General comments

The authors present observation system simulation experiments of regional CO_2 concentration data assimilation and analyze the impact of different station networks on the forecast quality. The work provides interesting insights for high-resolution, regional CO_2 forecasting based on in-situ observations. These insights deserve publication.

However, the discussion of the results should be extended with the aim of understanding the implications for other setups and possibly for forecasts with real observations. I do not expect a long discussion of OSSE in general, and I do see that the authors discuss and evaluate some important aspects of their setup, e.g., using the rank histograms. But I miss an overall discussion which aspects of the setup the authors deem crucial for the interpretability and transferability of the results. Just to provide an example, the construction of the true state in the OSSE might be a relevant aspect (see my comments concerning line 233). Ideally, I would expect a concise discussion of these aspects in the conclusions section.

The setup is in general well structured and well explained, aside from a few remaining questions listed in the specific comments. The presentation of the results is well understandable, but parts of the results could be presented more concisely (see my comment concerning lines 468–484).

Specific comments

line 139, beginning of section 2.3.2 In section 2.3.1, the authors define the self-sensitivity, which is defined in observation space. In section 2.3.2, a similar vector notation as for the observation space is used for the state space. A brief comment on the structure of the state vectors ($\boldsymbol{x}^f$ and $\boldsymbol{x}_t$) could guide the reader to immediately see this difference.

line 155, Eq. (13) The definition of $\delta\mathbf{y}_0$ is missing. I assume that $\delta\mathbf{y}_0 = \mathbf{y} - \mathbf{H}\bar{\mathbf{x}}_b$.

line 233, emissions in OSSE The choice of emissions for the OSSE true state and for the DA experiments is mentioned in sections 2.4.1 and 2.4.2. This choice is important to obtain meaningful results, as indicated by the authors when mentioning the identical twin problem (line 234). My impression is that the comparison of the emissions chosen for generating the true state and those use for the DA experiment deserves more attention.

The importance of the choice of emissions is illustrated by the following interpretation of the results: The simulation of the true state and the four DA experiments used VPRM and will therefore show very similar (or identical?) biogenic fluxes. The authors find that observation sites in regions with strong biogenic fluxes greatly reduce forecast errors (line 543). Is this because the fluxes in these regions were close to the true fluxes by construction? Or do the authors expect a similar improvement in an experiment with real observations?

I suggest to state explicitly how the emissions in the true state and in the DA experiments differ. If the authors agree on the relevance of this choice of emissions for the results, this aspect should be mentioned when presenting or discussing the results. (see also my general comments above)

line 191 The authors write: "The CO_2 variability of the true state in this study was included in the CO_2 variabilities of CT2022 and the Copernicus Atmosphere Monitoring Service (CAMS), ..." It is not clear to me how CAMS is used and how the variabilities of CT2022 are considered to create a reasonable deviation of the DA experiments from the true state.

line 256 The authors list the meteorological and chemical initial and lateral boundary conditions for the experiments without distinguishing between the single forecast run for the true state and the ensemble forecast for the DA experiments. Were the lateral boundary conditions identical

identical for all model runs? If yes, do the authors expect an underestimated deviation from the true state at the lateral boundaries that could influences the results?

How did the authors make sure that the spread of the meteorological ensemble remains approximately constant? Was an initial condition update cycling or similar technique used?

line 354 (minor comment) The authors provide many values with the unit $10^4\,\mathrm{ppm}^2$, to be interpreted as the sum of squared errors in all grid cells. A (very) brief guidance on how to interpret these numbers could possibly help the reader in section 3.3.1 and in Table 5.

lines 354 and 361 (minor comment) The authors provide many numbers and even lists of numbers in the text. I do not want to criticise what is a matter of style, but instead of reading lists of 8 numbers as in lines 354 and 361, I personally prefer the structure of a table. The paper contains many tables already and some of the numbers provided in the text could be included in the existing tables, e.g., by including averages in Table 5.

lines 424 to 484 (minor comment) Section 3.3.3 mainly discusses three separate aspects: The correlation of EFSO with self-sensitivity and CO_2 concentration variability; the impact of vegetation types around the observation sites; and the fraction of beneficial observations. For the reader it might be easier to follow if section 3.3.3 is split into two or even three parts (e.g. start a new section 3.3.4 on line 468 for the fraction of beneficial observations).

lines 468 to 484 (minor comment) This paragraph could be written more concisely. For example, the following sentences seem redundant: "The average fractions of beneficial observations ... were 68.9% and 66.3% ... On average, more than half of the observations contributed to reducing the forecast errors. ... Therefore, on average, more than half of the assimilated surface CO_2 observations in the four experiments contributed to the reduction of forecast errors."

In general, the presentation of the results is clear and understandable. But as the example shows, it could be more concise in a few paragraphs.

Technical corrections

line 104, Eq. (4) I do not understand how Eq. (4) leads to $A^\top$ in Eq. (3) (instead of just A).

Fig. 1 The figure is very illustrative, but on the horizontal axis, the label "t h" seems odd to me since t is usually considered a value including a unit.

line 284 I was confused by the word "multiple", which is meant here in the sense of "too many".

line 381 "forecast reduction" should probably be "forecast error reduction"

lines 481 f. Sites "A1, A4" should probably be "V1, V4"