General comments:

Atmospheric correction is an essential part of satellite remote sensing of land surface. Yin et al. describe and evaluate the Sensor Invariant Atmospheric Correction (SIAC) algorithm for atmospheric correction. The existing atmospheric correction methods can be improved and therefore further development of the algorithms is welcome.

SIAC is capable of carrying out atmospheric correction both for Sentinel 2 and Landsat 8 satellite data. Furthermore, SIAC is a Bayesian (statistical) algorithm so it can take advantage of prior information and it produces uncertainty estimates for the surface reflectances produced. The algorithm is tested and validated with ground-based AERONET and RadCalNet data. There are no significant steps taken in new method development in this work, but SIAC combines well existing methods. The results shown in the manuscript show that SIAC is capable of carrying out good quality atmospheric correction.

The selections made in the development of SIAC are mostly well justified and based on previously published literature. It is very good that the authors have shared the codes for others to be used.

My main criticism is in the presentation. The quality of presentation in the manuscript varies. Time to time the text is well written and smooth but quite often the text is difficult to read. The manuscript is quite long and structured so that first SIAC and results are explained in general terms followed by the Discussion and Conclusions, and then all the technical details are mostly included in the Appendices. As this is a method development manuscript, I find this a bit difficult for the reader as it is needed to browse back and forth while reading. Furthermore, the manuscript heavily relies on use of acronyms and symbols. It is good that the main symbols are explained in a table but this also makes the manuscript very slow to read, especially for a person who is not that familiar with the field. I would strongly recommend the authors to think the use of acronyms (even single letter acronyms) and symbols, and possibly shorten and re-organise the manuscript for improved readability. Also, there were some typos in the text so proofreading is recommended. Most of the font sizes in figures are too small and very difficult to read.

This paper describes the theoretical basis, practical implementation, and the validation results, which inevitably makes the paper long and relatively complex. We have decided to put implementation details and further testing in appendices to have a more readable main paper by a wider audience. We have added some text clarifying that the appendices are there to provide detail and additional testing towards the end of Section 2.2:

> *In this paper, we present the theoretical underpinnings of the method and major results, relegating details of the implementation and additional results to the appendices.*

Regarding acronyms, the ones used are in common use throughout the literature. Symbols are described in Table 2 and used within the detailed implementation sections of the paper, so we assume the reader is familiar with the paper and overall method at that stage.

We have updated the font size in the figures to improve readability.

Specific comments:

l.15 Abstract mentions efficient emulators. However, in the text this was a bit unclear how these were used. Could it be clarified?

l.129 In the abstract this was probably mentioned as "statistical emulation", neural networks are not really statistical emulation.

> We have removed the qualifier "statistical" from the abstract and modified the text to be clearer around L129:

> *Running the model atmospheric model many times is computationally costly, and is often approximated by using e.g, look-up tables. Here, we provide fast surrogate approximations to the full atmospheric model, called emulators (Gómez-Dans et al., 2016). These approximations are based on fully connected artificial neural networks (ANNs) and provide an estimate of the pi terms as a function of the model inputs Xac . Additionally, the Jacobian of the atmospheric model (needed for efficient gradient descent minimisation and for uncertainty propagation) is also approximated by the emulator making use of backpropagation techniques (Hecht-Nielsen, 1992)*

l.53 AERONET is based on remote sensing, not in-situ measurements.

> Changed to "AERONET estimates" in L203

l.83 ".., we calculate:". "Calculate atmospheric parameter estimation"?

Thanks for pointing this out; it has been changed to:

SIAC comprises two major steps:

l.84 The list is difficult to follow.

We modified the list to be a bit more readable.

l.133 "This may cause errors". Can you estimate the significance of these errors

We have changed the text to be clearer:
This choice of aerosol type may cause errors when conditions strongly depart from it, such as situations dominated by urban, maritime or biomass burning conditions. See (Tirelli et al., 2015; Shen et al., 2019) for analysis of the impacts of aerosol types.

