The core models utilised by VIGIL are DISGAS and TWODEE. DISGAS is an advection-diffusion model that transports a passive scalar emitted from a steady source within a prescribed wind field, refined using the DIAGNO model. It disregards gravitational effects. TWODEE, on the other hand, is a shallow water model for transporting a gravity current that interacts with prescribed wind and topography. DISGAS is appropriate for diluted flows and strong wind conditions (Richardson number, Ri, much less than 1), whereas TWODEE is used when gravitational effects dominate (Ri much greater than 1). Intermediate regimes, which can exhibit behaviours divergent from these two models, are more complex. While the authors correctly introduce and discuss this point, they do not quantify the epistemic uncertainty introduced by these approximations. Comparing the two models in intermediate regimes can help quantify this uncertainty. The authors use DISGAS when Ri < 0.25 and TWODEE when Ri > 0.25. I recommend comparing CO2 gas concentrations obtained with TWODEE and DISGAS for a scenario with Ri ~ 0.25 and a wind field from W-SW to quantify the differences between the two approaches under challenging conditions

This comparison is crucial also because the persistence values along the valley shown in Figure 4 likely result mainly from TWODEE simulations. The Richardson number, used to decide which model to apply, is calculated at the source. However, dilution due to entrainment and turbulent diffusion tends to decrease the local Richardson number, making the current lighter. Additionally, the Richardson number is averaged over a whole day, meaning some hours with Ri < 0.25 are modelled with TWODEE, which stretches TWODEE's capabilities in diluted regimes. I recommend adding a comment in the main text to describe these approximations.

The reviewer is right about the possible change of regime as the gas flow dilutes downstream, although the domain under analysis is rather small.

It's unclear what the domain is being described as "small" in relation to. The significance of a regime change is determined by the dilution rate, and no domain is inherently "small": its size depends on the specific variability scale of the phenomena being modeled. In my view, the possible change of regime as the gas flow dilutes downstream, along with the model's approximations, should be explicitly addressed in the description of the methodology to provide clarity on how the domain size relates to the physical processes being studied.

To overcome this issue, one should use a more complex model which handles both regimes, which is computationally much more demanding than DISGAS and TWODEE2. Such a model would be significantly more demanding in terms of computational resources, hence making it difficult to apply for hazard studies like the one here presented.

I agree this point and, at my advice, a sentence like this one should be put in the text.

Concerning the approximation used by VIGIL in the intermediate regime, to answer to the reviewer criticism, we use TWODEE2 for all the cases in which Ri > 0.25. This is a cautious approach from hazard assessment point of view, since TWODEE2 generally results in higher near surface concentrations. First, the number of simulations in this regime is 277, which corresponds to 27.70%; therefore, we agreed that an analysis of the outputs produced by the two models in this intermediate regime was needed.

Ok.

Specifically, we selected three simulations with the following values of Ri, chosen to be equally spaced between 0.25 and 1: 0.438, 0.625 and 0.812. For each Ri value, we conducted the simulation using both DISGAS and TWODEE. From the outputs, with a Python code we calculated the RMSE between the DISGAS and the TWODE2 solution in the domain for each vertical level and each time step, then we calculated the average of all the RMSEs for each Ri case.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N}\left(C_{i,DISGAS} - C_{i,TWODEE}\right)^2}{N}}$$

Results are summarized in Table 1 and show that the error is negligible from hazard assessment point of view, especially for CO2, since it is lower than the background value and the typical maximum concentrations easily overcome 1000-10000 ppm. To be further conservative, we considered only the part of the domain where at least one of the two models computed concentration values above the background concentration used in the simulations (400 ppm). In this case we obtained larger RMSE values, but still not significant.

It is not correct to conclude that an effect is insignificant by comparing RMSE values with maximum values. Instead, RMSE should be compared with the mean value above the background concentration across both the entire domain and the reduced domain. This mean value should be calculated as:

$$MV = \frac{\sum_{i=1}^{N} C_{i,DISGAS}}{N} - C_{background}$$

These values should be included in Table 5 of the revised manuscript as a reference point for the RMSE values provided. Furthermore, since the hazard is presented as a map rather than a spatial average, a map showing the difference between the two models in either the worst-case or a representative scenario is necessary. This will help illustrate the potential epistemic error in the hazard maps as a function of location. This difference map should be generated at the same height at which the hazard maps were extrapolated. Additionally, the manuscript should specify the height at which the MV and RMSE were calculated.

Nonetheless, in future versions of VIGIL we will improve the way VIGIL automatically handles this intermediate regime, e.g., by introducing the option to calculate a Ri-weighted output from the DISGAS and TWODEE2 outputs.

