T_{eff}})$
$\overline{\delta\phi}$	-0.016	$-8.0\mathrm{e}^{-5}$
$\overline{\lvert\delta\phi\rvert}$	0.13	0.0017
σ_ϕ	0.21	0.002

ILIUM, dwarfs, G=15, *estimated* $\mathrm{T_{eff}}\leq 7000$ K

which is no worse than when making cuts on the real APs. Of course, to measure residuals and performance statistics like this we would need some independent "truth", but it is still interesting to plot the residuals agaist the estimated APs, as in Fig. 8. The interpretation of the plots is left to the reader as an exercise $\ddot{\smile}$.

ILIUM estimated the AP uncertainties for each star. Returning to the results for the full AP range, Fig. 9 shows the ratio of the AP uncertainties to the absolute value of the true residuals. The error predictions and their distribution are reasonable, although there is a tendency to

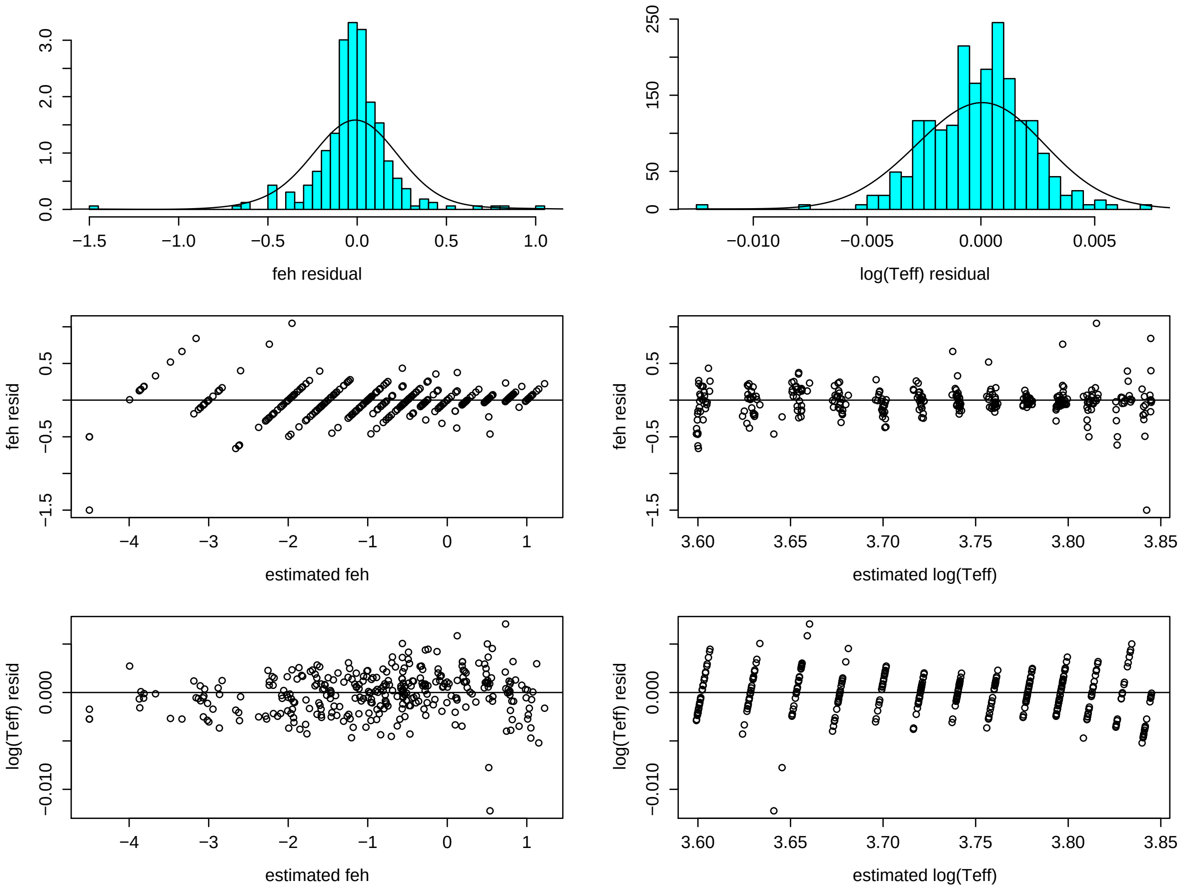

FIGURE 8: AP residuals for the dwarfs at G=15 shown just for stars with *estimated* $T_{\rm eff} \leq$ 7000 K, plotted as a function of the *estimated* APs

