Response to the comments of Drew Shindell

We thank Drew Shindell for a detailed and thoughtful review of our paper, which certainly helped in improving the paper. Before responding to Shindell's comments, we would like to point out that this paper presents the main findings of a recently finished European project and serves as an overview of the project outcome. While the paper shall be complete and is intended to stand alone, for details we often refer to more specialized papers written by smaller ECLIPSE research teams. Several comments of Shindell's review concern methodological details that are better addressed in the specialized papers (some are not finalized yet and the authors will consider the comments). On the contrary, adding too many technical details in this already long paper would run counter to the idea of providing a concise overview of the – in our opinion - innovative overall concept we followed in ECLIPSE.

It is also difficult to address some of the comments that would require a different set-up of the models, or different type of analyses, since after the completion of the ECLIPSE project, it is virtually impossible to run all the models again in a consistent way. The individual ECLIPSE model groups will certainly consider some of the advice as guidance for future model experiments.

In the following, we repeat Drew Shindell's comments in normal font and add responses written in italic font.

The authors have chosen an interesting topic and provided many valuable results in their analysis. A particular strength of this study is that it comes from a large multimodel project. Some of the results such as the characterization of regional and seasonal metrics for species other than BC are fully multi-model and therefore of great value to the community.

We thank Drew Shindell for the overall positive reception of our paper.

Some of the primary findings, however, do not come from the multi-model results but rather depend upon single models. In particular, all the human health impacts come solely from the GAINS model's source-receptor relationships, …

With respect to human health impacts, it is true that these come only from one model. However, we would like to point out that we have also provided multi-model mean changes in surface concentrations of, e.g., ozone and $PM_{2.5}$ (see Figure 10) and some of the modelling groups also investigated changes of air quality in Europe and China, using regional models (not presented in this paper).

… and the net climate impact of the BC-related measures is highly dependent upon the estimated importance of processes such as the semi-direct effect of BC and the BC albedo forcing that were also simulated by just one model. The reliance on single models for some of the main results suggests that those conclusions may not be as robust as others from this project and this should be made clearer to the reader I feel.

While the metrics and metrics-based model calculations used the explicit semi-direct effect from a single model, the coupled simulations implicitly included the semi-direct effect in all four models. The snow-albedo effect was indeed only considered in one model but on a global scale this is less important. However, we agree that we need to make this clearer and, thus, we have added at the end of section 2 the following statement: "In ECLIPSE, we used multi-model ensemble results wherever possible, as these are more robust than results from an individual model. However, certain calculations could only be performed by a single model and are, thus, presented as such." In section 3.5, we have also added: "Notice, however, that, in contrast to the changes of pollutant concentrations presented in Fig. 10, the quantification of human health impacts is based on results from a single model."

Much of the paper reads very well and provides detailed descriptions of the simulations and results. Two sections in particular could use improvement in my opinion. The first is section 3.2 on model evaluation. Given that aerosol forcing is key to much of this study, and although not directly observed it is most readily evaluated by examining aerosol AOD and AAOD, I was quite surprised that there is no evaluation of the models against AeroNet and/or satellite AAOD. Capturing surface BC (Fig 5) is a good thing, but tells us little about how aerosols influence climate in these models. The lone comparison in Figure 4 is with total AOD, so does not tell us how the models simulated absorbing aerosols vs reflective aerosols or even natural vs anthropogenic aerosols. The models should be compared with observational constraints for absorbing aerosols (AAOD), as was done in Bond et al (2013) for example, in particular since that study found that all models were systematically biased low (as had been shown in prior studies). Finemode aerosol AOD might also be a good test though admittedly the observations are not without their own issues.