We included these considerations in lines 197-200, 435-457,

Table 1. Results of RMSE calculations for the Richardson-dependency test.

Ri	RMSE All domain [ppm]	RMSE Reduced domain [ppm]
0.438	54.71	102.38
0.625	60.84	159.56
0.812	77.82	155.10

Another critical aspect needing better description is the effect of turbulent diffusion. In advection-diffusion models like DISGAS, turbulent diffusion is a key parameter determining *how much the flow dilutes and how gas concentration decreases with distance. This parameter is also important in TWODEE2, alongside entrainment. The authors should explicitly explain how these parameters are calculated in the two models and how turbulent diffusion depends on mesh resolution. I suggest adding a map (even in the supplementary material) showing the depth-averaged horizontal turbulent diffusion used by both models for scenarios with median, strong, and weak wind conditions, using the paper's resolution (3 m) and a refined resolution (1.5 m).*

Simulations with refined resolutions are also essential for quantifying the epistemic uncertainty due to numerical approximations in this specific application. Including the effect of resolution in a supplementary figure would be beneficial.

Even if the purpose of this paper is to use DISGAS and TWODEE2 to calculate the gas hazard and not to validate these models, whose details have already been published in a few papers reported in the reference list, we included further details on the turbulent diffusion parameterization used in DISGAS in lines 128-132. TWODEE2 only allows controlling the numerical diffusion with a coefficient, which was left to the default value of 0.2; in our opinion there is no need to go into this detail in this manuscript for TWODEE2. We believe that a thorough analysis of the turbulent (or numerical) diffusion modelling in DISGAS and TWODEE2 is outside the scope of this paper. However, on a selected simulation day (the one at Ri = 0.438 of the test mentioned above), we carried out two further simulations with spatial resolutions of 1.5 and 6 m, respectively. We then calculated the RMSE between the solution at the highest resolution (1.5 m) and the other two solutions (3 and 6 m):

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N}\left(C_{i,res} - C_{i,res\_1.5m}\right)^2}{N}}$$

Results of this analysis are summarized in Table 2:

Table 2. Results of RMSE calculations for the resolution-dependency test.

	DISGAS		**TWODEE2**	
resolution	**RMSE All domain [ppm]**	**RMSE Reduced domain [ppm]**	**RMSE All domain [ppm]**	**RMSE Reduced domain [ppm]**
3 m vs 1.5 m	12.02	33.10	64.79	267.02
6 m vs 1.5 m	23.09	58.13	130.73	361.82

Results show that overall, the RMSE obtained by decreasing the resolution increases but, although not negligible, is not significant. RMSE are higher with TWODEE2, but so are the maximum concentrations, therefore the two models behave similarly. Moreover, that stronger sensitivity of the results of TWODEE2 are likely due to the topographic control and not to the parameterization of the turbulence. Also, in this case to be further conservative we considered only the part of the domain where at least one of the two models computed concentration values above the background concentration used in the simulations (400 ppm). In this case we obtained larger RMSE values, but still below the atmospheric background value.

The same comment applies here as previously mentioned: the mean value (MV) and a representative difference map should be included and discussed in the paper. These additions will help provide a clearer understanding of the model's accuracy and potential epistemic errors across the domain.

All parameters used by the two models should be explicitly listed in a table to help readers understand the modelled conditions without needing to download the database and delve into the model configuration.

This is not possible as many of the parameters depend on the meteorological conditions, hence they change across the 1000 simulations. This is why we provided all the input data in the Zenodo repository.

Thank you for your clarification regarding the model's parameters. I understand that some of these parameters may vary significantly, and while this variability is important, I believe it would still be highly beneficial to present them in a table. For parameters with a wide range, simply reporting the range of variability should suffice. However, for parameters that are kept fixed, such as the numerical diffusion mentioned, it is crucial to include these explicitly in the manuscript. Providing a clear list of all parameters used in the simulation ensemble would greatly enhance the reader's ability to quickly understand the key assumptions and approximations underlying the model. This would also improve the transparency and reproducibility of the work.