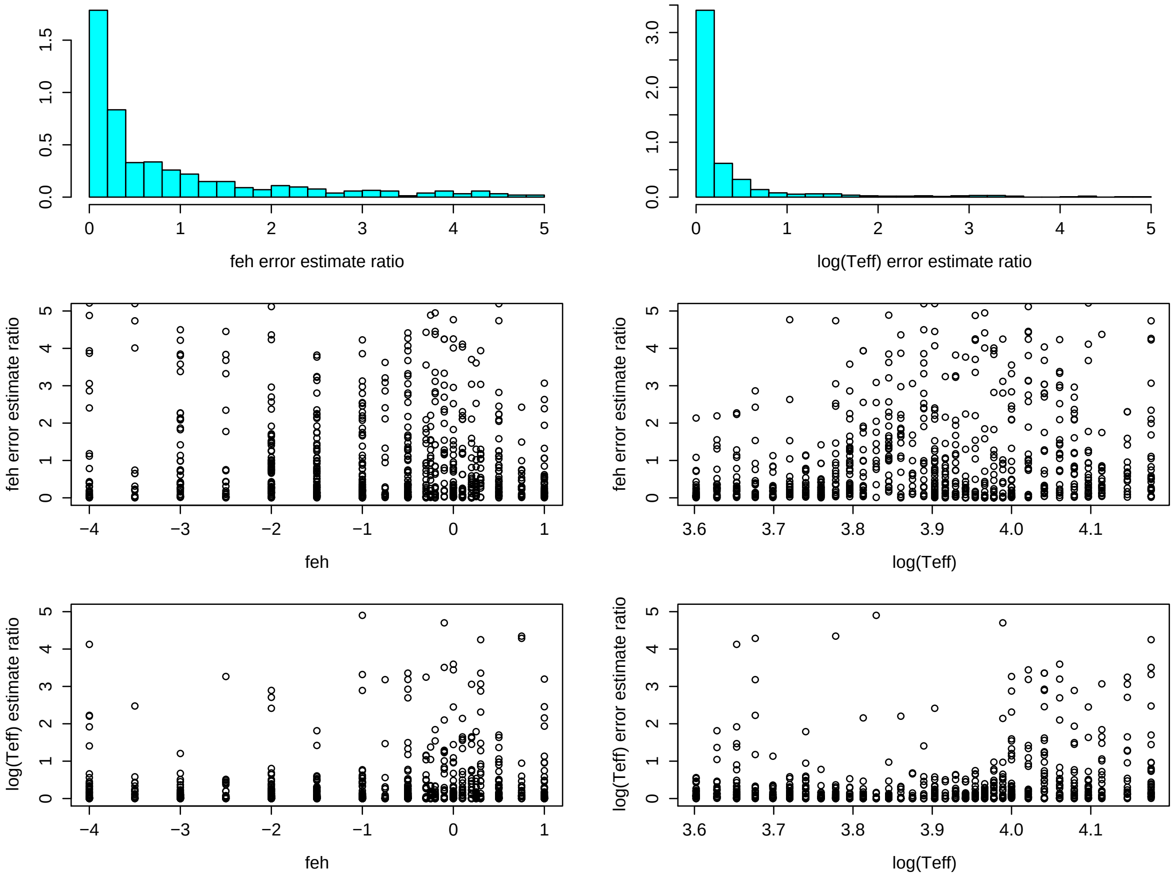

FIGURE 9: Estimated AP uncertainties for the dwarfs expressed as a ratio of the absolute value of the true residuals, for G=15

underestimate the errors, especially for $\log(\mathrm{T_{eff}})$.

2.2 G=18.5

Applying ILIUM to the same data at G=18.5, we again see poor statistics when averaging over the full range of $\mathrm{T_{eff}}$ and [Fe/H], so limiting to cool stars we get

	[Fe/H]	$\log(\mathrm{T_{eff}})$
$\overline{\delta\phi}$	-0.037	$-4.5\mathrm{e}^{-6}$
$\overline{\lvert\delta\phi\rvert}$	0.26	0.0024
σ_ϕ	0.42	0.0033

ILIUM, dwarfs, G=18.5, $\mathrm{T_{eff}} \leq 7000\,\mathrm{K}$

Even 3.5 magnitudes fainter than G=15, we still get very reasonable results, with little trend in the accuracy with $\mathrm{T_{eff}}$ or [Fe/H]. That the performance degrades little is not surprising, however, when we consider that the SNR per band is still over 20 for most of the spectrum (see Fig. 2).

2.3 G=20

We now apply ILIUM to stars at Gaia's magnitude limit. We do not necessarily expect the best science to come out these objects – the median SNR per band is just 10 – but as there will be so many of them it is important to assess how well we can estimate their APs. Furthermore the performance on G=20 end-of-mission data is roughly what we expect for a single transit spectrum on G=17.7 stars, a scaling which assumes that the noise is dominated by source noise. (That is, if the source delivers F photons per transit over N transits, I am assuming SNR $\propto \sqrt{FN}$, which can be contrasted with the case when background/readout noise dominate, in which case SNR $\propto F\sqrt{N}$.) In practice the results would actually apply to slightly brighter stars, because the flux limit for a given SNR scales more rapidly with the number of transits than $N^{-0.5}$, due to the source-independent noise terms and because additional noise will effectively be introduced by the spectral combination.

The summary statistics for cool stars at G=20 are

	[Fe/H]	$\log(\mathrm{T_{eff}})$
$\overline{\delta\phi}$	-0.033	$3.6\mathrm{e}^{-4}$
$\overline{\lvert\delta\phi\rvert}$	0.82	0.0070
σ_ϕ	1.14	0.009

ILIUM, dwarfs, G=20.0, $\mathrm{T_{eff}} \leq 7000\,\mathrm{K}$

and the residuals are plotted in Fig. 10. Over the whole metallicity range $\mathrm{T_{eff}}$ accuracy is still very good at 1.5%. Removing in addition the hot stars hardly improves this, decreasing it by about 7%.