Comparisons such as suggested by Drew Shindell were done within ECLIPSE, building on the Aerocom platform. For instance, deliverable report 2.1 of ECLIPSE (available from http://eclipse.nilu.no/) presents correlation coefficients, normalised mean bias and root mean square errors calculated for AOD, fine-mode AOD and absorption AOD, for the ECLIPSE models against Aeronet sun photometer data. These are comparisons routinely performed for aerosol climate models and, as such, not particularly novel. Therefore, we decided to present statistical comparisons against data that are "less off the shelf" and thus, we believe, are more novel and complementary to information published in similar form elsewhere. Furthermore, the reliability of absorption AOD derived from Aeronet data is debated and may be questionable. Our own comparisons of BC concentrations from different observation platforms are much less conclusive and suggest that it is dangerous to use a single observation type as a reference for model comparison. However, we realize that we should have mentioned these comparisons. Therefore, we added the following:

"The improved ECLIPSE models were evaluated against global data sets such as AOD, fine-mode AOD and absorption AOD derived from data of the Aeronet sunphotometer network, as well as against measurements of aerosol and gas-phase species (Schulz et al., 2015). Here, we focus on a more detailed regional model evaluation for Eastern Asia, Europe and the Arctic …"

Secondly in this section, and along similar lines, the models should be compared with observationally-constrained estimates of forcing by absorbing aerosols (showing values of about 0.9 W/m2 for all present-day (natural plus anthropogenic) BC (e.g. Ramanathan and Carmichael, 2008; Bond et al., 2013). The modeled values should also be compared with values reported in recent comprehensive assessments (IPCC AR5 and Bond et al., 2013 in particular) to give the reader some context for the models used in ECLIPSE. For example, for the direct radiative forcing of preindustrial to present-day BC increases, the reported values from recent Assessments are 0.45 (0.30- 0.60) W m-2 in UNEP/WMO (2011), 0.71 (0.08-1.27) W m-2 in Bond et al. (2013), and 0.54 (0.10-1.03) W m-2 in IPCC AR5 (2013).

The forcing values mentioned above include a mixture of forcing mechanisms, e.g. the Bond et al. (2013) *value includes only the direct aerosol effect, whereas the IPCC AR5 value includes direct aerosol effect, semi-direct effect and BC impact on snow and ice.*

One difference between these numbers and the ECLIPSE forcing is that BC emissions from biomass burning are not considered in ECLIPSE, so only BC sources from fossil fuel and biofuel emissions are included. In Bond et al. (2013) and IPCC AR5 the direct aerosol effect of BC from biomass burning was estimated to +0.2 *W m^{-2}*. *When scaled to 100% BC reductions and reported annually, ECLIPSE BC radiative forcing and its range are 0.28 (0.02 to 0.46) W m^{-2}. Neglecting semi-direct effects, but including indirect aerosol effect of BC those numbers become 0.35 (0.11 to 0.48 or ± 0.24) W m^{-2}. ECLIPSE models are therefore at the stronger end of the multi-model ACCMIP estimate of 0.24 ± 0.1 W m^{-2} (Shindell et al., 2013) and the AeroCom estimate of 0.18 ± 0.07 W m^{-2} (Myhre et al., 2013). In constrast, ECLIPSE estimates are in the middle or weak range of the UNEP assessment and Bond et al. (2013). So in the context of global modelling, ECLIPSE BC forcing estimates are quite representative of both magnitude and range.*

As stated by Drew Shindell in a later comment, the reasons why ECLIPSE estimates, and more generally modelled estimates, do not reach the stronger ends of the ranges listed by Drew Shindell above, and those mentioned in section 3.3 of the manuscript, is that the latter include observational or empirical scaling. Those reasons however are not sufficient to warrant an upward revision of ECLIPSE estimates.

The stronger estimates in the UNEP/WMO range (section 3.3 and Table 3.1 of the UNEP/WMO report) stem from a series of difficult choices, discussed in details in the UNEP/WMO report itself. Its direct forcing is skewed towards stronger values by accounting for semi-empirical estimates, which include a brown-carbon contribution and need to be corrected in a very uncertain way for the pre-industrial baseline. The UNEP/WMO semi-direct forcing estimate is also skewed towards stronger values by including the possibility that BC semi-direct forcing is the same sign as direct forcing. In most global models, however, semi-direct forcing systematically opposes direct forcing on the global scale. The UNEP/WMO estimate of BC deposition forcing is also made stronger by accounting for a very large efficacy of 5, essentially making it an effective radiative forcing – a step that has not been taken in the ECLIPSE estimate.