Meteorological conditions used in the simulations should be presented more clearly. The spatial resolution is 30 km, correct? Where is the center of the cell used for this specific application located? What is the temporal resolution (I assume it is 1 hour)? I recommend *including this information in Section 3.2.1. Additionally, I suggest adding a figure (even in the supplementary material) showing an example of the vertical profiles downloaded from the cited datasets up to the height of the numerical domain. Highlight the value used for the quantification of the Richardson number and the Monin-Obukhov length scale, as well as the value used to draw the wind speed distribution shown in Figure 7 (at 10 m above the ground). How much variability does DIAGNO introduce to the original wind profile? I recommend showing the median and standard deviation of the vertical profiles obtained by applying DIAGNO to the example meteorological condition in the same figure.*

We improved the description of the approach used by VIGIL in Section 3.2.1; specifically, we explained that the ERA5 data that are the nearest to the computational domain centre are retrieved and then interpolated in the centre of the computational domain (line 177). Furthermore, we clarified that the meteorological data are at 1 hour resolution, as in the original ERA5 dataset (line 236). Since the Richardson number is calculated at the source, the Richardson number is calculated using meteorological data calculated by DIAGNO at 10 m above the ground. As far as the Monin-Obukhov length scale is concerned, this is calculated internally by DISGAS and TWODEE2 using the temperature at 2 m above the ground and at the soil in the centre of the domain coming from the ERA5 dataset.

We think we need to clarify the meaning of Figure 7. It is the domain-averaged wind speed at 10 m above the ground, so there is no information on the wind vertical profiles here. We made it clearer in the manuscript (lines 355-356). We think there is no need to show the vertical profiles of the ERA5 dataset and DIAGNO for the simple reason that in DIAGNO we restricted our analysis up to 500 m above the ground (which is what needed for these applications). Within the first 500 m above the ground one has only a couple of points in the ERA5 dataset, making a comparison of the vertical profiles hard and, in our opinion, not meaningful.

OK. I suggest explicitly stating that the regime is selected by calculating the Richardson number using the wind field at 10 m above the ground. Additionally, details regarding the data used to evaluate the Monin-Obukhov length scale should be included in the manuscript to enhance transparency.
Regarding the vertical wind profiles, I understand that Figure 7 represents data at a fixed height and that the available meteorological data have relatively low vertical resolution. However, for this very reason, it is crucial to comment on the variability of the original dataset. For instance, if the dataset includes only two vertical points, it would be useful to illustrate this variability using Rose diagrams. This would help explain how the limited data are translated into the much more detailed wind field produced by the DIAGNO model.
To improve clarity, one option could be to present Rose diagrams side by side: one for the heights where meteorological data are available and another for the wind field at 10 m above ground level, as generated by DIAGNO. Additionally, I believe it would be helpful, if possible, to show an example of a vertical wind profile from the DIAGNO model, overlaid with the two original points from the ERA5 dataset. This would provide a clearer view of how DIAGNO refines the coarser input data into a more resolved wind profile.

It is unclear to me how many simulations are performed, how the source distribution is sampled, and how much computational time is required. Specifically, it appears that 24 simulations for each of the 1000 sample days are performed to account for the variability of meteorological conditions, resulting in a total of 24,000 1-hour simulations with fixed *source conditions. How is the source variability taken into account? If the normal distribution introduced in the manuscript is sampled with, say, 10 points, the total number of simulations would be 240,000, correct? Please clarify these points in the main text.*

We run 24 hours-long 1000 simulations, we made it clearer in the manuscript (lines 240-250, 469-470).

OK. The fact that the gas emission rate is varied by randomly sampling it across the 1,000 simulations should be explicitly stated also in lines 240-250 to prevent any confusion. In its current form, Section 4.1 gives the impression that 1,000 days with different meteorological conditions were selected independently of the source conditions. However, as I now understand, this is not the case. Each day could potentially have a different source condition, randomly sampled as briefly described later in the conclusion. This raises a question about the stability of the results: if the same workflow were run again, different outcomes could emerge due to the random sampling. While I recognize the constraints imposed by computational resources and am comfortable assuming this effect is likely negligible, I believe this point should still be explicitly addressed in the manuscript. Clearly describing this aspect will help readers understand the inherent variability in the results and the model's robustness.

Specific comments

Web links: Numerous web links are present in the main text. Is it possible to move them to a specific reference section with the date of last access?

To our knowledge there are no prescriptions to authors about the URLs. Anyway we cut several URLs by including new references.

Zenodo Dataset: I was unable to access the Zenodo dataset.

Indeed there were problems with the link. We restored it.

OK

Section 2: The source data are from several years ago (Chiodini et al. 2010 and Rogie et al. 2000). Please add a comment on the expected variability at the source after 14 years (for the mass flow rate) and 24 years (for the composition).

We agree that a new measurement campaign would be useful, but it is quite difficult and dangerous since the high gas concentrations in the area. However, as recorded by the impacts on the local vegetation and historical chronicles, the Mefite area has been characterized by stable emission rates, similar to those dating back to the Roman era. Therefore, we do not expect a significant variation outside the range used in our work, which, in any case, fully include the statistical uncertainties estimated after the campaings by Chiodini et al. (2010).

