200 (see Efron & Tibshirani 1993 for a detailed discussion and examples where this does not hold). Note that the 'error on the error' is proportional to $\frac{1}{\sqrt{(B-1)}}$.

The presented bootstrap strategy is also called *Bootstrap pairs sampling* and should not be confused with *Bootstrap residual sampling.* For the latter, the model residuals $(y_i - \hat{y}_i)$ are taken as the sampling units. As outlined in Dybowski & Roberts (2000), residual sampling uses the assumption that the residuals are independent of the inputs which need not be the case. Due to this and other reasons (see especially Tibshirani 1995), the more robust pair sampling procedure was chosen for these tests.

Above, it is assumed that the ensemble of networks yields unbiased estimates of the (true) regression function. As discussed in Heskes (1997), this is not really the case (neural networks are biased estimators) since models trained on a limited number of patterns will always tend to oversmooth sharp peaks in the data. We here follow Heskes (1997) and assume that the bias component of the errors is small as compared to the variance component (see Sect. 3).

2 The data set and bootstrap simulations

The data used were those of Blind Testing Cycle 2 for G=15 and 19 mag. For completeness, a graphical representation of the astrophysical parameter (AP) combinations in the training and validation sets is shown in Fig. 1 (but see also Brown 2003).

A training set at a given magnitude has 20000 different APs with the corresponding 11 filter fluxes. There are 20 different noise versions (noise in the filter fluxes) of the training set. We performed several tests with different numbers of noisy training templates for a given astrophysical parameter combination (see Sect. 4). This procedure is motivated by the fact that additional noise can help to regularize the network, e.g. prevent overfitting (see e.g. Bishop 1995). The random resampling was done with a random number generator as given in Press et al. (1992).

For the general case (**case1**), we chose 1 of these 20 noise versions with 20000 inputs of filter fluxes (plus corresponding stellar parameters) and randomly sampled 80 bootstrap sets. For such a large number of patterns, 80 bootstrap samples may seem rather small. To allow for a sufficiently large amount of resampling (by which we mean that a pattern does not appear in the training set), we therefore ensured that 1000 patterns of the original 20000 were indeed missing in each bootstrap sample (*on average*, each training pattern is missing four times for the 80 bootstrap samples). The tests for this case were done for the 1X and 2F photometric system.

Additionally, we calculated the bootstrap errors from the same networks' outputs, but only for 20 bootstrap replications (instead of 80). This is referred to as **case1b**.

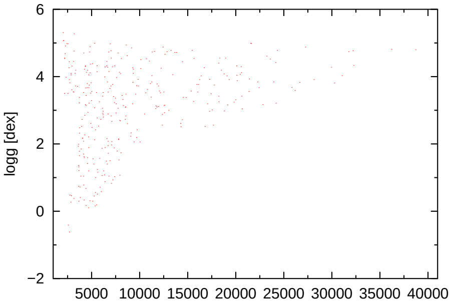
For the second case (**case2**), we did the same as above but, in case that an AP combination was randomly chosen into the bootstrap sample, we chose 5 noisy (filter flux) versions for this specific AP, i.e. each bootstrap sample is made up of 20000 × 5 (number of noise versions for given AP) training templates. In total, we created 20 bootstrap samples in this way for both magnitudes. Only the 1X system was considered for this case. Note that the random resampling was done over the APs and not over the noise versions!

For the third case (**case3**), we did the same as above, but chose in total 15 noise versions for each randomly resampled AP, i.e. each bootstrap sample is made up of 20000 × 15 inputs (number of noise versions for given AP). As before, we created 20 bootstrap samples in this way for both magnitudes but only for the 1X system.

The choice of the number of bootstrap replications is here mainly motivated by the limited amount of available computer time. Especially for the last case with 300000 templates per training set we had to choose a smaller number of replications, given that the training for one of these network takes very long.

For the validation, we chose only one pattern out of the 20 noisy versions available. For each bootstrap sample a network with architecture 11:13:13:4 was trained (11 filter flux inputs, two hidden layers each having 13 neurons and 4 outputs, one for each parameter). In addition, we trained networks with the same architecture but on the whole training set of 20000 AP combinations (i.e. without resampling). For case1, a committee of 5 networks was trained, while for the two other cases only single networks were set up. The code used was that of Bailer-Jones (2000).

Note that these networks are different from those used in Blind Testing Cycle 2 (results of Kaempf & Willemsen) which were specialized networks trained on specific ranges in parameter space and which had a larger number of hidden neurons.