The large observationally-constrained estimates of BC forcing by Bond et al. (2013), which influenced the IPCC estimate, have recently been called into question by several studies (Wang et al., 2014; Samset et al., 2014; Wang et al., 2015). There are several reasons for this. Wang et al. (2014) and Samset et al. (2014) argue that BC residence time has to be relatively short to fit observed remote concentrations. Furthermore Wang et al. (2015) showed that the fairly low resolution of global models induces an artificial negative bias when comparing to AERONET stations in Asia. Accounting for *both representativeness and observation errors results in a reduced best estimate for the BC direct radiative forcing.*

To address Drew Shindell's comment, we have added a shortened version of the above discussion to the paper.

Section 3.6 on climate impacts is the other section that is lacking a substantial amount of information. In particular when it comes to describing the model results for the impact of BC, the paper does not give enough information for the reader to judge the credibility of the results. The authors discuss two possible reasons for the small values and wide range in the climate response, unforced variability and differences in processes. For the former, the authors refer to the possibility of "unforced responses of climate system components, especially sea ice, that happen to counteract the small temperature response".

Comprehensive information on the models was provided by Baker et al. (2015a). This manuscript was previously in Open Discussion but is now already in press for Atmos. Chem. Phys. We do not see the need to repeat all this information also for the overview paper. However, we realize that it might not have been so obvious for the reader where this information can be found. Therefore, we have added explicitly:
"This experiment is described in detail in Baker et al. (2015a), where also more detailed descriptions of the models used can be found. Here, we only provide a synthesis of the results."

These results come from models, not observations, so one can and certainly should test this to find out. The models need to be run long enough and have enough ensemble members to reduce unforced variability to a small enough magnitude to see the forced response or at minimum bound it rather than presenting a response that may consist largely of internal climate 'noise'.

In principle, this could be tested. However, in practice the computational effort is too large. Notice that in contrast to previous studies (which have used perturbations of 10 times the real BC emissions), we have only removed the existing BC emissions. This provides a signal approximately 10 times weaker than in previous studies, which is much more difficult to detect in practice. Much longer (computationally prohibitively long) model runs would have been needed in order to extract a signal of comparable statistical significance as in previous studies. This leaves us with some signal-to-noise detection problems. On the other hand, our perturbation is much more realistic than that used in previous experiments and we have added ensemble members for the final *version of this article and Baker et al. (2015a). For more discussion, see Baker et al. (2015a).*

Given the difference in the two CAM4 ensemble members shown in Figure 12, it seems that the model sampling may be inadequate, in which case the conclusions of the study in this regard need to have substantially expanded caveats. For the question of processes, the authors write that differences in the response to BC "could be due to the different sizes of the indirect and semi-direct effects of BC in different models". Previously the paper described how only one model included the semi-direct effect, one included BC albedo forcing, and 3 included the first aerosol indirect effect. Given the strong differences in the response to BC shown in Figure 12, it is important to state more clearly which processes are included in which of these models.

As pointed out earlier, in the climate perturbation runs, all models implicitly include the semi-direct effect. To make this clear, we have added at the beginning of section 3.6:
"Note that all ESMs implicitly include the semi-direct effect, while for the radiative forcing calculations this was calculated explicitly by only one model."
More details about the individual models are given in Baker et al. (2015a).

It is also important to present the effective radiative forcing due to BC in these models as without this it's very difficult to understand how, for example, the HadGEM model produces a warming in response to BC removal. It seems likely that HadGEM has an overall negative forcing from BC given its response, which is a surprising result, and ECHAM seems to have positive forcing from OA (maybe due to strong 'brown carbon' absorption) but without diagnostics of the forcing from various components, a clear description of which processes are included and a comparison of total present-day forcing with observationally-constrained estimates it's difficult to have an understanding of how credible these results are. There is clearly a strong negative BC semi-direct effect in the one model that simulated this in ECLIPSE, yet in the Assessment of Bond et al (2013) multiple BC-cloud impacts were analyzed including indirect effects on liquid and ice clouds, semi-direct effects, cloud inclusions, and mixed-phase clouds. In particular, Bond et al estimated that, like the one ECLIPSE model with results, the semi-direct effect was likely negative but that the effects of both BC cloud inclusions and BC's mixed-phase cloud forcing were positive and larger than the semi-direct effect (the latter without inclusions, as I'm assuming was done here). Thus the overall indirect BC forcing used in this paper appears to come from only a subset of known processes and hence to be of the opposite sign to the Assessement of Bond et al (summing over all their quoted values), so it's important to know if this is in fact the case by specifying clearly which processes are included and their forcing values. Including only negative indirect effects and neglecting positive ones might be a reason that the net climate impact of BC in the models is so small, though without a clearer description of the models' processes and forcings it's impossible to be certain.

