I would like to thank the editor for the opportunity to review this interesting paper and the authors for their nice work. First and foremost, the quality of figures should be denoted, notably the maps.

In its current state, I found that the study suffers from reporting loopholes and sometimes inappropriate or not sufficiently supported statistical choices. In light of this, I found that some authors' statements were too clear-cut.

- **Introduction:**
  - I found the introduction interesting but maybe a bit too long. I would suggest shrinking the conclusion and maybe including some of the § in the supplementary files for readers, interested in getting additional background.
  - At this step, it is not clear why the authors focused only on 2020 and did not consider subsequent years. Similarly, I did not get why they restricted the analysis to the Teesside. What is preventing authors from a larger spatial field?

- **Comments related to the 'Data collation' part:**
  - I found the description of the data sources clear and informative.
  - **Major concerns - Line 160 :** 'contemporaneously for both pillars (from week of year 19 to 53) and ran a correlation on the two pillars.' What do authors mean by 'ran a correlation on the two pillars'? Did they simply compute a correlation coefficient between the two time series? I assume both series are non-stationary both in mean and variance? In addition, there is probably a high level of temporal correlation between the data points. In this context simply computing a correlation coefficient is not appropriate and basic principles of time series analysis should be considered.

- **About model selection in the 'Disease mapping' part:**

In the 'disease mapping' part, authors compared a Poisson and Negative binomial regression within a Bayesian framework using the R-INLA package. Model selection was performed using the DIC.

The modern Bayesian approach instead recommends relying on Cross-validation. Cross-validation is a family of techniques that try to estimate how well a model would predict unseen data. This is especially appropriate in a Bayesian framework in which the goal is to model the full data-generating process. A must-read for a better understanding of this topic is: https://mc-stan.org/loo/articles/online-only/faq.html.

A classical assessment metric in this context is the CPO (conditional predictive ordinates) or the LOO-PD (leave-one-out log pointwise predictive density). This metric is defined for all observations in the dataset and measures the leave-one-out model predictive capabilities of the model. It is a special case of cross-validation in which only one observation is left out (leave-one-out cross-validation). These CPO/LOO-PD are then usually used along a logarithmic scoring rule to get a summary of the model's quality in terms of probabilistic forecasts, see https://doi.org/10.1198/016214506000001437.

The LOO and LGOO-CV can be routinely computed in R-INLA by specifying in the inla() call:

```
control.compute = list(
dic = TRUE,
cpo = TRUE,
```

```
waic = TRUE,
control.gcpo = list(enable = TRUE),
config = T
)
```

Because this is more consistent with the current recommended approach and as easy to implement as the DIC-based model selection, I recommend using the CPO along with a log-score approach, instead of the current one, if they really do want to stick with their current model selection.

I however don't think such a selection is relevant here. In the trade-off between Poisson and Negative binomial, two key elements are at stake:

1. The Poisson distribution constraints the mean to be equal to the variance, which might be a strong assumption and prevent to specify of a good generative model;
2. The Negative binomial distribution allows to relaxation this assumption by introducing potential overdispersion (Var(.)>E(.)) but at the cost of additional parameters which might however lead to overfitting;

It is known that the Negative binomial can naturally arise from a Poisson-Gamma model. In this model, the standard Poisson distribution appears as a limiting case.

R-INLA provides a class of priors (Penalised-Complexity prior, PC-prior) aiming to tame the risk of overfitting for nested models having a natural base model.
In the case of the Negative binomial: it penalises the departure from a Poisson distribution. For the BYM2 model: first departure from the case in which there is no heterogeneity and then departure from the case where the whole heterogeneity is not spatially structured.

Key references include: DOI: 10.1214/16-STS576. For the Poisson-Negative binomial case, the full derivation of the PC-prior and additional details are provided on https://dansblog.netlify.app/2022-08-29-priors4/2022-08-29-priors4.html by Dan Simpson, also provided in DOI: 10.1214/17-STS576REJ

I would therefore recommend to just stick with the Negative binomial distribution along with a PC-prior (which is the R-INLA default). If authors really do want to strongly penalise the deviation from the Poisson case, priors parameters might be changed accordingly (see https://inla.r-inla-download.org/r-inla.org/doc/likelihood/nbinomial.pdf). Note that doi: 10.2807/1560-7917.ES.2021.26.28.2001401, which you are referring to multiple times only used the Negative binomial distribution.

Critical reporting issues:

- **About priors in the 'Disease mapping' part:** Please specify in the main text or supplementary files the priors (distribution and values) for all parameters, notably the Negative binomial hyperparameter and the BYM2 latent effect.
- **About the adjacency matrix in the 'Disease mapping' part:** Please, specify in main text or supplementary files which adjacency matrix was used for the BYM2 latent effect. Did you build it using the Queen contiguity definition? (i.e., what definition of 'adjacent spatial units' did you use ?)
- **About R-INLA in the 'Disease mapping' part:**
    - Please specify in the main text or supplementary files the version of R-INLA that you used. This is especially important because R-INLA is a package with very active development.
    - Please specify in the main text or supplementary files how the results were approximated. Especially, to increase the accuracy, please use the 'Laplace' strategy.