We have also added a clarification of this in Sect 5.3 ("Future developments"):
In this paper, the atmospheric composition is set by a model (6S in this case), and by a choice of aerosol optical properties (continental aerosol model). The use of emulators of the RT model makes it easy to change the RT model entirely in the code, or to use a different configuration of the model used. We can also extend the scheme to retrieve independent aerosol species concentrations by both modifying the RT model (and thus extending the number of parameters that go in the inference), and by using data on species distribution available from CAMS and extending the prior to cover these. A similar approach has been implemented in the MAJA processor (Rouquié et al., 2017), which uses the CAMS aerosol species data to define the aerosol types for the atmospheric correction and has found improved atmospheric correction results over deserts. This approach may well be valuable in areas of high dust aerosol loading, or in situations where biomass burning results in an important contribution to aerosol concentrations.

l.169 "Matrix D" was this defined?

It is defined as first-order spatial difference constraint, and the matrix is given in the updated version of the paper (Eq. 7)

l.183 "We assume that mean atmospheric parameters...are constant..." Can you estimate the significance of this assumption?

We have added support for this statement:

We calculate the pa,b,c with the mean atmospheric parameters x and the auxiliary data (Ozone and elevation) at MODIS spatial grid Gc. A linear interpolation is then used to re-sample the pa,b,c to the target sensor grid Gm, which is then used to derive the mean surface reflectance r with Eq. 8. The simple linear interpolation method used to resample the pa,b,c to sub-MODIS scale is justified as atmospheric parameters are known to exhibit much larger correlation lengths (100s of km) (Anderson et al., 2003; Chatterjee et al., 2010).

l.221 Can you give an example of "other artifacts"

We have added a list of likely artifacts:

[...](such as saturation of pixel value, cloud shadow, modelling error from the RadCalNet surface reflectance to TOA reflectance, etc.)

l.223 Why the tolerance of 10% was selected? How this filtering affects the evaluation of the results?

We have added a justification in the text:

5% is used as the target uncertainty of the S2/L8 TOA reflectance and the RadCalNet TOA uncertainty is around 2-5% for non-absorption bands (Wenny and Thome, 2022), which lead us to choose 10% as the threshold to filter out bad samples

l.300 "corrected to R", what is R?

R is surface reflectance (changed in the text)

l.333 Why different TCWV gamma values are used for S2 and L8?

We have pointed out that this is discussed in one of the appendices:

We use γ values of of 5 for S2 and L8 AOT, 5 for S2 TCWV and 0.1 for L8 TCWV in SIAC. Cross-validation studies suggest that there is a wide range suitable values for γ (Sect. G for additional details on this choice and its implications).

p.23 Figure 13. What are the colors of the bars?

p.26 Figure 15. What are the colors?

We have added a clarification on the boxplots colors

Boxplot colors are the same as colors in Fig. 2.

l.374 Is IQR defined?

It has been updated as the interquartile range (IQR) with the reference added.

l.375 Validation of surface reflectance uncertainty was a bit unclear. Should be clarified.

We have added a summary statement to this paragraph.

In summary, our tests show that we have a conservative estimate of uncertainty. The overestimation of the variance in surface reflectance is more marked for L8 than S2.

l.390 In many sentences (especially on page 25) you use "that" & "this" and it is unclear to which word these are referring to

We have updated the text to be clearer in that paragraph.

I reviewed the manuscript years ago submitted to another journal + I did look forwarding to the revision which I did not see. Now I saw it in a different journal with significantly improvement in the experimental design and presentation style.

I recommended for publication as the study

- presented a novel way for medium resolution (Landsat-8 and Sentinel-2) satellite data atmospheric correction
- showed the presented method significantly improves the surface reflectance and aerosol property retrieval accuracy
- systematically validated the presented method using Landsat and Sentinel-2 data over globally distributed AERONET

Impressive work and I have to say this is a hard paper to review as there are many techniques used in the study. I appreciated the authors make the codes public available. I like many ideas in the paper but in particular the spatial smoothness prior and the validation of the uncertainty (many geostationary satellite Bayesian AOD estimation output uncertainty without validation). I really hope the authors later (not in this manuscript) can experimentally show that the proposed method is better than LaSRC and Sen2Cor – by running LaSRC and Sen2Cor software on the datasets used in this study (this maybe boring to the authors but I believe it is beneficial to the community and to the authors if the algorithm more convinced). Or simply participate ACIX and hope to see the algorithm will be an unambiguous winner in ACIX.