Effects like BC cloud inclusions and BC's mixed-phase cloud forcing were not included. They are far too speculative to be included in global models, at least for our type of study. Observational constraints are essentially non-existent, and to rely on those to argue for a stronger BC radiative forcing seems misleading.

Effective radiative forcing calculations have not yet been done for all models involved. However, as emphasized earlier, all models included the semi-direct effect in the climate response calculations. The smaller warming due to BC compared to Bond et al. is thus not due to ignoring the semi-direct effect.

In coupled model simulations, the sign of the response is not always determined by the sign of the effective radiative forcing (ERF). The fact that HadGEM warms when BC is removed does not mean that BC ERF is negative. It means that the sum "ERF + internal variability in radiative fluxes", which is dominated by internal variability when ERF is small, is positive. To obtain more robust climate response results, several additional model ensemble members were run in the meantime. Notice, in particular, that the "surprising" small warming in HadGEM due to BC removal (Fig. 12) is not present in an added other ensemble member of the same model, which gives a small cooling, as the other models (new version of Fig. 12). However, overall, the addition of ensemble members did not change our overall results.

The obvious way to figure out the question begged by the results as to whether the BC forcing in the models is relatively small or the response to BC is relatively weak compared to prior literature is for the authors to add a table with forcing values (ERF) and responses per unit forcing for their equilibrium removal of individual aerosols simulations and compare with prior results. It seems a major oversight not to have this and I strongly recommend such an addition.

We agree that ERF values would have been a valuable addition to this paper. ERF calculations have been done already for individual ECLIPSE models and will be published *separately but are, at this time, not available for all the models. Therefore, ERF values will likely also not be part of the revised version of this paper.*

Finally, it seems likely that these models have smaller BC forcing than that in the recent Assessments of UNEP/WMO, Bond et al and IPCC AR5 at least in part because those three all used a blend of observations and models (adjusting modeled values upward to account for systematic biases relative to observations) whereas ECLISPE used native model results. If indeed that's the case, that deserves greater prominence in the discussion of these results which currently only mentions the semi-direct effect (not noting other indirect effects) and reduced BC lifetimes as possible reasons for differences relative to prior studies (the comparison with observations suggested previously would clarify if these models, with BC lifetimes and emissions that differ from prior studies, are likewise still biased low in AAOD). The prior CICERO groups' study (Hodnebrog et al., Nature Comm., 2014) indicated that lowering BC lifetime while simultaneously increasing emissions gave a better match to observations of the BC distribution, but had little effect on radiative forcing. It is difficult to reconcile those results with the suggestion here that the impact of BC is comparatively small due to the shorter lifetime used unless the reader assumes that the emissions are too small to allow the models to reproduce observations.
I reiterate that there is a lot of good material in this paper, but I think it could nonetheless be substantially improved by addressing these larger issues as well as some additional specific comments listed below. I believe it would be well worth the additional effort required.

Indeed, ECLIPSE used native model results rather than results scaled with selected measurement data sets. See our reply above on the BC radiative forcings from recent assessments. The scaling used, for instance, by Bond et al. relies heavily on the accuracy of Aeronet retrievals. However, both the representativeness and accuracy of the AAOD retrievals from Aeronet may be questioned and we believe it remains uncertain how to scale model results with a single measurement data set. Our BC comparisons, for instance, indicate that comparisons depend strongly on which measurement data set is chosen (e.g., eBC vs. rBC), with obvious (partly explainable) biases between the various measurement data sets. However, in our opinion it does not currently seem justifiable to scale model results based on a single observation type. In addition, measurements of the vertical profiles of BC lead to an important constraint of BC lifetime and abundance of BC in the middle and upper troposphere. Previous BC forcing results based on models *scaled to observations rely on models for the vertical profile of BC. These models have large abundance of BC in the middle and upper troposphere globally and observations show that such high BC abundances are not present in the upper troposphere. Recent observations and analysis within ECLIPSE and by other research groups show that such high BC abundance substantially overestimate the direct aerosol effect of BC.*

