Supplement of Earth Syst. Sci. Data, 13, 5273–5292, 2021
https://doi.org/10.5194/essd-13-5273-2021-supplement

Supplement of

The global and multi-annual MUSICA IASI {H_2O, δD} pair dataset

Christopher J. Diekmann et al.

Correspondence to: Christopher J. Diekmann (chjdiekmann@gmail.com) and Matthias Schneider (matthias.schneider@kit.edu)

User Guide for the Level-2 and Level-3 datasets of the MUSICA IASI {H_2O, δD} pair product

Christopher J. Diekmann, Matthias Schneider,
Farahnaz Khosrawi, Benjamin Ertl

Institute of Meteorology and Climate Research,
Atmospheric Aerosols and Remote Sensing

29.09.2021

Version 1.0

Document version	Date	Specification
0	13.07.2021	Initial draft
1.0	29.09.2021	First finalized version, used as supplement for Diekmann et al. (2021a)

Table of Contents

1 Introduction

This document serves as user guide for the datasets of the MUSICA IASI {H_2O, δD} pair data as presented in Diekmann et al. (2021a). It comprises the description of the output files of the Level-2 (L2) {H_2O, δD} pair post-processing (Section 2) as well as the re-gridded Level-3 (L3) {H_2O, δD} pair products (Section 3). The datasets are publicly available via the Digital Object Identifier (DOI) 10.35097/415 for the L2 {H_2O, δD} pair dataset and 10.35097/495 for the re-gridded L3 {H_2O, δD} pair dataset.

These datasets evolved from the MUSICA project (*MUlti-platform remote Sensing of Isotopologues for investigating the Cycle of Atmospheric water*, Schneider et al., 2016), whose aim was to create calibrated tropospheric profiles of water vapour H_2O and its isotopic composition δD (given in permille, ‰)

$$\delta D = \left(\frac{HDO/H_2O}{R_{VSMOW}} - 1\right) * 1000 \tag{1}$$

$$\text{with } R_{VSMOW} = 3.1152 * 10^{-4}$$

for data from ground-based and spaceborne remote sensing data. In this framework, a retrieval processor was developed for retrieving {H_2O, δD} pairs together with interfering trace gases from infrared spectra measured by the IASI sensor (*Infrared Atmospheric Sounding Instrument*, Blumstein et al., 2004) onboard the polar-orbiting Metop satellites from EUMETSAT. Building on that, further post-processing steps focusing on individual target species were established and concatenated in a quasi-operational processing, which was optimized for the use of high-performing cluster systems (see Figure S1). In that way, MUSICA IASI {H_2O, δD} pair data can be generated efficiently on a daily and global basis. Detailed information about the MUSICA IASI processing chain can be found in Schneider et al. (2021a). All output data are stored in netCDF4 files, which are compliant with the CF metadata naming conventions (Version 1.7, https://cfconventions.org/).