As expected, metallicity performance is much worse than at G=18.5, although we can still distinguish metal poor stars ([Fe/H] < -2.5 dex) from solar metallicity ones at three times the mean absolute error (or at "2 sigma" if we use the RMS). However, the most metal poor stars suffer from systematic errors: stars with [Fe/H]$=-4.0$ have a systematic metallicity error of

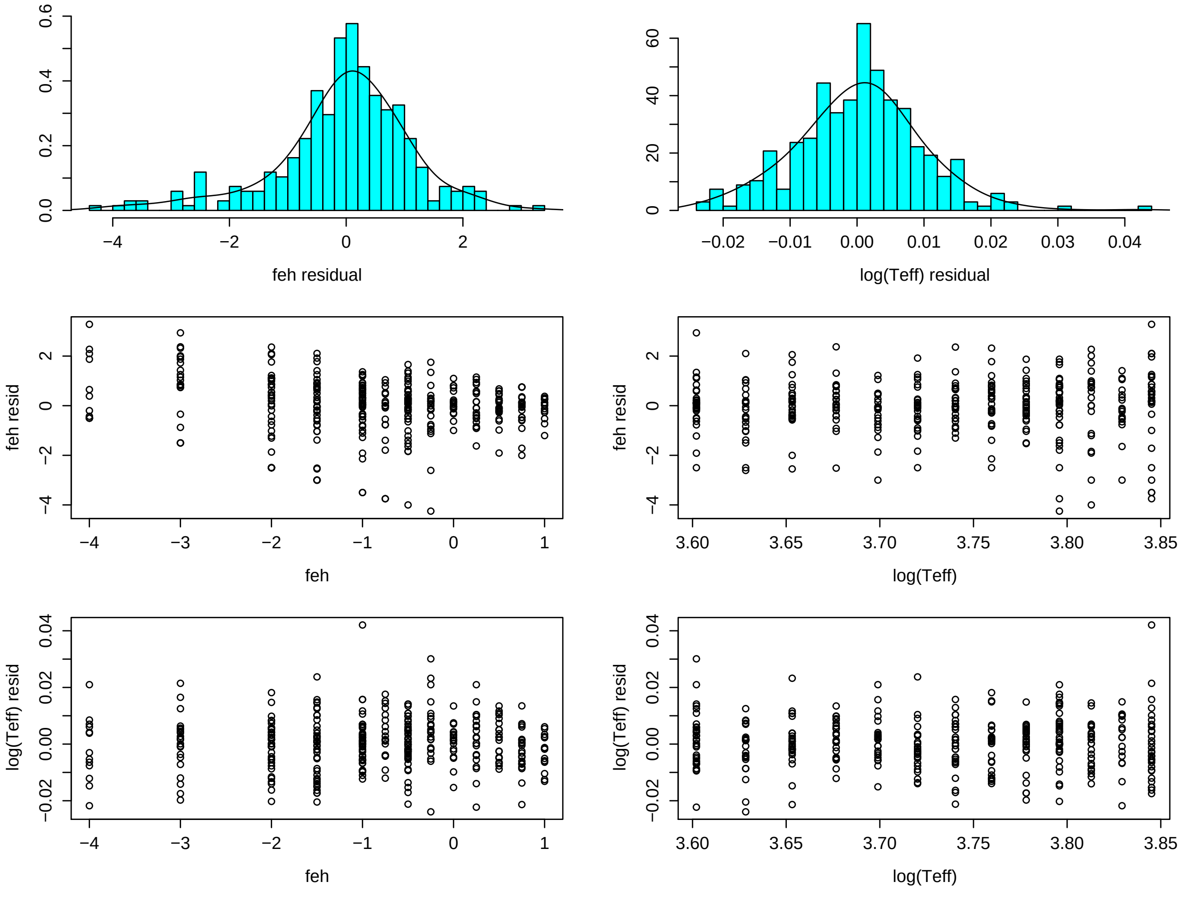

FIGURE 10: AP residuals for the dwarfs at G=20 shown just for stars with true $T_{\rm eff} \leq 7000$ K, plotted as a function of the true APs

$+0.9$ dex (overestimated), and a standard deviation about this of 1.3 dex. At [Fe/H]=-3.0 the systematic is $+0.6$ dex with a standard deviation about this of 1.3 dex. We might think that we could correct for these systematics, but it turns out that we cannot (see Appendix A).

