Reviewer #1

Thank you for your encouraging comments. We are glad that you are able to recognize and understand the main contributions in our paper.

We have more to say about nonlinear models (including further experiments), which we can add to the camera-ready version. Basically, a version of the constraints in the regression formulation of sensor fusion can be also imposed for nonlinear models. We suggested this at the end of our paper in equation (17), but have since discovered a more tractable way to write the constraints in terms of local linear approximations.

Reviewer #2

Thank you for your thorough and helpful review. We appreciate all of your feedback. Bayesian viewpoint: this is a fair point, and we essentially agree with everything you say here. While we did not find something like Theorem 1 stated explicitly in the Faragher paper, nor in the Sarkka book (however, it is possible we missed it and we will look carefully through this entire book, before making this claim in any official way), we agree that the Bayesian/Gaussian viewpoint offer another lens from which we can understand Theorem 1. We are happy to add this along with appropriate references and discussion in a camera-ready version of our paper.

Thus by itself, Theorem 1 is not of high originality (for the reasons just noted, and to repeat, we will revise the text surrounding it appropriately). But we still think its role in our paper is significant, and hence it should be kept in the paper. This is because Theorem 1, combined with the insight in Theorem 2 that we can reformulate sensor fusion via a *forwards* or *direct* regression, suggests new and promising methodological possibilities. For example, we can throw in multiple candidate process models as sensors, then perform feature selection in the regression formulation to adaptively select some subset of them. This is explained in the "sensor selection" paragraph at the end of the paper and demonstrated empirically in Section A.7 in the supplement.

We are glad that you understand and appreciate the significance of Theorem 2. Empirical results/better demonstrations on real-world data: we have since rerun our experiments in Table 1 at the *US state level*, over more seasons. (This gives a test set with over 50x more observations: Table 1 only reports results at the national level over 4 seasons, which is much sparser in terms of a test set evaluation.) Sensor fusion combined with shrinkage now consistently displays a clear advantage over random forests (and all others, though random forests its closest competitor) throughout. This is important because it shows that even just a simple linear model with the "right" hierarhical constraints, encoding the measurement map, can perform well in comparison to a much richer, nonlinear/nonparametric method like random forests. It also suggests that with the "right" constraints put in place, a nonlinear method should do very well.

We are happy to add another sensor selection experiment to the supplement in a camera-ready version, rerunning an experiment like that in Section A.7 on real data. For example, we can try multiple process models on the flu data.

Reviewer #3

Thank you for your feedback; we regret that you were not able to appreciate our contributions and we will revise the paper accordingly to try to make the main points more salient. We hope that in the meantime, our response here will help clarify some things. First, we agree that Theorem 1 is simple and in it of itself of major originality. But we still believe that its implications when combined with Theorem 2 are significant, and believe that as such it belongs in our paper. Please see our comments in response to a similar point raised by Reviewer #2, above.

The name "sensor fusion": we are aware that this is a very broad term and describes a whole class of methods, not just equation (8). We apologize for the confusion. We simply needed something to call (8) in order to cleanly refer to it in our paper. See the bottom footnote on page 2 of our paper. We can definitely change this in a camera-ready version of the paper and are happy to hear alternative suggestions for a name. But to be clear, the fact that "sensor fusion" is a broad term (and has conferences associated with it) should not take anything away from the equivalences derived in our paper. We could have simply called this something else ("process-agnostic Kalman filter", or "linear inverse MLE").

Convex optimization view of Kalman filtering: we appreciate you raising this point, and sending an example reference, but we are in fact quite familiar with the optimization view of many estimation/tracking/control/prediction tasks, including Kalman filtering (the Boyd and Vandenberghe book, for example, makes this view ubiquituous). We ourselves work actively in this area as well. However, we must be perfectly clear that *this is not the same as the equivalence* as that we derive in Theorem 2, and the convex optimization view of KF has no bearing on the originality nor significance of this result. The work by Boyd and others just poses planning as a convex optimization problem, which is quite natural (and apparently, effective). Our Theorem 2 is completely different. It reformulates a *backwards* or *indirect* model of $z_{t+1}|x_{t+1}$ in terms of a *forwards* or *direct* model of predicting x_{t+1} from z_{t+1}. This has significant implications because it allows us to bring in the entire "ML toolbox" for this prediction problem. We can expand on the simulation in Section A.7 of the supplement, if you think this would help (see our comments to Reviewer #2 about this, too).