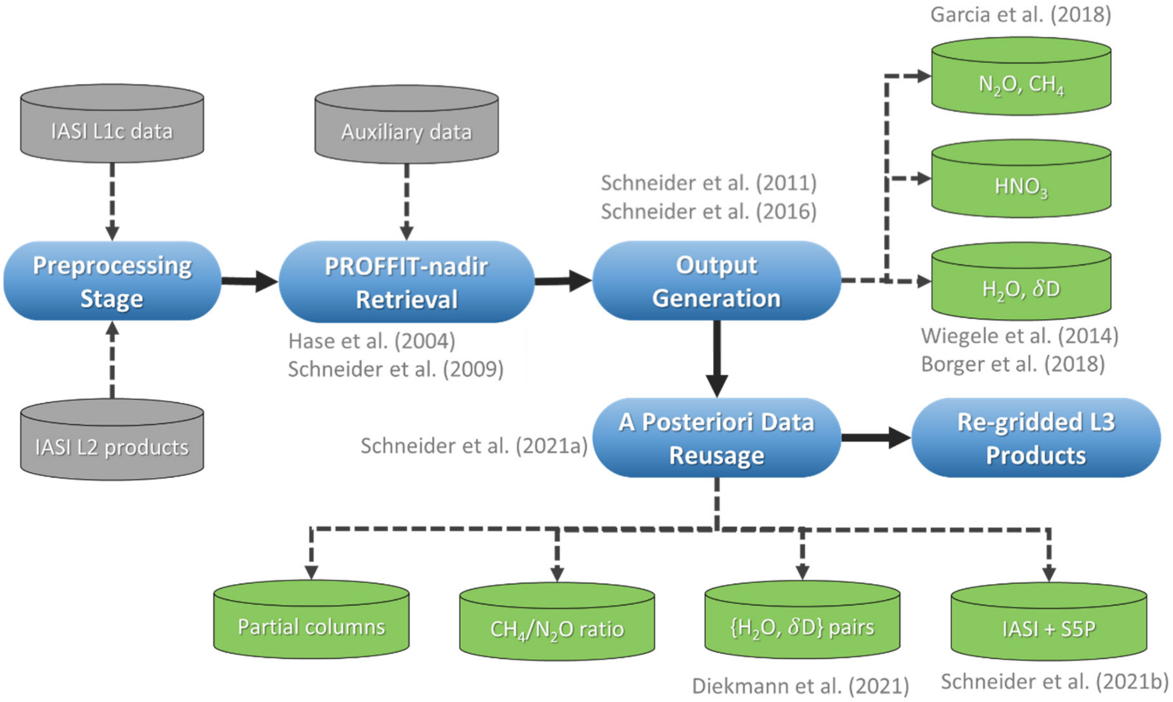

Figure S1 *Flow diagram of the different components of the MUSICA IASI processing chain as described in Schneider at al. (2021a).*

This user guide addresses the data structure and recommended usage of the output files of {H_2O, δD} pairs created within the L2 post-processing step "A Posteriori Data Reusage" as well as the "Re-gridded L3 Products" from Figure S1. In order to support and facilitate the use of the data, in this user guide some minimal code examples written in Python3.6 will be provided. For completeness, the common preamble used for the code examples is:

Listing 1 *Preamble for Python3.6 code snippets used throughout this document.*

```
# read modules
import xarray as xr
import numpy as np

# read iasi file
file_l2pp_iasi = '<path_to>/<musica_iasi_l2pp_file>.nc'
ds = xr.open_dataset(file_l2pp_iasi)
```

Accordingly, this document will stick to the zero-based indexing convention of Python, i.e. the first index of a list or array is 0.

2 Level-2 Post-processing {H_2O, δD} Pair Data

This section describes the output files of the L2 {H_2O, δD} pair post-processing, which is documented in terms of methods, data quality and spatio-temporal representativeness in Diekmann et al. (2021a). Here, an overview will be given about the output filenames (Section 2.2), the main product (Section 2.2), dimensions and variables of the output files (Section 2.3) together with some guidelines and recommendations for the use of these data (Section 2.4).

2.1 Output filenames

The {H_2O, δD} pair data are sorted into global and daily files, with considering a further separation into morning (local overpass around 09.30 am) and evening (local overpass around 09.30 pm) files for each day. The file naming convention is as follows:

`'IASI<multipleS>_MUSICA_<V>_L2pp_H2Oiso_<VPP>_<YYYYMMDD>_<OP>_global.nc'`

- `<multipleS>`: information about the considered sensors
    - Files between 2007-07-10 and 2013-02-19: `'A'`
    - Files between 2013-02-20 and 2019-10-23: `'AB'`
    - Files from 2019-10-24 on: `'ABC'`
- `<V>`: MUSICA retrieval processor version
    - Files until 2019-06-30: `'030201'`
    - Files from 2019-07-01: `'030300'`
- `<VPP>`: MUSICA IASI {H_2O, δD} pair post-processing version
    - All files: `'v2'`
- `<YYYYMMDD>`: day of observation (universal time format)
- `<OP>`: overpass information
    - Files with data with local overpass times around 09.30 am: `'morning'`
    - Files with data with local overpass times around 21.30 am: `'evening'`