Additional comments:

P15158, L23-26: The text states "The climate response from BC reductions in our study is smaller than reported previously, largely because our study is one of the first to use fully coupled climate models, where unforced variability and sea-ice responses may counteract the impacts of small emission reductions." It is not correct to say that unforced variability may counteract the impacts of forcing as that confuses forced and unforced climate change. Unforced variability doesn't reduce the impact of forcing, rather it may mask it if statistics are inadequate to remove it. This should be analyzed (see further comments below).

We have added the words "may counteract <u>(and, thus, mask)</u>" to make clear what we mean.

P15160, L8-9: The text states "Methane is a greenhouse gas roughly 26 times stronger than CO_2 on a per molecule basis at current concentrations." This is an incomplete description as this is assuming integrated RF over 100 yrs, and with a different physical quantity or time horizon the ratio would be quite different.

Our statement is factually correct. What we are referring to here is radiative efficiency, which is an instantaneous value (in W m^{-2} $ppbv^{-1}$) and does not require a timescale, in contrast to the Global Warming Potential (GWP), which Drew Shindell is obviously referring to in his comment. Coincidentally, the GWP of methane for a 100-year horizon (28) is numerically similar to the radiative efficiency relative to CO_2 (26) (see table 8.A.1 in IPCCC WG1 Assessment Report Five), probably explaining the reviewer's confusion. In order to make our statement more clear we have changed the wording to: "Methane is a greenhouse gas with a radiative efficiency (in W m^{-2} $ppbv^{-1}$) roughly 26 times greater than that of CO_2 at current concentrations."

P15164, L8: I provided additional analyses supporting the claim that even cooling agents can cause damaging climate effects via changed in precipitation in Shindell, Climatic Change, 2015.

Thank you for the additional reference, which we have added.

P15167, L5-7: Reference to later and broader studies such as Anenberg et al., Env. Health Pers., 2012 or Lim et al, Lancet, 2012 would be better than our earlier, simpler model-based study (Unger et al.)

OK, we have replaced the reference with Anenberg et al. and Lim et al.

P15176, L26-28: There is substantial debate about how much of the difference between the regional temperature changes during these time periods is due to aerosols vs oceanic circulation changes. This makes this a fairly weak test of a model's ability to capture the climate response to aerosols.

Well, we did not only want to test the models' ability to capture the climate response to aerosols, but more generally to test the models' ability to reproduce trends – even though in the presented case aerosols likely contribute strongly.

P15178, L9-11: The text states "Quantifying the semi-direct effect has large uncertainties, however, because internal variability in the climate system masks tropospheric adjustments to BC perturbations." So how long were the simulations run to diagnose this and what is the resulting uncertainty? Please also explain how this uncertainty is incorporated into the total SRF uncertainty (shown in Fig 7) given that you only had one model diagnosing the semi-direct and BC albedo forcings vs multiple models for direct forcing?

As noted earlier, all models include the semi-direct effect in their climate response calculations, whereas only one model was used to calculate the SRF due to the semi-direct effect. Therefore, no uncertainty values were available for the SRF values for the semi-direct effect. The model calculations were run over 50 years.

P15179, L7-9: The text states "ECLIPSE is therefore able to state with confidence that NOx exerts a negative SRF, because the O3 response is not sufficient to offset the combined CH4 and nitrate response." I think it'd be worth pointing out that this conclusion is consistent with that of the IPCC AR5 (Myhre et al., 2013).

We have slightly reformulated the text to read: "…able to confirm with confidence the earlier quantifications (Myhre et al., 2013) that…"

P15179, L11-13: The text explains that the SRF for methane differs across models due to the different lifetimes for methane in the models. These should be compared with observationally-constrained values (the Prather et al values cited previously) and if some models are unrealistic then they should be excluded.

None of the models is "unrealistic". Some of the uncertainty also arises via methane's effects on ozone and aerosols. We only wanted to emphasise that the uncertainties are not due to uncertainties in absorbing properties, but realize that our text may have been misleading. We have shortened this to simply read: "…, reflecting the differences in methane lifetime and methane's effects on ozone and aerosols." Radiative properties are not mentioned anymore.

