This paper discusses the Bayesian approach to inverse problems within the context of the NNPDF fits, and presents new statistical estimators for assessing the behaviour of the fitting procedure under closure tests. These are then applied to the NNPDF4.0 fit, in order to evaluate the faithfulness of the uncertainties within such a closure test context. This provides an in depth study of this important issue, and certainly merits publication. There are a few questions and clarifications I would like to see resolved before this however. These are presented below in the order they appear in the paper:

1. Introduction, first paragraph. Reference to Hadamard possibly needs a citation.

2. Page 2, left column, bottom paragraph. The question of how closely a closure test validates the uncertainties in the context of a real world fit (i.e. with data/theory and data/data inconsistencies) is discussed in the conclusion, and is in my view an increasingly important one in the context of the NNPDF4.0 fit. I think it would be worth also adding a brief comment here as well (e.g. after 'how faithful our uncertainties are') to make clear that what is being tested in terms of faithfulness is (while clearly important and non–trivial) not necessarily the end of the story with respect to a global fit.

3. Section 3.2. I was naively a little surprised that this sort of demonstration had not been given in previous NNPDF literature, given it seems rather key to the approach that the NNPDF methodology reproduces the posterior distribution of the model given the data (at least in some limits, as described in the paper). Was this just assumed to be true before? Possibly it is just a question of what language you use; clearly the MC replica approach intuitively makes sense, and certainly e.g. many comparisons between the Hessian and MC replica approaches have been made previously, with good consistency found. Some clarification would be good on this point.

4. Page 12, (107). The notation '$E_{y_0,y_0'}$' could do with being explicitly defined. The statement below this that the two estimators are independent of the test data would be better clarified as well, as it is not immediately obvious this is true. It seems pretty clear that the bias depends on the true values f', so presumably what is meant is that they are independent of η'.

5. Page 12, below (108). Please expand a little on what is meant by 'single replica proxy fits were used'.

6. Section 4.2. I believe what is considered is the case that the training and test datasets are the same (i.e. both the 2D data space), but as far as I can see that is not explicitly stated, whereas in the previous sections this is not assumed. So this could be worth stating explicitly.

7. Page 15, below (124). y_0 is mentioned, but not present in the equation above it?

8. Section 5, general comment, and the most significant one I have. Having derived these data space estimators it would have been nice to see a direct comparison made between the 3.1 and 4.0 methodologies with respect to these. In footnote 3 it is stated that the improved efficiency of the 4.0 methodology makes this possible. Well, in total $30 \times 40 = 1200$ replica fits are done here, and from 2109.02653 Table 3.3 an average run time per replica of 15 hours is quoted for the 3.1 case. Clearly that is very far from something you could do many times, but with a reasonable batch computer system does not seem beyond the realms of possibility, for the purposes of a single direct comparison as in this paper. Though I may have missed a key further obstacle in the above.

   Ideally, that would be good to see, though I understand if it is not judged to be feasible. But some comparison, perhaps using the older 3.1 statistical estimators, should be made here, and some discussion on this point given in the main body of the text. In particular, what I would like to see discussed/clarified is the extent to which these estimators can or cannot distinguish between the 3.1 and 4.0 cases, given 4.0 gives much smaller uncertainties than 3.1 for the same data. My understanding is that the bias–variance ratio will not do this, i.e. essentially in the Fig. 3 example the 3.1 case would give a larger blue circle, but one (hopefully) that still has the true value within its (larger) uncertainties 68% of the time. I could be wrong though, and at the moment as far as I can see this is not explicitly addressed. Given that, one might be left wondering how exactly the 3.1 methodology has passed the previous closure tests apparently so well, and yet is so different in result from 4.0. Note this is a separate question to the comparisons made in Fig. 5, which while very useful do not make use of any statistical estimators.

9. Page 17, bottom left. The statement about improved methodologies aiming to reduce the L0 and L1 errors. It would be interesting to elaborate on that a little. In particular, in Fig. 5 (top) we can see that there is still a pretty non–negligible L0 and L1 error in the 4.0 case. Could/should we expect a future improved release to reduce this further, or are we reaching the fundamental limit already?

10. Fig. 8 is not completely clear to me. What is the x axis showing? And do we expect the green and orange curves to follow each other in shape? They do to some extent, but clearly not completely. Some clarification would be good here.

11. Pages 19-20, starting last paragraph of page 19. This is I agree a key point. I think that even to say it is 'likely not the case in real worlds fits' is not strong enough. We know many examples of very poor fit quality to certain datasets, and even disregarding the real outliers, the global fit quality for all PDF fitters is not from a statistical point of view good. For example, in 2109.02653 Table 5.1 the $\chi^2/N_{\rm pts}$ is 1.16 for the NNLO fit, which while 'close' to one by eye is actually many σ away from it, given the large number of points. So in my view the statement should be even stronger. We know that textbook statistics are not obeyed in global fits (i.e. that there is not complete data/theory and

data/data consistency), and so the key question as we aim for accuracy and precision is how accurate our PDF uncertainties are within that context. In particular, the closure tests presented here are an important and necessary check, but are arguably not sufficient in the final analysis to say that the NNPDF4.0 errors are faithful in the global context (see e.g. how much the gluon moves in Fig. 4.2 of 2109.02653 when a single CMS dijet dataset is included). So in my view some expansion on that in line with the comments above, and strengthening of the language is in order.

12. In addition to the above, I think the statement on p.20 that 'Most of the observed inconsistency...is likely due to missing theoretical uncertainties' is definitely too strong. There is quite a bit of evidence for inconsistencies within datasets (e.g. ATLAS jets, differential top...) as well as between them, and no clear indication that MHO uncertainties will resolve these. Indeed in a NNLO fit the eventual MHO uncertainties will be rather small. So that is certainly one element, but I do not think there is any reason to say that most of the inconsistency will be due to this. That really remains to be seen.