2.2 Product Specification

The L2 {H_2O, δD} pair product is based on the MUSICA IASI retrieval from Schneider et al. (2021a), which processes the EUMETSAT IASI Level-1 (L1) spectral measurements and uses (and processes) some EUMETSAT IASI L2 products (such as temperature profiles, cloud cover and land surface information). The retrieval processer of Schneider et al. (2021a) provides independently retrieved H_2O and δD profiles, their a priori settings, retrieval constraints and error covariances. In fact, their retrieval processor performs an optimal estimation retrieval of the water vapour proxy state vectors wv_1 and wv_2:

$$wv_1 = \frac{\ln(HDO) + \ln(H_2O)}{2}$$
$$wv_2 = \ln(HDO) - \ln(H_2O) \qquad (2)$$

which are found to constitute reliable proxies for variations in H_2O and δD (Schneider et al., 2012). Already during the MUSICA IASI retrieval, HDO is normalized with R_{VSMOW}. Therefore, by combining the retrieval results of the two proxy states wv_1 and wv_2, profiles of H_2O and δD can be extracted:

$$H_2O = \exp\left(wv_1 - \frac{wv_2}{2}\right)$$
$$\delta D = (wv_2 - 1) * 1000 \qquad (3)$$

The aim of the subsequent L2 {H_2O, δD} pair post-processing is to harmonize the vertical sensitivity (i.e. averaging kernels) of wv_1 and wv_2 and in that way to create an optimal estimation product for {H_2O, δD} pairs, such that H2O and δD can be jointly interpreted (for details about this harmonization step please refer to Schneider et al., 2012, Wiegele et al., 2014, Barthlott et al., 2017 and Diekmann et al., 2021a). Therefore, analogous to the MUSICA IASI retrieval, also the L2 {H_2O, δD} pair post-processing internally requires the consideration of the proxy states wv_1 and wv_2. For the final output, the post-processed wv_1 and wv_2 states are included as reference in the netCDF4 files, but are additionally transformed to profiles of H_2O and δD according to Eqn. (3), which represent the final {H_2O, δD} pair product as main output of the L2 {H_2O, δD} pair post-processing.

2.3 Dataset Documentation

This section serves as technical documentation of the output files generated with the L2 {H_2O, δD} pair post-processing.

2.3.1 Dimensions

Table 1 lists all dimensions included in the output files of the L2 {H_2O, δD} pair post-processing.

The dimension `observation_id` is the main observation identification dimension for the output files of the L2 {H_2O, δD} pair post-processing. For each individual file, it starts from 0 and counts all observations that are included in the corresponding file.

Various additional dimensions are required for storing variables that describe properties and diagnostics of the water vapour retrieval. As discussed in Section 2.2, the MUSICA IASI retrieval considers the water vapour proxy states wv_1 and wv_2 as retrieval state vectors. To distinguish between those two states for variables that are given in the water vapour proxy state base and therefore include values for both wv_1 and wv_2, the dimension `musica_species_id`, is introduced. For instance, this dimension is used for the averaging kernel matrices and quantities derived from them (degree of freedom for signal (DOFS), resolution, measurement response (see Section 2.3.6).

The dimension `error_parameter` serves to distinguish between the different error states considered during the L2 {H_2O, δD} pair post-processing (see Section 2.3.7). `resolution_parameter` points to the different metrics that describe the properties of the averaging kernels (see Section 2.3.6.2). `regularisation_parameter` is used for providing parameters necessary for the retrieval regularisation (see Section 2.3.9). `fit_quality_parameter` points to different metrics evaluating the quality of the spectral retrieval fit quality (see Section 2.3.7). `wv_avk_rank` and `wv_xavkat_rank` provide information about the matrix compression considered for the averaging kernel matrices (see Section 2.3.6.1).

Table 1 *Dimensions included in the output files of the L2 {H_2O, δD} pair post-processing.*

Name	Specification
`observation_id`	Dimension indicating ID of each IASI observation (starting from 0 in each file)
`atmospheric_levels`	Dimension indicating the number of atmospheric altitude levels
`musica_species_id`	Dimension indicating the two water vapour proxy states (see Sect. 2.2): = 0: first water vapour proxy state (wv_1) = 1: second water vapour proxy state (wv_2)
`error_parameter`	Dimension indicating the different error variables (see Sect. 2.3.7)

`resolution_parameter`	Dimension indicating the different vertical resolution variables (see Sect. 2.3.6.2))
`regularisation_parameter`	Dimension indicating the parameter used during the retrieval regularisation (see Sect. 2.3.9)
`fit_quality_parameter`	Dimension indicating the different retrieval fit diagnostics (see Sect. 2.3.7)
`wv_avk_rank`	Dimension indicating the maximum rank of the decomposed averaging kernel matrix for the water vapour proxy state (see Sect. 2.3.6.1)
`wv_xavkat_rank`	Dimension indicating the maximum rank of the decomposed cross averaging kernel matrix for the water vapour proxy state with respect to atmospheric temperature (see Sect. 2.3.6.1)

2.3.2 Sensor information

Table 2 lists the variables that contain information about the IASI sensor and the viewing geometry.