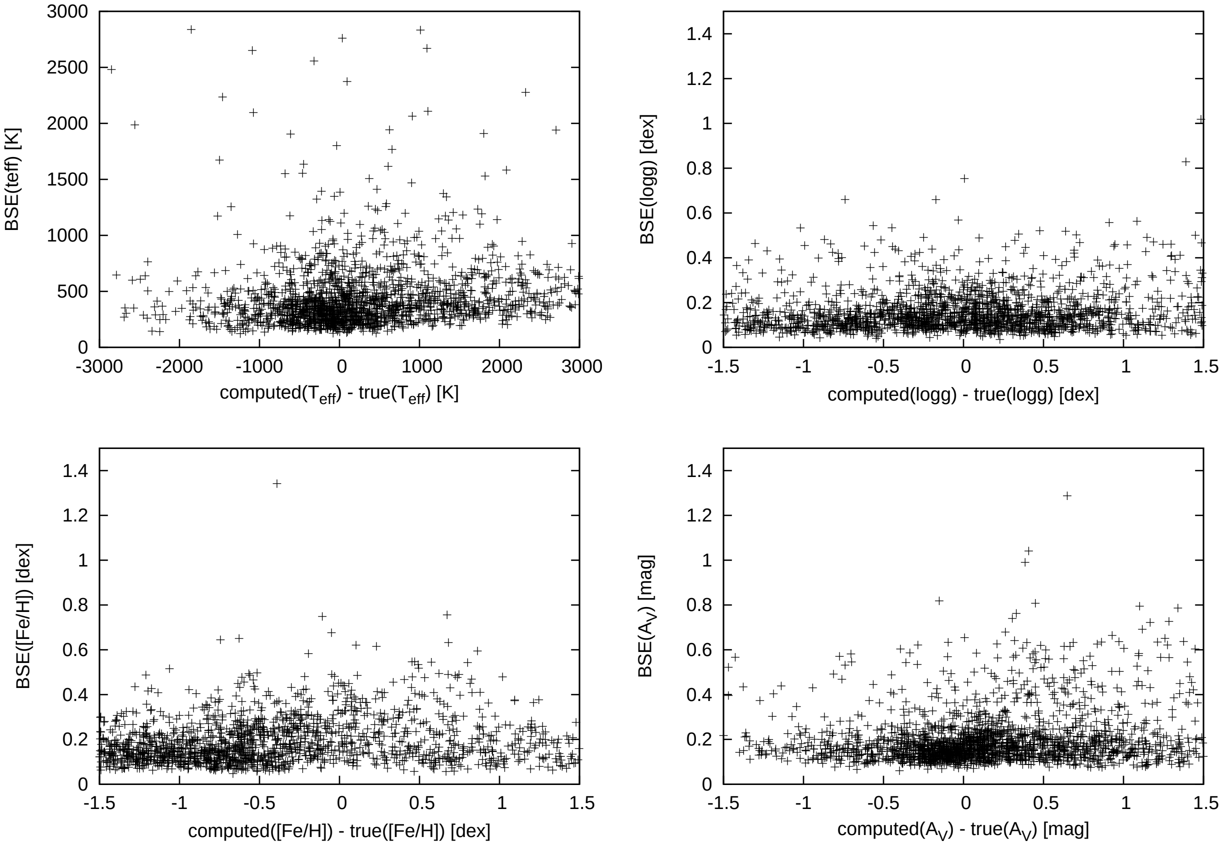


Figure 13: The same as in Fig. 12 but for $G = 19$ mag.

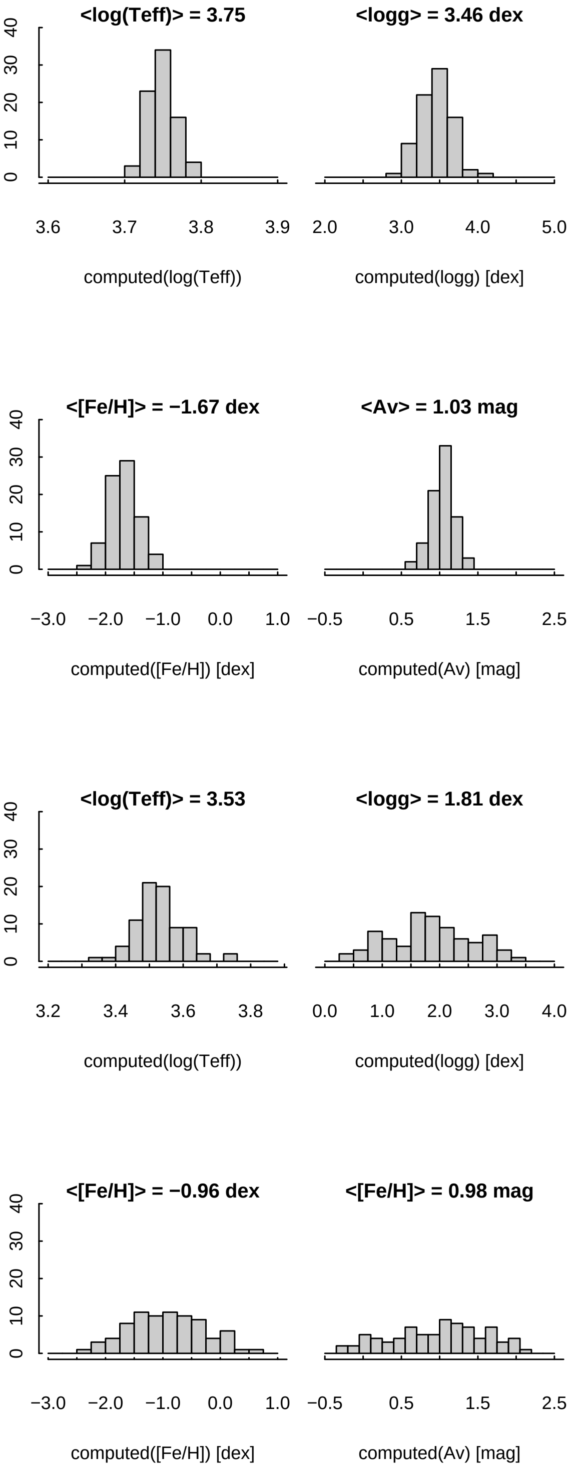


Figure 14: The distributions of 80 bootstrap neural network replications (case1, 1X) for two stars at G=15, with $\log g$ = 4.21 dex and $\log(T_{\rm eff})$ = 3.75 (top four panels) and $\log g$ = 1.73 dex , $\log(T_{\rm eff})$ = 3.55 (lower panels). A_V and [Fe/H] were fixed to 0.95 mag and −1.39 dex.