        - Please, specify in the main text or supplementary files which integration strategy was used. Especially, note that the 'eb' strategy severely underestimates uncertainty and is therefore not recommended.
- **About the link used in the model:** The link used for the Poisson and Negative binomial distribution should be stated in the main text to avoid any ambiguity. I got from Figures and Supplementary that authors used the log link but it should be included in the methods part.
- **About the posterior distribution for the hyperparameters of the 'Disease mapping' part:** Please report in supplementary files the posterior distribution of all hyperparameters for the model (notably the two hyperparameters from the BYM2 model, which have an intuitive interpretation, as denoted in https://doi.org/10.1177/0962280216660421).
- **About posterior predictive check in the 'Disease mapping part':** In line with the view that the aim of a Bayesian model is to specify a *generative model*, a key step in the Bayesian analysis workflow is to compare the posterior predicted values against the observed one ('posterior predictive check, see https://doi.org/10.48550/arXiv.2011.01808 page 30). Please provide the posterior predictive check in the supplementary files. Note that with R-INLA, you must use inla.posterior.sample and inla.posterior.sample.eval to draw from the joint posterior distribution.
- Authors should be praised for using the BYM2 latent effect.

- **Minor elements in the 'Disease mapping part':**
    - Line 201: 'To help us define the spatial dynamics of COVID-19 in Teesside' this does not seem appropriate as your model does not include any interaction between space and time. As such, it cannot help you assess any 'spatial dynamic'. Please, consider changing 'dynamic' by 'heterogeneity'.
    - Line 208-209: 'We included the expected number of cases for each postcode district […]', am I right to assume that by 'included' you meant 'included as offset' ?
    - I would recommend a mathematical description of the full model **(including all priors)** in supplementary files to avoid any misunderstanding.

Other major concerns: What about sensitivity analyses (adjacency matrix+prior) ?

- **About the 'Mixed-effect modeling' part:**
    - Please, explain why you did not consider any spatially-structured prior for the postcode district. Especially, why did you not consider a BYM2 latent effect, which might reduce toward the I.I.D. case that you imposed on your model?
    - Line 239, the authors wrote: "which demonstrated negative binomial to be a superior fit for the data in all". Please, explain how you came up with this conclusion. Please report in supplementary files all the diagnostics provided by Dharma that you may have looked at.
    - Is there any reasons to assume a fixed 'effect' of time (i.e., to include time as a fixed coefficient ?). Why did you not consider penalised smoothing splines (which may in addition shrink toward the case you imposed) allowing the handle time in a more flexible way?
    - What do authors mean by 'random gradient for a week for each postcode district' ? Is it a time-specific intercept varying by postcode district? Are these intercepts temporally correlated (e.g., AR(p) or RW(p)) process? Is that a smoothing spline? I don't understand if there is any assumption about the space-time interaction? Please clarify the relation between space and time in your model using Knorr-Held's classical typology (https://pubmed.ncbi.nlm.nih.gov/10960871/). To avoid any ambiguity for

readers, please provide the mathematical counterpart of your model in the supplementary files.

- Line 235: 'To account for variation in symptom onset and testing delay after infection, we applied a two-week time lag to all temporal variables'. In the cited reference doi: 10.2807/1560-7917.ES.2021.26.28.2001401 the baseline analysis relied on a 7-day delay. Here, the authors used a 2-week delay. The choice of a 2-week delay seems very arbitrary without any theoretical support in the cited reference.
- Please make it clear that you used the log link for your model (this is specified in Figure 4 but should be specified in the method section)
- Authors should be praised for avoiding AIC-based variables selection, which would have introduced post-selection bias.

General comment about the method:

- Mixing both Bayesian and Frequentist approaches is very weird as parameters have fundamentally not the same nature in these two framework and methodologies are different.
- Authors restricted the spatial domain to Teesside, what about possible border effects ?

Results:

- Figure 4: ', asterisks indicate significance', please add the matching between number of stars and confidence level.

Discussion:

- **Line 462: '**Our study has demonstrated the effects of weather and government interventions on the spatio-temporal'. In the meantime, the authors wrote, line 362: '. These results must be interpreted with caution due to several limitations.'. 'Demonstrated' seems way too strong given the empirical approach and the data (plagued with reporting bias and other issues as denoted) on which the analysis relies on. This statement seems even more inappropriate given the lack of sensitivity analyses to priors and adjacency matrix.

  Note that many more limits could be raised regarding the analysis, among which imperfect ascertainment (false positive and negative due to tests' imperfect clinical performances), reverse causality (e.g., endogeneity issues between Tier and number of positive tests), unplausible underlying assumption that the observed variables are uncorrelated with the unobserved variables (see Mundlak's model). All these could bias the coefficients, which, again, call for greater care in model interpretation.

- After reading the whole study, I do not believe that labelling the study as 'spatio-temporal' is really appropriate. First, the disease mapping model does not include any temporal or spatiotemporal component (everything being aggregated over the entire period). Second, the GLMM model does not include any terms aiming to handle residual spatial autocorrelation or residual interaction between space and time. Spatiotemporal models can be broadly classified into 4 types, each assuming a type of interaction between space and time,
  see https://pubmed.ncbi.nlm.nih.gov/10960871/. Type IV models are usually the right type of spatiotemporal model when modelling infectious diseases spread in a spatio-temporal disease mapping context. The 'group' argument in the latent effect in R-INLA can help set such spatio-temporal model. Else, space-time interaction can be easily

handled through the INLAspacetime package: https://cran.r-project.org/web/packages/INLAspacetime/index.html.