The variables `across_track`, `along_track` and `platform_zenith_angle` give information about the viewing and pixel geometry of the IASI sensor and are provided by EUMETSAT. The latter also holds for `orbit_id`, which indicates the number of the corresponding Metop orbit. Thereby the orbits from each Metop satellite are counted individually. The variable `instrument` allows for determining the corresponding Metop/IASI instrument of each observation.

Table 2 *Variables providing information about the IASI sensor.*

Name	Dimension	Specification
`across_track`	`observation_id`	across track observation index (consistent to EUMETSAT input files)
`along_track`	`observation_id`	along track observation index (consistent to EUMETSAT input files)
`instrument`	`observation_id`	Integer variable indicating the corresponding IASI instrument: 0: IASI-A (onboard of Metop-A) 1: IASI-B (onboard of Metop-B) 2: IASI-C (onboard of Metop-C)
`orbit_id`	`observation_id`	Integer variable indicating the corresponding orbit number (every Metop satellite has a separate orbit numbering)
`platform_zenith_angle`	`observation_id`	Satellite instrument viewing angle (zenith), given in degree (°)

2.3.3 Spatio-temporal information

Table 3 lists all variables containing information about the spatio-temporal properties of each observation. The time information is given in UTC (`time` and `time_string`), which is converted into local solar time (`time_local_solar`) during the MUSICA IASI retrieval processing according to following equation:

$$t_{LST} = t_{UTC} + lon * \frac{24h}{360°} + \frac{t_{corr}}{60} \tag{3}$$

This transforms the UTC time information t_{UTC} according to its longitudinal position lon into the local solar time information t_{LST}, with both given in hours. The additional term t_{corr} corrects for the eccentricity of the Earth's orbit and the Earth's axial tilt according to following empirical equation (Milne, 1921):

$$t_{corr} = 9.87\ \sin 2B - 7.53 \cos B - 1.5\ \sin B \tag{4}$$

$$\text{with } B = \frac{360}{365}(d - 81)$$

where d is the number of days since the start of the year. As t_{corr} is given in minutes, it has to be divided by 60 when considered for correcting t_{LST}. At last, in the final output files, the local solar time is transformed into seconds relative to solar noon, i.e. negative (positive) values refer to morning (evening) IASI observations.

The horizontal spatial information of each measured data point is given via `lat` and `lon` and the vertical dimension is `atmospheric_levels`. The atmosphere is discretised between the surface and the top of the atmosphere, which is set to 55.6 km, and the nominal altitudes of the vertical grid considered during the MUSICA IASI retrieval are shown in Table 4. As the altitudes are provided as "above sea level (a.s.l.)", negative altitudes will arise for geological depressions. In case of the surface height being higher than the lowest retrieval grid level, those grid level below the surface height will be set to NaN, and the grid level below the lowest still included grid level is filled with the surface altitude. For instance, if the surface height is at 1000 m, than the grid level 26 - 28 will be NaN and grid level 25 will contain the surface height (1000 m). This leads to a differing amount of vertical grid levels for each observation depending on the surface height, which is indicated by the variable `musica_nol`.

The PROFIT-nadir retrieval has heritage of the PROFFIT retrieval used for the retrieval with ground-based solar absorption spectra measured with the NDACC (Network for the Detection of Atmospheric Composition Change). Therein it is required to provide the profile retrieval products on an altitude grid. The retrieval procedure requires a constant gas amount between two grid levels for each iteration. If temperature is fitted the pressure-altitude relation changes (at all levels above the altitude where the temperature is modified). This means that the altitude levels change weakly between each iteration step. The pressure levels do not change between each iteration step. However, the pressure levels change between different observations, because for each location and time there is a specific pressure-altitude relation (according to the hydrostatic equilibrium), which leads to slightly varying altitude levels for each observation (e.g. for the upper troposphere, the typical standard deviation of the altitude levels is around 10 m). The resulting altitude levels for each observation are then stored in the variable `musica_altitude_levels`.