OK. Please incorporate this reasoning into the manuscript for clarity.

Line 113: Is the momentum coupling one-way or two-way? Please specify explicitly.

It’s one way, we clarified this in the manuscript (lines 122-123).

OK

Section 3: Please include a comment about the potential chemical reactions affecting H2S during its transport and their time scales.

The main sink of H2S in the atmosphere is the reaction with OH radicals (e.g., Watts, 2000), other minor sinks can be found on a local scale, during and subsequent to rainfall events (Kristmannsdottir et al., 2000, Thorsteinsson et al., 2013) or under the action of lakes, soils, and vegetation (Bussotti et al., 1997; Cihacek and Bremner, 1990). However, these interactions typically do not have a first order control. For example, Olafsdottir et al. (2014) conducted ad hoc measurement campaigns in Iceland showing that the depletion of H2S from the atmosphere is insignificant compared to the emissions within a 35 km distance from the sources. Neglecting such reactions could imply an overestimation of the H2S concentration, probably not significant for our restricted domain. We clarified this point in the text (lines 255-264, 396-400).

OK, thanks.

Lines 147-151: The difference between Forecast and Reanalysis mode is not perfectly clear to me. Why can't the ERA5 dataset be used in Forecast mode and NCEP in Reanalysis mode? Please clarify the differences between these two datasets.

NCEP is a dataset that includes GFS daily forecasts for the next 384 hours. ERA5 is a reanalysis, therefore it cannot be used for forecasts (the corresponding dataset is called IFS: https://www.ecmwf.int/en/forecasts/documentation-and-support/changes-ecmwf-model).

OK

Line 176: ECDF is not defined when first introduced. It is defined later at line 186.

We fixed this in the manuscript (line 193)

OK

Section 4.1: The height of the numerical domain and the vertical size of the cells are not reported.

This is because there is no vertical domain and discretization along the vertical direction in TWODEE2, being a shallow layer model (apart from vertical levels used to print the interpolated outputs), whilst there is in DISGAS and DIAGNO and they don’t need necessarily to coincide. Anyway we provided necessary details in lines 243-244,

Ok.

Figure 3: Please add a box highlighting the emission area.

It is already shown in fig. 2c, we tried adding the box in the output figures and the resulting picture was not optimal in our opinion.

OK

Figures 3, 4, 5, 6, 8: Please indicate, for each figure, how many scenarios come from TWODEE2 and how many from DISGAS simulations. This information is essential to understand the regimes producing hazardous conditions.

This was clearly stated in lines 250-251, in our opinion there is no need to repeat this information in the figure captions and descriptions.

OK.

Line 285: Change "finds blowing" to "winds blowing."

We fixed this in the manuscript (line 327).

OK

Figure 5: It is not clear whether the curves represent 24-hour averages or hourly results.

It's 24-hour averages, we clarified this in the manuscript (line 358, 369).

OK

Figure 4 and Comments in the Text: The method for calculating the persistence maps is not completely clear to me. For example, when you say CO2 persistence > 5000 ppm for 8 hours, does this mean selecting (cell by cell) all the 24-hour scenarios with 8 consecutive hours satisfying the condition? If the condition is satisfied for 4 hours, then concentration decreases for 1 hour and then increases again for 4 hours, do you keep or discard the scenario?

The persistence is defined as the probability to overcome a certain concentration threshold (in your example, 5000 ppm) for a duration (consecutive hours in this application) of at least N hours (in your example, 8 hours). In your example, the simulation would be discarded in the calculation of the persistence.

OK.

Figure 7: Link the figure to the discussion at line 225.

This request is not clear.

I believe Figure 7 could be highly useful for readers attempting to understand how many scenarios fall into one regime or another. To enhance clarity, I suggest adding a reference to Figure 7 in support of the discussion around line 225. This would help readers better visualize the distribution of scenarios across different regimes. Additionally, I would like to point out that the reasoning around line 225 is based on the mean gas emission rate. Please ensure this is made explicit in the text to avoid any ambiguity.

Line 387: It is unclear how a lower estimate in a probabilistic hazard map (Figure 4def) can be described as "safe."

We agree, the term “safe” is misleading. We removed it.

OK

Lines 402-403: It is not specified how many points were used to sample the source variability.

For each simulation, a value of the source emission rate is sampled from the normal distribution. We made it clearer in the manuscript (line 466).

OK

Lines 404-406: State explicitly that this approach disregards intra-day variability of the Richardson number.

We fixed this in the manuscript (line 468).

OK.