P15181, L24-26: Discussing the RTP coefficients, the text states "Even though the coefficients are likely model-dependent, we had to use these values because they are not available from any other model (and specifically not from the ECLIPSE models)." It would strengthen the reader's confidence in the use of these single-model values to point out that they appear to be fairly robust in comparison with the response to historical aerosol forcing in several other models (Shindell, Evaluation of the absolute regional temperature potential, Atmos. Chem. Phys., 12, 7955–7960, 2012).

Thanks, we have added: "The coefficients also seem fairly robust in comparison with the response to historical aerosol forcing in several other models (Shindell, 2012)."

P15183, L11-13: This text discusses the NMVOC-related solvent measures, but I didn't see previously any discussion of metrics for NMVOCs (though it did say metrics were calculated for 'others'). It's not obvious how one would calculate those given the large number of species. Could the authors please explain how these measures fit within the ECLIPSE methodology given that measures are selected based on their GTP? Were there in fact NMVOC GTPs, and if so, for which species?

Specific radiative forcing values and GTP20 values were calculated for a larger range of species than shown in Fig. 7+8, including also NMVOC emissions. Details are presented in Bellouin et al. (2015) and Aamaas et al. (2015 (submitted) but were deemed not suitable for this overview paper.

Figure 12: It seems the uncertainty on the multi-model means is something like the average of the uncertainty of each individual model (or perhaps propagating those individual ranges mathematically) but it doesn't include the difference between the models. If so, this doesn't seem a sensible approach to me.

Indeed, the multi-model mean does not include the differences between the models. However, the number of models is too small to calculate a statistically robust uncertainty range. By showing all individual models, we assume that the reader is able to judge the uncertainty arising from the differences between the models.

Additional references:

Samset, B. H., Myhre, G., Herber, A., Kondo, Y., Li, S.-M., Moteki, N., Koike, M., Oshima, N., Schwarz, J. P., Balkanski, Y., Bauer, S. E., Bellouin, N., Berntsen, T. K., Bian, H., Chin, M., Diehl, T., Easter, R. C., Ghan, S. J., Iversen, T., Kirkevåg, A., Lamarque, J.-F., Lin, G., Liu, X., Penner, J. E., Schulz, M., Seland, Ø., Skeie, R. B., Stier, P., Takemura, T., Tsigaridis, K., and Zhang, K.: Modelled black carbon radiative forcing and atmospheric lifetime in AeroCom Phase II constrained by aircraft observations, Atmos. Chem. Phys., 14, 12465-12477, doi:10.5194/acp-14-12465-2014, 2014.

Shindell, D. T., Lamarque, J.-F., Schulz, M., Flanner, M., Jiao, C., Chin, M., Young, P. J., Lee, Y. H., Rotstayn, L., Mahowald, N., Milly, G., Faluvegi, G., Balkanski, Y., Collins, W. J., Conley, A. J., Dalsoren, S., Easter, R., Ghan, S., Horowitz, L., Liu, X., Myhre, G., Nagashima, T., Naik, V., Rumbold, S. T., Skeie, R., Sudo, K., Szopa, S., Takemura, T., Voulgarakis, A., Yoon, J.-H., and Lo, F.: Radiative forcing in the ACCMIP historical and future climate simulations, Atmos. Chem. Phys., 13, 2939-2974, doi:10.5194/acp-13-2939-2013, 2013.

Wang, X., Heald, C. L., Ridley, D. A., Schwarz, J. P., Spackman, J. R., Perring, A. E., Coe, H., Liu, D., and Clarke, A. D.: Exploiting simultaneous observational constraints on mass and absorption to estimate the global direct radiative forcing of black carbon and brown carbon, Atmos. Chem. Phys., 14, 10989-11010, doi:10.5194/acp-14-10989-2014, 2014.

Wang, R., Y. Balkanski, O. Boucher, P. Ciais, G. L. Schuster, F. Chevallier, B. H. Samset, J. Liu, S. Piao, M. Valari, S. Tao, Quantifying the role of black carbon in the Earth's climate system, *Nature Scientific Reports*, submitted, 2015.