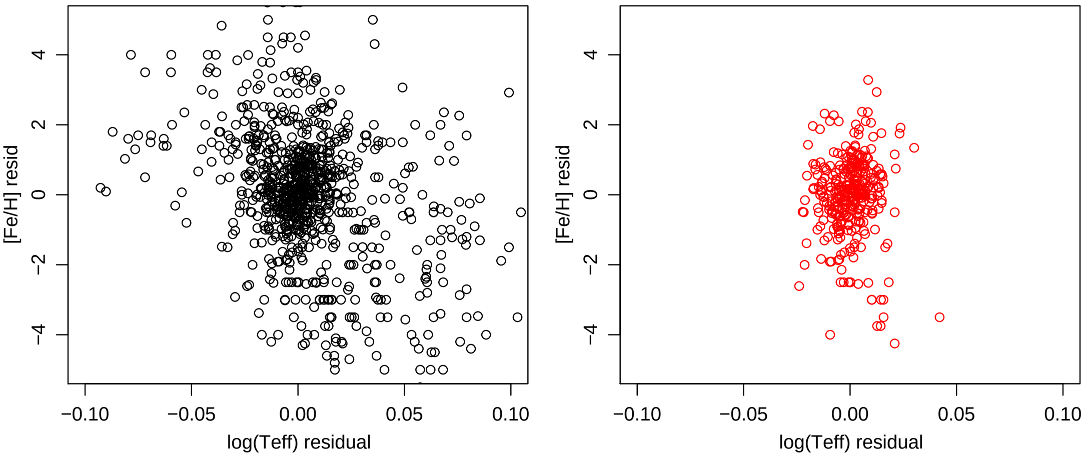

FIGURE 11: Correlation between the residuals on the dwarfs at G=20 for the full sample (left) and for the cool stars ($T_{\rm eff} \leq 7000$ K) (right). The Pearson correlation coefficients are -0.37 (left) and -0.003 (right)

Fig. 11 plots the correlation between the residuals. On the full sample there is a small but significant anticorrelation. This is presumably related to the systematic errors introduced by the hot stars, because there is no significant correlation once we remove these from the analysis (right panel).

It is almost as important to have a measure of uncertainty in an AP estimate as it is to have the AP estimate itself, as only then do we know whether (and to what degree) we can trust the estimate. Statistics based on test sets (i.e. those shown in the table above) are important, but it is desirable to have object-specific uncertainty estimates which take the actual measurement into account. ILIUM can do this, as was decsribed in CBJ-042. These error predictions are shown in Fig. 12, plotted as a ratio over the true residuals for each object. The distribution is "better" than we saw for G=15: it extends over a larger range of values and is not skewed towards frequent underestimation.

In addition to error estimates, we need to know whether a presented unlabelled object fits into the domain of the classifier's training grid. We assess this via a Goodness-of-Fit (GoF) measure between the observed spectrum and the spectrum which ILIUM predicts. ILIUM currently measures this via the reduced χ^2 for this, whereby a value of 1 is expected for a good fit. The distribution is shown in Fig. 13 and has a mean of 0.97 (median of 0.45).