I have a few comments before the publication of the manuscript

SIAC has been part of the ACIX-II and comparisons will be published soon. The results from SIAC compared in the ACIX-II were submitted more than 2 years ago and many updates used in the current paper had not been implemented by that time.

Major issues

1, The readers may wonder what is the computational cost of the proposed method and the iteration nature of the gradient descent (L-BFGS-B algorithm) may boost the computation time rapidly. Add a paragraph for discussion.

We added a line at the end of Section 5.1:

The entire process (this includes state estimation, surface reflectance estimation, cloud masking and uncertainty quantification) of a single S2 or L8 scene on a standard workstation with a 3.1 GHz i7 CPU and 16 Gb RAM, takes between 20-30 minutes on a single process.

2, The authors omitted the deep blue bands for both Landsat-8 and Sentinel-2 and the band is most sensitive to aerosol among all the bands, any particular reason for doing so. MODIS does not have this band? use MODIS blue to replace not work ?

Discuss the possible shift from MODIS to VIIRS considering the dying of MODIS.

We added a comment on these two points in Sect. 5.3 ("Further developments"):

SIAC relies on the surface reflectance constrain from the MODIS MCD43 product, but the current MODIS satellites are approaching the end of their mission (Skakun et al., 2018; Xiong and Butler, 2020). VIIRS is designed to produce a continuity data record to MODIS, so it may be possible to use the VIIRS VNP43 product (Justice et al., 2013) as an alternative surface constraint, but this has not been tested. VIIRS land products additionally include bands within the deep blue spectral region (M1, M2), which are more sensitive to the aerosol variation. This information may be used to constrain the S2/L8 deep blue bands to retrieve AOT over bright surfaces (Hsu et al., 2013).

3, Spatial smoothness is interesting but unclear mostly related to how D matrix is derived. Is the spatial difference applied to the intermediate AOD and WV results derived for each iteration of L-BFGS-B algorithm. If so, Cxb is changing with iterations? The spatial difference is calculated over neighbor 8 pixels or over an entire 40 km window? How the row and column direction are averaged? Simply give an example of deriving D would solve my concern. After the line of 775, "over the whole of Gc" or of sub-40 km? Looks like the sub-40 km is used as a block to solve the equation which means the solution of each sub-40 km is independent to the neighbor sub-40 km block (but I am not sure).

We have tried to clarify the concept prior spatial correlation by rewriting parts of Section 2.4. We hope it is clearer (after L166):

This gives a constraint based on a background (a priori) estimate of atmospheric state, Xb. In SIAC, the prior mean comes from the European Centre for Medium-Range Weather Forecasts (ECMWF) CAMS Near-real-time services (Morcrette et al., 2009; Benedetti et al., 2009) for estimates of atmospheric composition parameters AOT at 550 nm, total column water vapour and total column of Ozone in Xb and Xac. These are at a coarse spatial resolution on a 40 km grid, but we need Xb on a 500m grid to match with rb, so the data are interpolated to 500 m resolution.

The role of Cxb is to encode the prior expectation of variance and spatial correlation of the AOT and TCWV fields. AOT and TCWV variances are reported in the CAMS global validation report (see Sect. B2 for a detailed derivation).

We expect the AOT and TCWV fields to have long correlation lengths (Anderson et al., 2003), which result in non-zero off-diagonal elements in Cxb . This spatial correlation structure is defined using a Markov process covariance after Rodgers (2000). This has two free parameters, the variance σ^2 xb and relative length scale. Since the covariance appears in the constraint Eq. (5) in inverse form, we use a fast approximation this is derived from Rodgers (2000) and implemented as a first-order spatial difference constraint defined in matrix D.

