Author answer to the referee report of EPJC-18-04-021

We thank the referee for the positive report and for her/his insightful comments. Below, we address all the questions and suggestions contained in the report in order. To ease the reading, the points raised by the referee (quoted in italic) are followed by an explanation of the actions undertaken to address them.

1. *At the beginning of Sec. 2 "LO" is defined as the three possible coupling combinationas α^6, $\alpha_s\alpha^5$, $\alpha_s^2\alpha^4$. However for most of the paper "LO" is more and more taken equivalent to only mean the EW α^6 contribution. (At first it still says "LO at α^6", then it becomes LO i.e. α^6, and eventually in the LO+PS section the coupling order is dropped completely. This is of course just semantics, but it was quite confusing on the first reading and it would be good to make this a bit clearer from the start.*

   In section 2, before the last paragraph, we have added the following sentence:

   > In the rest of this article, the notations LO or NLO(-QCD) without any specification of coupling powers refer to the contributions at order $\mathcal{O}(\alpha^6)$ and $\mathcal{O}(\alpha_s\alpha^6)$, respectively.

   For clarity, we have also re-stated the coupling orders at the beginning of section 6 (Matching to parton shower).

2. *The ordering of the various programs in tables 1, 3, and 5 is very helpful. I can perhaps understand the reasoning behind choosing this alphabetical ordering, but I think it is misplaced here and makes it quite hard to get anything out of this. It would be much more useful to have programs with similar levels of approximations next to each other. So keep Phantom and Whizard together (both having the same approxmations), then Bonsay, Powheg, VBFNLO (all without interference with VBFNLO adding the s-channel), then MG5_AMC (also adding interference), and last MoCaNLO+Recola having the full result.*

   We have changed the order as suggested. For consistency, we have also changed the order of the description of the various codes in section 3.2, and of the results in tables 3 and 5.

3. *In table 3: Please indicate which ones are exact at LO, and which ones are not. Also, regarding the discussion in Sec. 4.3 and the fact that several full LO predictions differ far outside their statistical MC uncertainties. This I find extremely puzzling and the offered explanations are quite unsatisfactory. After all, at LO, MC statistics is just statistics, I don't understand how statistical uncertainties can be aggressive or not, they are what they are. Also, the complex-mass scheme is a well-defined procedure, so how can different implementations of it lead to numerical differences? (If there are somewhat different approximations being employed, it would be good to spell this out clearly.) The fact that the differences are at the 0.5% level cannot hide the fact that they are much much bigger than the MC statistics. Taken at face value, that means that there are clear systematic differences where from what I understand there shouldn't be any, so the reader is left to wonder what is going on? To play the devil's advocate, if there are systematic differences at LO that cannot be understood, how can I trust the NLO comparison?*

   In the caption of table 3, we have added the sentence

   > The complete $2 \to 6$ matrix-element, without any approximation, is employed by PHANTOM, WHIZARD, MG5_aMC, and MoCaNLO+Recola.

   In order to have a better understanding of the origin of the differences among the four codes, we have compared the matrix-elements, finding very good agreement. Other possible sources of the differences may originate from the factorisation scale setting or the cuts, or, more likely, from the non-perfect or non-ideal coverage of the VBS phase space, which is rather complex (in particular when also non-resonant contributions have to be included). Another issue of concern, as we already have written in the submitted version, is whether one should trust or not the uncertainties quoted by the Monte Carlo integrator. We have modified the text in which we discuss the results at LO accuracy:

Two things should be highlighted here: first, despite the different underlying approximations, the two most-distant predictions (POWHEG-BOX and MG5_aMC) are only 0.7% apart. This simply means that the details of the various VBS approximations have an impact below 1% at the level of the fiducial cross section at LO for a typical phase-space volume used by experimental collaborations. This is in agreement with the findings of Refs.[Denner:2012dz,Oleari:2003tc]. Second, the four complete predictions (WHIZARD, PHANTOM, MG5_aMC, and MoCANLO+RECOLA) are not in statistical agreement. While we have checked the point-wise agreement of the matrix-element, we cannot exclude other reasons for the disagreement, for example a non-representative (*i.e.* too-aggressive) estimate of the Monte Carlo uncertainty or a non-perfect mapping of the six-body phase-space. However, the level of ambiguity is at the 0.5% level, which we deem satisfactory compared to the larger differences observed at NLO or when including matching to parton shower.

4. *page 16, 2nd column, line 13, typo: "elment" → "element"*

   It has been corrected

5. *page 18, 2nd column, line 48: "on the other ..." the part of the sentence is nonsensical.*

   We have corrected the sentence.

   > which clearly suggest not to rely on a single tool/parton shower, and on the other make...

   to

   > which on the one hand clearly suggest not to rely on a single tool/parton shower, and on the other make...

6. *page 19, 1st column, line 14: I think it is important to not mistake "more realistic" for "conservative". The scale variations are clearly underestimating the actual theoretical uncertainties. This does not mean that another method, which implies larger uncertainties, is "conservative", it is simply (hopefully) "more realistic". Conservative is easily (mis)interpreted as possibly overestimated, which I don't think is the case here. (In fact, there is no guarantee that even the comparison of different tools will provide an estimate that covers the next order, so calling this a conservative estimate is really misplaced.*

   We have replaced "conservative" with "more realistic", as suggested.

Further minor changes have been performed, not related to the referee's comments:

1. The references to experimental measurements in the second paragraph of the Introduction were not consistent with the text. The sentence has been corrected.
2. The program PHANTOM was sometimes referred to as PHANTOM, with all capital letters. All occurrences have been changed to PHANTOM. The same applies for WHIZARD.
3. The left plot of Figure 5 (dijet invariant mass at LO) has been updated, as the WHIZARD histogram was not present in the inset because of an error in the plotting script.
4. We have modified the sentence: "combined measurements which are better theoretically defined" to "combined measurements which are theoretically better defined".
5. We have updated the LO+PS plots. It has been found that the predictions for WHIZARD+P8 suffered from a bug which has now been fixed. The agreement with other Pythia parton-shower predictions is now even better. In addition, the plot at LO+PS for the Zeppenfeld variable of the third jet suffered from a miss-match of the PHANTOM+H7-default due to bug in the plotting routine. Again, the predictions of Phantom+H7-default are now in better agreement with Herwig parton-shower predictions. No modifications to the text have been made.
6. In Ref. [81] (Yu, Bardin, Leike, Riemann, 1988), the hyperlink was pointing to a wrong DOI. It has been fixed.