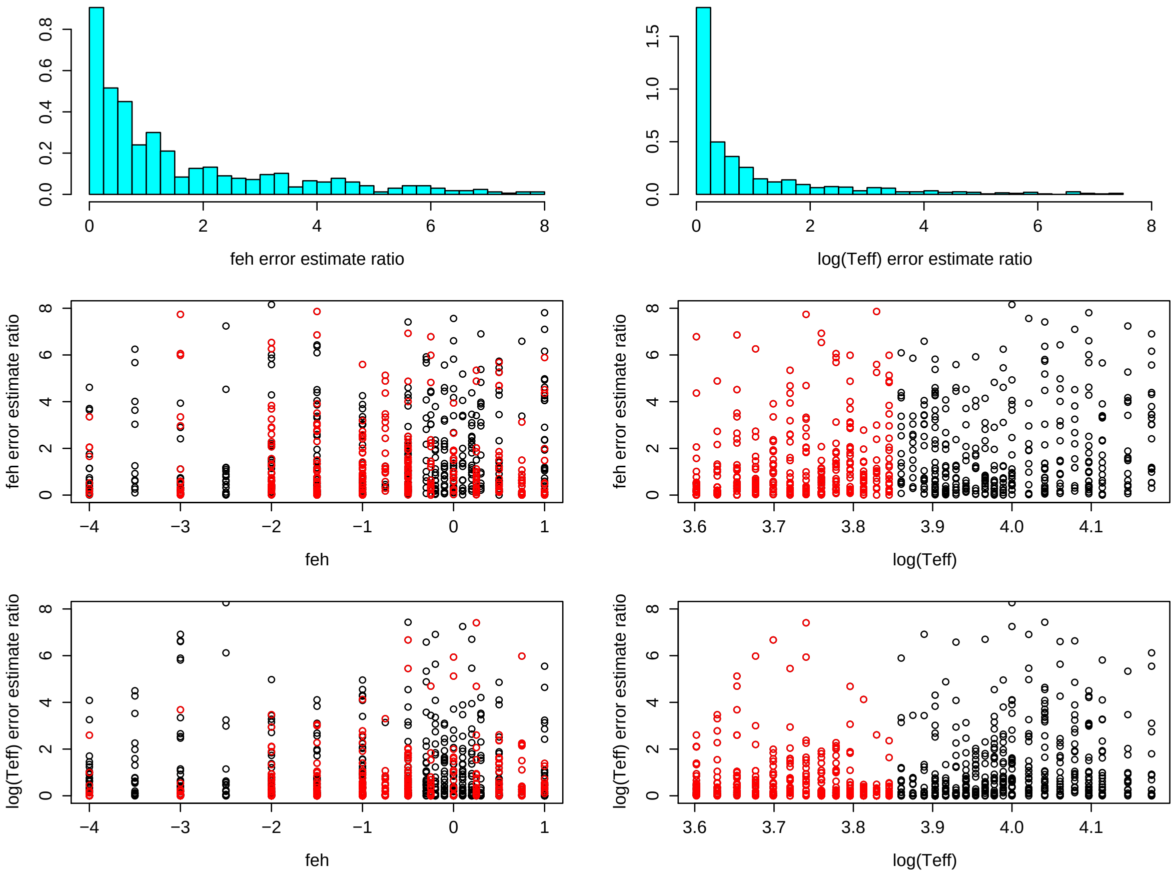

FIGURE 12: Estimated AP uncertainties for the dwarfs expressed as a ratio of the absolute value of the true residuals, for G=20, for the full range of APs. The red points are for objects with $T_{\rm eff} \leq 7000$ K: of these 327 objects, 24 (7%) have error estimate ratios greater than 8.

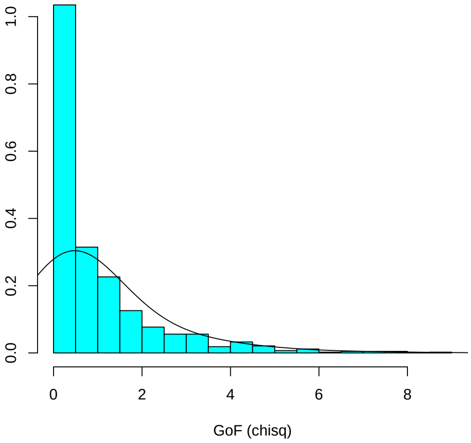

FIGURE 13: Distribution of the goodness-of-fit (reduced χ^2) values for the dwarf sample at G=20

3 Application to giants

The forward model fits for the giants are shown in Figs. 14 and 15. Recall that the forward models are fit to noise-free data (the black crosses in the figures). To get an idea of how much the noisy data deviates from them, I overplot G=20 data as blue points in Figs. 14 and 15. This can be compared to the blue points in the dwarf forward model plots, which were for G=15 data.

The fits (red points) against $\mathrm{T_{eff}}$ are quite similar to what we found for the dwarfs (Fig. 3) , which is not surprising as $\mathrm{T_{eff}}$ is a strong parameter and so the different $\log g$ selection has little impact. (In both cases [Fe/H] was held constant at -1.0 dex). At first glance the [Fe/H] fit looks very different from the dwarf case (Fig. 4), but this is mostly because of the different scales on the ordinate. Yet there are differences, indicating that the metallicity dependence of the flux depends on the surface gravity.

At G=15 the performance on the test set for the full $\mathrm{T_{eff}}$ and [Fe/H] range can be summarized as

	[Fe/H]	$\log(\mathrm{T_{eff}})$
$\overline{\delta\phi}$	−0.089	$-1.0\mathrm{e}^{-3}$
$\overline{\lvert\delta\phi\rvert}$	0.62	0.0048
σ_ϕ	1.03	0.0072

ILIUM, giants, G=15, full AP range

This is very similar to what we found with the dwarfs. We likewise see here that we get better performance on the cooler stars

