Table 1: Requested additional comparisons. SAT: Satellite, LAQN: Ground stations. Random seed of 0.

Model	Data Sources	sRMSE ($\mu \pm \sigma$)	RMSE ($\mu \pm \sigma$)	NLPL ($\mu \pm \sigma$)
Single GP	LAQN only	1.04 ± 0.04	23.02 ± 11.26	12.0 ± 12.22
MR-GP	SAT only	0.72 ± 0.41	14.87 ± 9.34	16.7 ± 23.14
VBAgg-Normal	LAQN & SAT	0.82 ± 0.48	16.24 ± 9.15	9.78 ± 11.97
MR-GPRN w/o CL	LAQN & SAT	0.69 ± 0.43	14.03 ± 8.93	9.24 ± 14.35
MR-GPRN w/ CL	LAQN & SAT	0.69 ± 0.42	14.45 ± 9.09	8.83 ± 12.92
MR-DGP	LAQN & SAT	$\mathbf{0.39 \pm 0.13}$	$\mathbf{8.65 \pm 4.93}$	$\mathbf{4.54 \pm 4.12}$

We thank the reviewers for their time and detailed, constructive feedback. We are glad to see our application-motivated methodological contributions and narrative were well-received. We denote with e.g. **R1.2.3** our response to Reviewer 1, Section 2, Paragraph 3. (**Joint**): As requested we are offering additional baselines (Table 1) that strengthen our results and are discussed below. As suggested by **R3** we will move inducing point material to the appendix. The additional page will allow us to improve clarity: we will expand on the MR-DGP model, we will add the additional baselines, suggested by **R1** and **R4**, with further discussions of results and the uncertainty quantification benefits of the CL corrections. We will also improve the description of the experiments and lighten the use of inline equations.

R1.2.4: We will improve the motivation for the composite likelihood (CL). The estimated CL ensures that the asymptotic posterior $p(\mathbf{Y}|\mathbf{f}, \mathbf{X}, \theta)$ converges to the misspecified asymptotic MLE distribution [24, 30]. The CL cannot be set as a free parameter because otherwise we would not obtain this theoretical guarantee. The MR-DGP learns dependencies between the layers $p(\mathbf{f}_1|\mathbf{f}_2, \dots)$ and hence between resolutions. **R1.2.5**: We have rerun all our experiments and additional requested baselines on the real world example, see [**Joint**] above. **R1.2.7** The size of the std is due to variability across the 42 sites in London, we will also offer site-standardized results (e.g. sRMSE) in the appendix. We are happy to follow alternative standardizations if reviewers express a preference for the final version.

R3.2.3: We agree that the distinction between multi-fidelity and multi-resolution would be beneficial and we will offer that. See [**Joint**] above to see how we will improve the explanation of MR-DGP (including the choice of kernel). MR-DGP arises very naturally in real world examples and is able to successfully handle data from biased sensor networks; as shown in our experiments the performance is significant. **R3.2.10-11**: We will reduce the number of inline equations by moving the standard results into the appendix. **R3.2.13**: This is an interesting paper that simply takes the formulation of [27] and applies it to the multi-task setting through the LCM formulation. The model proposed by the authors is a special case of MR-GPRN where the latent GPs $\mathbf{W}$ are constant. We have also presented a very natural and principled method to dealing with bias whereas they consider a very ad-hoc solution through data normalization. **R3.5.1**: It is indeed natural that specific contributions will be more or less interesting to different readers. In this paper we have tackled some of the underlying issues of previous approaches and offer the state-of-the-art.

R4.1.1: We respectfully disagree with "generalization of multi-task ... straightforward", in fact we have challenged two very common assumptions by correcting (MR-GPRN) or accounting (MR-DGP) for dependent observations and have provided a principled way to deal with biases and multi-resolution. Through these contributions we have shown impressive results on a very complex problem. **R4.1.2 + R4.2.Originality**: Both are latent variable models and with MR-DGP we use the latent structure to model the obs.dependency instead of correcting via CL. Further extensions or special cases of this framework we leave for future work. **R4.2.3**: Indeed the composite weight comes out of the expectation (See Eqn 14 in appendix). We do not model the cross-resolution dependencies via CL in the MR-GPRN model, the weight corrects for posterior contraction due to loss of that dependency as done in similar settings [24]. We will further demonstrate the UQ benefits of CL in the extra page through coverage and pred.densities. (**R4.2.4**) VBAgg is the only published work that is a suitable baseline. We have shown that both handle the same types of multi-res data in Sec. F of the appendix. As also suggested by **R3** we will merge this into the main text. We do not use [6] as baseline because, despite the name, it is unable to handle multiple observation processes, see ($\ell : 170 - 173$). **R4.2.5**: See [**R1.2.4**] above. **R4.2.6**: The dimension of the Hessian is $|\hat{\theta}| \times |\hat{\theta}|$ where θ are the hyper parameters. The size is very small and is dominated. We will clarify this in the main text. (**R4.2.Clarity.3**): Thank you for pointing out the typo in bolding MR-DGP, we will fix this. MAPE is an asymmetric loss that penalizes overestimation. As shown in Fig. 2 the prediction from MR-DGP is slightly over estimating whereas VBAgg-Normal is severely underestimating, hence MAPE over penalizes MR-DGP. **R4.2.Clarity.4**: Different sensor networks are calibrated differently, hence comparing raw values is not viable. The information theoretic corrections are from the composite weights in Sec. 3, we will clarify. Hyper-local is higher resolution than typical LSOA area estimates. (**R4.2.Clarity.5-7**): Thank you for pointing out these typos, we will fix them. The subscript m_k is meant to be m_a which represents the output for each layer. The p is used to denote the multiple tasks and because we have presented MR-DGP in the general case the ordering of tasks between layers is a user-choice. We will improve clarity throughout the paper based on all reviewers' suggestions.