Table 3 *Variables displaying the spatio-temporal information of each observation.*

Name	Dimension	Specification
`time`	`observation_id`	UTC time given in seconds since 2000-01-01 00:00:00
`time_string`	`observation_id`	UTC time given in ISO 8601 UTC date-time string
`time_local_solar`	`observation_id`	local solar time (considering eccentricity of the orbit of the Earth) given in seconds relative to solar noon
`lat`	`observation_id`	geographical latitude, relative to equator; positive (negative) values indicate North (South)
`lon`	`observation_id`	geographical longitude, relative to prime meridian; positive (negative) values indicate East (West)
`musica_altitude_levels`	`observation_id, atmospheric_levels`	Atmospheric altitude above sea levels (a.s.l.), given in meters (m)
`musica_nol`	`observation_id`	Number of atmospheric grid levels above the surface for each observation

Table 4 Nominal altitudes (above sea level) of vertical MUSICA IASI retrieval grid.

Level	0	1	2	3	4	5	6	7	8	9
Altitude (km)	55.60	48.55	42.37	36.34	30.69	26.22	22.10	18.31	15.96	13.66
Level	10	11	12	13	14	15	16	17	18	19
Altitude (km)	12.00	10.92	9.78	8.01	7.18	8.88	6.38	5.62	4.90	4.22
Level	20	21	22	23	24	25	26	27	28	
Altitude (km)	3.57	2.95	2.37	1.82	1.30	0.83	0.39	0	< 0	

2.3.4 Water vapour state variables

This section shows the variables describing the retrieved and post-processed MUSICA IASI water vapour profiles.

The variables `musica_h2o` and `musica_deltad` represent the vertical profiles of the water vapour states H_2O and δD and thereby constitute the final {H_2O, δD} pairs as main output product from the L2 {H_2O, δD} pair post-processing (see Table 5). Additionally, their a priori assumed profiles considered during the retrieval processing and the L2 {H_2O, δD} pair post-processing are part of the final output (marked with the suffix `_apriori`).

For completeness, the vertical water vapour profiles are provided also in the water vapour proxy state base wv_1 and wv_2 (indicated by the syllable `wvp`, see Table 6), as these are the state vectors used during the L2 {H_2O, δD} pair post-processing (see Section 2.2). Both proxy states are stored in the variable `musica_wvp` and can be identified by means of the dimension `musica_species_id`. By using the conversion terms from Eqn. (2) and (3), the water vapour states (`musica_h2o` and `musica_deltad`) can be converted to the water vapour proxy states (`musica_wvp`) and vice versa.

Table 5 *Variables indicating the retrieved water vapour state profiles.*

Name	Dimension	Specification
`musica_h2o`	`observation_id, atmospheric_levels`	Retrieved H_2O state, given in volume mixing ratios (ppmv)
`musica_h2o_apriori`	`observation_id, atmospheric_levels`	A priori profiles for H_2O, given in volume mixing ratios (ppmv)
`musica_deltad`	`observation_id, atmospheric_levels`	Retrieved δD state, given in permille (‰)
`musica_deltad_apriori`	`observation_id, atmospheric_levels`	A priori profiles for δD, given in permille (‰)

Table 6 Variables indicating the retrieved water vapour proxy state profiles.

Name	Dimension	Specification
`musica_species_id`	–	Dimension variable indicating the two water vapour proxy states (see Sect. 2.2): = 0: first water vapour proxy state (wv_1) = 1: second water vapour proxy state (wv_2)
`musica_wvp`	`observation_id, musica_altitude_levels, musica_species_id`	Retrieved water vapour proxy states, given in natural logarithms of mixing ratios (ln(ppmv))
`musica_wvp_apriori`	`observation_id, musica_altitude_levels, musica_species_id`	A priori profiles for water vapour proxy states, given in natural logarithms of mixing ratios (ln(ppmv))

2.3.5 Atmospheric state variables

The output of the L2 {H_2O, δD} pair post-processing also includes various information about the meteorological state of the atmosphere (see Table 7).

The retrieved results of the atmospheric temperature are provided by `musica_at` and the retrieved surface temperature by `musica_st`. The variable `altitude_tropopause_climatological` is a measure for the climatological altitude of the temperature lapse rate tropopause (according to the definitions of the World Meteorological Organisation). It is derived from monthly output between 1979 – 2014 provided by the coupled chemistry climate model CESM1/WACCM (*Community Earth System Model version 1 / Whole Atmosphere Community Climate Model*, Marsh et al., 2013) and is constructed with a latitudinal dependence (Schneider et al., 2021a). `musica_pressure_levels` consists of the atmospheric pressure profiles provided by

EUMETSAT and interpolated to the MUSICA IASI retrieval grid altitudes (`musica_altitude_levels` from Table 3) after considering hydrostatic equilibrium (Schneider et al., 2021a).