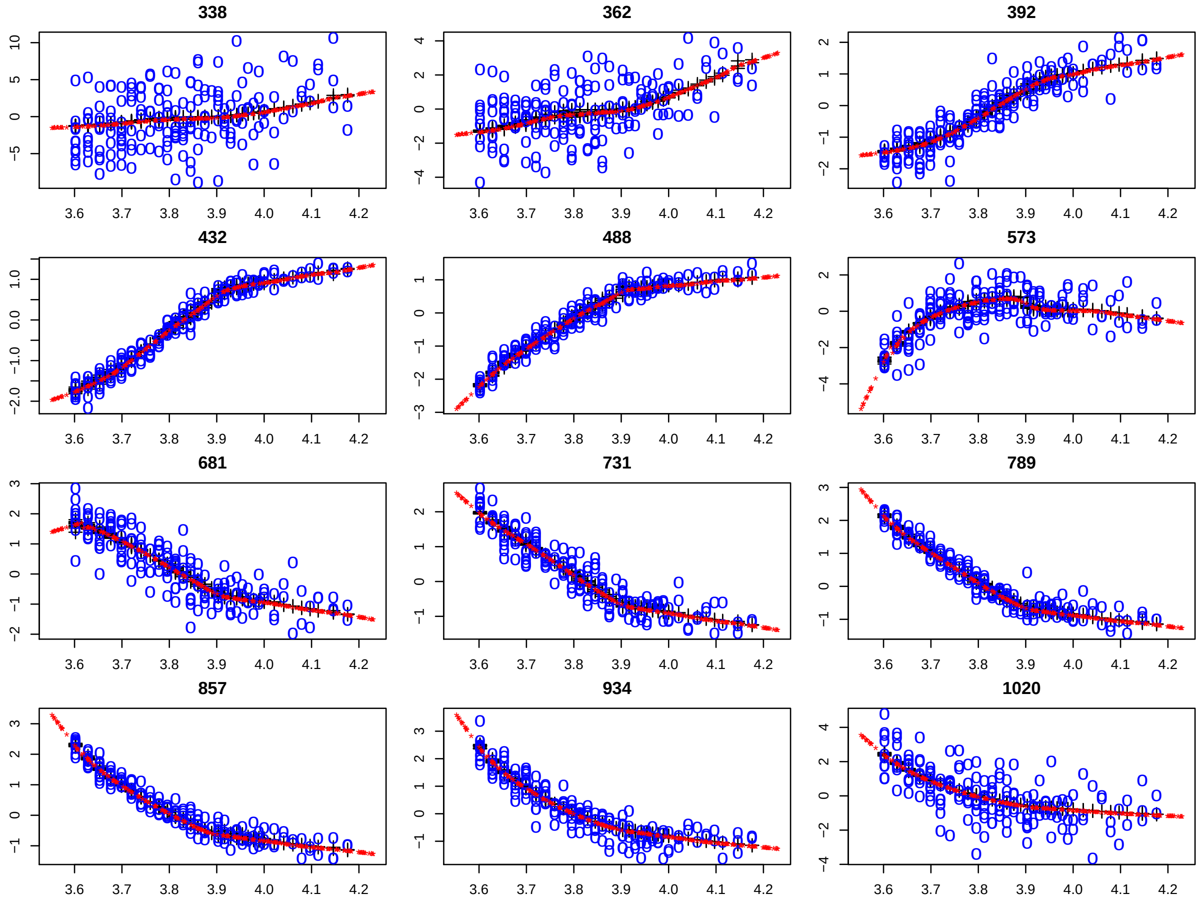

FIGURE 14: Predictions of the full forward model for the dwarfs as a function of $\log(T_{\rm eff})$ at constant [Fe/H]=-1.0 in 12 different bands (with wavelength in nm at the top of each panel). The black crosses are the (noise-free) grid points, the red stars are the forward model predictions (at randomly selected AP values) and the blue circles the noisy (G=20) grid points. The flux plotted on the ordinate is in standardized units.

	[Fe/H]	$\log(T_{\rm eff})$
$\overline{\delta\phi}$	-0.01	2.3e^{-4}
$\overline{\lvert\delta\phi\rvert}$	0.22	0.0028
σ_ϕ	0.34	0.0037

ILIUM, giants, G=15, $T_{\rm eff} \leq 7000$ K

The systematic in [Fe/H] is again reduced and no longer significant. Both [Fe/H] and $T_{\rm eff}$ can be estimated accurately for giants at this magnitude, although the errors are about 50% larger than could be achieved with dwarfs.