We have also added a small summary towards the end of the Section describing the effect of these choices, as well as a definition of the D matrix:

The matrix D is given as an example for a variable length of <u>n</u>:(7)

$$
\begin{bmatrix}
-1 & 1 & 0 & 0 & \cdots & 0 \\
0 & -1 & 1 & 0 & \cdots & 0 \\
0 & 0 & -1 & 1 & \cdots & 0 \\
\vdots & \vdots & \vdots & \ddots & \ddots & \vdots \\
0 & 0 & 0 & \cdots & -1 & 1
\end{bmatrix}_{(n-1)\times n}
$$

Having this prior inverse covariance matrix allows a flow of information from regions in the scene that are well constrained by observations, to areas that are poorly constrained or missing.

The study need a consistent check of the writing style. For example, lines 102-103.

Line 105, how to treat b09. Is MODIS band 2 reflectance (after D2 transformation) used as a prior for b09 reflectance?

Changed the text to accommodate both comments:

The output of the spectral mapping provides an estimate of reflectance from 400 to 2400 every 1. This provides a surface reflectance estimation for S2 B09, a band strongly sensitive to TCWV, even though MCD43 does not include this spectral region. Uncertainty from the spectral mapping is explicitly treated in the SIAC framework.

Line 675, there is another MCD43 PSF estimation paper worth citing (Che et al. 2021).

Che, X., Zhang, H. K., & Liu, J. (2021). Making Landsat 5, 7 and 8 reflectance consistent using MODIS nadir-BRDF adjusted reflectance as reference. Remote Sensing of Environment, 262, 112517.

Added.

Line 130, not quite familiar with the L-BFGS-B algorithm. But does the L-BFGS-B algorithm have its own way to calculate the gradient or the authors really used the back propagation of ANN to calculate ANN (ANN uses an automatic differentiation method) to feed into L-BFGS-B algorithm. Or back prorogation gradient is only used in B4 for RBF uncertainty but not used in L-BFGS-B?

We have added some text to describe the use of emulators as providers of the Jacobian of the cost function:

The cost function and its partial derivatives exploit the ability of the emulators to provide accurate approximations to the atmospheric RT model and its partial derivatives (Gómez-Dans et al., 2016).

Line 140, the LaSRC cannot be considered as DDV method anymore as each single pixel on the global have a visible/SWIR ratio (not just DDV pixels) see Vermote et al. 2016.

Yes, LaSRC is a variant of it, but strictly speaking it cannot be treated as DDV. So, we have removed it from the list.

Line 155, the framework can tolerate incomplete coverage of Y^, how? If incomplete Y^, then the incomplete inversion of aod and wv will be spatially interpolated? Or get value using spatial smooth prior.

This is the role of the spatial correlation in the prior pdf (Sect 2.4). We have already expanded on this as a response to another comment, and hope it is clearer now.

Line 200, "a estimates" – grammar

Changed

Line 260, how the σxaero is set

We have added a reference to Sect. 3.2, where these magnitudes are defined.

Line 310, which TOA band, the blue band should definitely greater than 10% mostly

We disagree with this comment. The filter is applied to each band of the 6 compared bands. Those two sensors agree better than 5% for most of the time even the blue and deep blue bands, please refer to paper mentioned in the previous line (309). Another paper by Helder, et al., (2020): https://doi.org/10.3390/rs10091340

Line 330 we are already in Section 4.3

Removed this line

Please label and re-arrange figures 10-12 correctly. Some are clearly surface but labelled as TOA. Check the labels of the plots. Rearrange them so that the TOA spectra and TOA plot are at the same row.

After checking figures 10-12 (11-13 in the new version), they are in the correct order. You may say that the surface reflectance cannot have a better statistic than the TOA, but it can happen as the surface reflectance has a better contrast and stretches the reflectance values to a wider range than the TOA reflectance. We have added some clarification on this:

The fact that sometimes the BOA statistics are better than the TOA is probably due to the better dynamic range of the surface reflectance signal compared to the TOA one.

Line 470, yes, in many Landsat calibration literature, the calibration accuracy is 3%. Should this linked to appendix H?

Yes, this has been linked to Appendix H.

Line 580. There is no Schaff 2021 in reference list

It is in Line 1110.