Table 7 *Variables providing information about the atmospheric conditions.*

Name	Dimension	Specification
`musica_at`	`observation_id,` `musica_altitude_levels`	Retrieved atmospheric temperature state, given in Kelvin (K)
`musica_st`	`observation_id`	Retrieved surface temperature state, given in Kelvin (K)
`altitude_tropopause_climatological`	`observation_id`	Climatological tropopause altitude, given in meters (m)
`musica_pressure_levels`	`observation_id,` `musica_altitude_levels`	Atmospheric pressure levels, given in Pascal (Pa)

2.3.6 Averaging kernels

This section provides information on how the averaging kernels of the water vapour states are treated in the output of the L2 {H₂O, δD} pair post-processing.

An averaging kernel matrix A characterizes how the retrieved solution $\hat{\boldsymbol{x}}$ relates to the true atmospheric state $\boldsymbol{x}$ with respect to an a priori assumed profile $\boldsymbol{x_a}$:

$$\hat{\boldsymbol{x}} = \boldsymbol{A}(\boldsymbol{x} - \boldsymbol{x_a}) + \boldsymbol{x_a} \quad (5)$$

As described in Section 2.2, the MUSICA IASI retrieval works for the water vapour proxy states wv_1 and wv_2, analogously the corresponding averaging kernel matrix $\boldsymbol{A}_{wv}$ also refers to the water vapour proxy state base. Thereby $\boldsymbol{A}_{wv}$ is defined as follows:

$$\boldsymbol{A}_{wv} = \begin{pmatrix} \boldsymbol{A}_{11} & \boldsymbol{A}_{12} \\ \boldsymbol{A}_{21} & \boldsymbol{A}_{22} \end{pmatrix} \quad (6)$$

A_{11} is the averaging kernel matrix for wv_1, and A_{22} for wv_2, whereas $\boldsymbol{A}_{12}$ and $\boldsymbol{A}_{21}$ are the respective cross-responses between wv_1 and wv_2. The rows of the different kernel matrices depict the individual averaging kernels for the different grid level. Figure S2a displays the averaging kernels for the retrieval grid level from the surface to 10.9 km from all four entries of $\boldsymbol{A}_{wv}$. As the focus of the post-processed L2 {H₂O, δD} pair product is on the free troposphere, the averaging kernels of higher grid levels will show decreasing quality (not shown). Additionally, Figure S2b shows the averaging kernels from the cross-response of the atmospheric temperature retrieval to the water vapour proxy states:

$$\boldsymbol{A}_{xwv} = \begin{pmatrix} \boldsymbol{A}_{x1} \\ \boldsymbol{A}_{x2} \end{pmatrix} \quad (7)$$

Here, $\boldsymbol{A}_{x1}$ refers to the cross-response from the atmospheric temperature state to wv_1 and $\boldsymbol{A}_{x2}$ to wv_2.

2.3.6.1 Reconstruction of averaging kernel matrices

If the averaging kernels are given in the matrix form of Eqn. (6), then $\boldsymbol{A}_{wv}$ has the dimension (2*`musica_nol`, 2*`musica_nol`), i.e. at maximum (58,58), if `musica_nol` equals the considered 29 vertical grid levels (see Table 4).As including the averaging kernels for each observation will increase the output file size significantly, the matrices $\boldsymbol{A}_{wv}$ and $\boldsymbol{A}_{xwv}$ are decomposed by means of a singular value decomposition:

$$\boldsymbol{A} = \boldsymbol{UDV^T} \quad (8)$$

The components $\boldsymbol{U}$ and $\boldsymbol{V}$ are the leading singular vectors. $\boldsymbol{D}$ is a diagonal matrix, where the diagonal entries are called singular values. The number of singular values is called rank. In case of $\boldsymbol{A}_{wv}$, both components $\boldsymbol{U}$ and $\boldsymbol{V}$ have the shape (2*`musica_nol`, `musica_nol`), whereas $\boldsymbol{D}$ has the shape (`musica_nol`, `musica_nol`). In case of $\boldsymbol{A}_{xwv}$, the difference is that $\boldsymbol{V}$ has the shape (`musica_nol`, `musica_nol`).