The performance at G=18.5 and G=20 is as follows

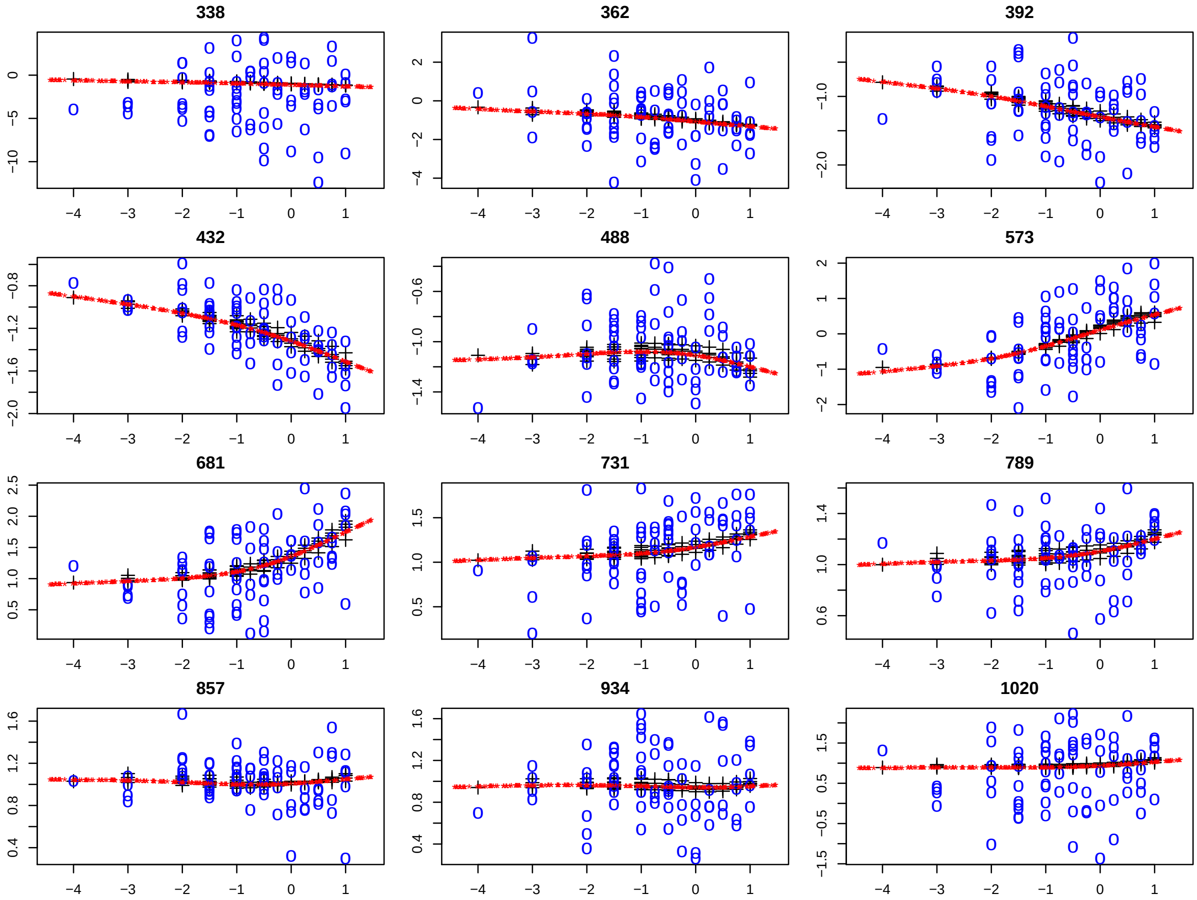

FIGURE 15: As Fig. 14, but now showing predictions of the full forward model as a function of [Fe/H] at constant T_{eff}=5000 K.

	[Fe/H]	log (T_{eff})
$\overline{\delta\phi}$	−0.03	3.4e^{-4}
$\overline{\lvert\delta\phi\rvert}$	0.31	0.0035
σ_ϕ	0.50	0.0045

ILIUM, giants, G=18.5, $T_{eff} \leq$ 7000 K

and

	[Fe/H]	log (T_{eff})
$\overline{\delta\phi}$	−0.11	1.1e^{-3}
$\overline{\lvert\delta\phi\rvert}$	0.74	0.0073
σ_ϕ	1.08	0.0092

ILIUM, giants, G=20, $T_{eff} \leq$ 7000 K

With respect to G=15, the [Fe/H] errors are 1.4 and 3.4 times higher at G=18.5 and G=20 respectively. T_{eff} can be estimated to 0.8% and 1.7% respectively (1.3 and 2.6 times higher than at G=15). In all three magnitude cases, removing the metal poor stars ([Fe/H] < -2.0) barely improves the results over what we see in the above tables, the largest reduction being of

0.05 dex in $\overline{|\delta\phi|}$ in [Fe/H] at G=20, just as we saw for the dwarfs.

4 Application to stars with unknown log g

So far we have trained ILIUM using stars with a restricted $\log g$ range and applied it to stars with the same $\log g$ range. However, there is no reason why we have to do this. For example, if we had a magnitude limited sample and were unable to estimate $\log g$, then we might decide to use the ILIUM model trained on dwarfs, because the majority of stars are dwarfs. This would presumably introduce larger errors on the giants. Alternatively, we might train ILIUM on a uniform sampling in $\log g$. This we do here, fitting ILIUM on the full range of $\log g$, from -0.5 to 5.0 in steps of 0.5 (see Fig. 5 of CBJ-042). The full $\mathrm{T_{eff}}$, $\log g$ and [Fe/H] grid comprises 4361 stars, of which 75% are randomly selected for training and the remaining 25% for testing. The performance on data at G=18.5 is

	[Fe/H]	$\log(\mathrm{T_{eff}})$
$\overline{\delta\phi}$	−0.02	1.0e^{-4}
$\overline{\lvert\delta\phi\rvert}$	0.40	0.0052
σ_ϕ	0.60	0.0070

ILIUM, all $\log g$, G=18.5, $\mathrm{T_{eff}} \leq 7000$ K

The performance is slightly worse in both APs than when we were restricted to either dwarfs or giants.

5 Estimating all three astrophysical parameters

The model in the previous section (call it $\mathrm{ILIUM_{feh}}$) allows us to estimate [Fe/H] and $\mathrm{T_{eff}}$ without knowing $\log g$. Having estimated [Fe/H], we could then use a $\mathrm{T_{eff}}$–$\log g$ version of ILIUM (call it $\mathrm{ILIUM_{logg}}$) at the appropriate [Fe/H] to estimate $\log g$, and thereby come up with a solution for all three APs. In CBJ-042 I demonstrated that $\log g$ could be estimated at G=18.5 to an accuracy of 0.35 dex (mean absolute error). That assumed [Fe/H] =0, but we could of course build several models of $\mathrm{ILIUM_{logg}}$ at different metallicities and choose the appropriate one based on the estimated [Fe/H]. To check the viability of this approach, I have trained and tested an $\mathrm{ILIUM_{logg}}$ model now using a small range of metallicities, namely [Fe/H] $\in \{-2.5, -2.0, -1.5\}$, on G=18.5. This range simulates having identified a metal poor star with some uncertainty in the metallicity estimation. The full data set contains 874 stars, randomly split into equal-sized train and test sets.

