[[Reviewer 1]] Thank you for the excellent comments and suggestions; we have updated the paper after taking all your comments into account. The 2-week performance of Imperial model for the US was mistakenly missed in Table 1, it is now provided in Updated Table 2. ■ **Evaluation metric:** We agree that evaluation on daily deaths is a more accurate metric for a model's generalization performance. We have amended Tables 1 and 2 by replacing the accuracy of predicting cumulative deaths with that of daily (incident) deaths—the updated results are summarized in Updated Table 2. With the new metric, our model still outperforms the baselines in the same countries and performs competitively in countries where it is not the best. More importantly, our key conclusions and insights regarding global hierarchical modeling are still preserved under the new metric. ■ **Uncertainty intervals:** Based on your suggestion, we evaluated the average *continuous ranked probability score* (CRPS) on daily deaths in Updated Table 2. Our model's probabilistic forecasts performed competitively compared to the baselines in all countries; we will also add results on coverage probabilities and CI length in the final version of the paper. ■ **Figures:** Fig. 3 (b) depicted the goodness of fit for daily deaths in the UK. In the final version of the paper, we will use the extra space to add similar figures for all countries in Table 2.

Updated Table 2: Accuracy of daily deaths predicted by baselines. (The Imperial model does not provide 30-day forecasts.)

Country	Mean Absolute Error on Daily Deaths (CRPS: continuous ranked probability score)						
	14-day Forecasts				30-day Forecasts		
	CGP	Imperial	IHME	YYG	CGP	IHME	YYG
US	139 (0.076)	149 (0.282)	753 (0.164)	**50** (**0.073**)	481 (0.196)	957 (0.260)	**365** (**0.164**)
UK	**58** (0.089)	164 (0.248)	288 (**0.088**)	178 (0.224)	231 (0.291)	259 (**0.156**)	**140** (0.176)
Italy	78 (**0.090**)	**63** (0.226)	202 (0.298)	87 (0.192)	**55** (**0.119**)	179 (0.324)	90 (0.184)
Germany	**30** (**0.100**)	51 (0.247)	54 (0.151)	70 (0.249)	**45** (**0.197**)	46 (0.230)	91 (0.273)
Spain	125 (**0.121**)	88 (0.236)	133 (0.197)	**82** (0.183)	83 (**0.168**)	140 (0.273)	**81** (0.170)
France	**26** (**0.075**)	85 (0.239)	148 (0.216)	124 (0.161)	**104** (0.190)	150 (0.282)	153 (**0.170**)
Netherlands	**11** (0.131)	29 (0.298)	83 (**0.112**)	34 (0.220)	**32** (0.277)	—	45 (**0.241**)
Sweden	**11** (0.098)	34 (0.271)	35 (**0.082**)	32 (0.218)	**34** (**0.210**)	118 (**0.210**)	38 (0.228)
Portugal	**1** (**0.092**)	2 (0.176)	7 (0.186)	10 (0.260)	**3** (**0.174**)	10 (0.275)	12 (0.263)

[[Reviewer 2]] Thank you for your feedback. We will fix the typo in Table 1. ■ **Broader consequences:** We agree that the model can be used to analyze liberal/conservative lockdown policies in developing countries. In fact, Table C4 and Table C5 in the Appendix already present an analysis on how country features impact the effectiveness of lockdown. We have collected more data since the time of submission and will update and augment this analysis in the final manuscript. Moreover, the model can be used to conduct counterfactual analysis as shown in Fig. 3.

[[Reviewer 4]] Thank you for the excellent comments and valuable suggestions. We will include all the suggested references in the final version of the paper. We would like to clarify that our model was trained on the archived data capture of May 8; in the final manuscript, we will also add a robustness analysis to examine the model performance on subsequent data updates. ■ **Long-term forecasts:** We focused on 2-week forecasts to enable comparisons with all baselines as some of the benchmarks do not issue long-term predictions (e.g., the Imperial model). As shown in Updated Table 2, our model performs equally well when tested on 30-day forecasts; it provides the same patterns of accuracy gains achieved on the 2-week forecast. ■ **Evaluating uncertainty measures:** We evaluated the quality of our probabilistic forecasts in terms of the average continuous ranked probability score (CRPS) in Updated Table 2. Please also refer to Lines 8-11 of our response to Reviewer 1. ■ **Evaluation metric:** We apologize for the typo in Line 233—in the original submission, accuracy was evaluated on predicted *cumulative* deaths rather than *incident* deaths. This is why we were able to evaluate the accuracy of the weekly forecast by the CDC-ensemble. In Updated Table 2, we evaluate the performance of all baselines with respect to the mean *absolute* error in the predicted daily deaths, i.e., $\mathcal{E} = \frac{1}{T}\sum_{k=1}^{T} |Y_i(t+k) - \widehat{Y}_i(t+k)|$. We will release the code for reproducing Updated Table 2. ■ **Model specification:** We use a standard radial basis function (RBF) kernel with a variance (amplitude) parameter. The data $Y_i(t)$ is assumed to be normal. We will provide the precise expression of the the distribution of $Y_i(t)$ and expand the kernel parameter set in lines 122 and 140 of the revised manuscript. ■ **Ablation study:** Your description of our ablated baseline is accurate; we will clarify the details in the final paper. The benefits of hierarchical modeling are multifaceted: (a) policy heterogeneity across countries regularizes *factual* fits enabling better generalization on *counterfactual* inferences, (b) asynchronicity of the pandemic across countries enables better generalization *over time* for lagging countries, and (c) countries with similar features share the epidemic parameters. While it is hard to disentangle these effects analytically, we will add more ablated baselines with clusters of countries (with similar policies to the US, similar features to the US, and pandemic onsets synchronized with the US) removed one at a time to empirical assess these effects separately.

[[Reviewer 5]] Thank you for your feedback. We will fix the typo in Line 61. ■ **Model Inspection:** Table C4 and Table C5 in the Appendix already show the ranking of country features with respect to their impact on R_0. Based on your suggestion, we will move these results to the main manuscript given the extra space allowed in the final manuscript.