	$\log g$	$\log(\mathrm{T_{eff}})$
$\overline{\delta\phi}$	−0.077	4.2e^{-4}
$\overline{\lvert\delta\phi\rvert}$	0.49	0.0058
σ_ϕ	0.79	0.0081

ILIUM, G=18.5, [Fe/H] $\in \{-2.5, -2.0, -1.5\}$

$\mathrm{T_{eff}}$ can be estimated just as accurately as the [Fe/H] = 0 case (CBJ-042). $\log g$ is slightly worse (it was $\overline{|\delta\phi|} = 0.35$ with [Fe/H] = 0) but still reasonable. This shows that a two-stage approach to estimating all three APs is viable: (1) use ILIUM$_{\rm feh}$ to estimate [Fe/H] and $\mathrm{T_{eff}}$; (2) use ILIUM$_{\rm logg}$ to estimate $\log g$ and $\mathrm{T_{eff}}$. In principle we could even iterate this and re-estimate [Fe/H] again with ILIUM$_{\rm feh}$ model and thereby achieve better accuracy. Maybe we would have the first ILIUM$_{\rm feh}$ trained only on dwarfs to get best accuracy on most stars. There are many alternatives. Of course, it is still quite possible that ILIUM can be extended to multiple weak and strong APs, as described in section 5 of CBJ-042.

References

Bailer-Jones C.A.L., 2008, ILIUM*: An iterative local interpolation method for parameter estimation*, GAIA-C8-TN-MPIA-CBJ-042

Sordo R., Vallenari A., 2008, *Description of CU8 cycle 3 simulated data*, GAIA-C8-DA-OAPD-RS-002

Zaldua I., et al., 2008, *Interface Control Document for GOG v2.0.2 (cycle3)*, GAIA-C2-SP-UB-IZ-001-02

A Why we cannot correct for the systematic errors

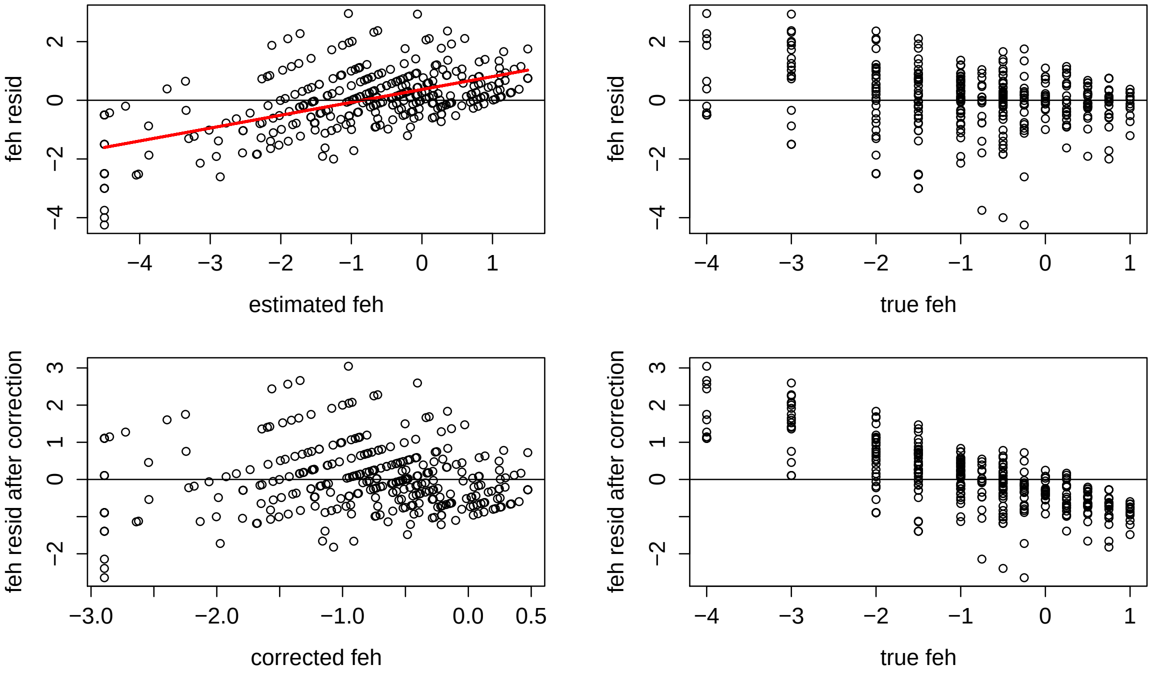

FIGURE 16: Metallicity residuals for the G=20 dwarf experiment in section 2.3 for $\mathrm{T_{eff}} \leq 7000\,\mathrm{K}$. *Top right*: [Fe/H] residuals against true [Fe/H]. *Top left*: [Fe/H] residuals against estimated [Fe/H]. The red line is a linear fit to achieve a correction. *Bottom left*: [Fe/H] residuals after applying the correction vs. the corrected [Fe/H]. *Bottom right*: The corrected residuals plotted against the true [Fe/H].

Fig. 16 demonstrates why we cannot correct for systematic metallicity errors, at least not for the G=20 case discussed in section 2.3. First, we cannot deduce from the plot of the residuals against the *true* [Fe/H] whether or not a correction is possible (top right panel): The true [Fe/H] cannot be the basis of a correction of unlabelled data! If we plot against the estimated [Fe/H] (top left panel), then we see a systematic trend which we can fit, e.g. with the red line shown. We then subtract this from each estimated [Fe/H] to give the *corrected* [Fe/H]. We can then analyse how well this correction has performed. The bottom left panel shows the residual in the corrected [Fe/H] (i.e. corrected minus true) plotted as a function of the corrected [Fe/H]. Comparing to the plot above it, we can see how the correction has worked. However, if we now plot the residuals against the true metallicity, we see that the systematic has actually got worse (compare to the plot above it).