The actual storage need reduction is achieved through an additional compression by considering only the leading singular values in $\boldsymbol{D}$, such that also the number of singular vectors in $\boldsymbol{U}$ and $\boldsymbol{V}$ reduces. In other words, the compression is achieved by reducing the rank of the matrix components. In the context of the L2 {H2O, δD} pair post-processing, singular values lower than 0.1 % of the maximum singular value in $\boldsymbol{D}$ are neglected, which is individually performed for every observation. The resulting rank is stored for each observation in the variable `musica_wvp_avk_rank` for $\boldsymbol{A}_{wv}$ and `musica_wvp_xavkat_rank` for $\boldsymbol{A}_{xwv}$ (see Table 8). In that way, for $\boldsymbol{A}_{wv}$ the shape of $\boldsymbol{D}$ becomes (`musica_wvp_avk_rank`, `musica_wvp_avk_rank`) and $\boldsymbol{U}$ and $\boldsymbol{V}$ reduce to (2*`musica_nol`, `musica_wvp_avk_rank`), whereas the components of $\boldsymbol{A}_{xwv}$ turn into (2*`musica_nol`, `musica_wvp_xavkat_rank`) in case of $\boldsymbol{U}$, (`musica_nol`, `musica_wvp_xavkat_rank`) for $\boldsymbol{V}$ and (`musica_wvp_xavkat_rank`, `musica_wvp_xavkat_rank`) for $\boldsymbol{D}$. The maximum rank from all observations within a single file defines the length of the corresponding dimensions `wv_avk_rank` and `wv_xavkat_rank` (see Table 4).

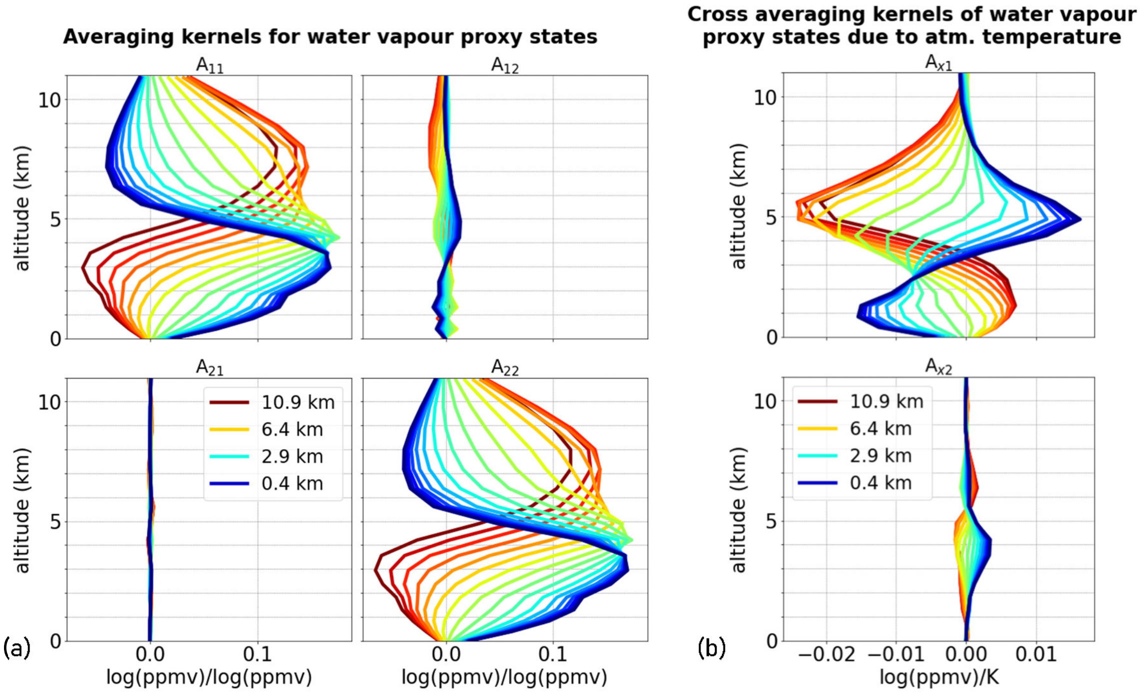

Figure S2 *Examples of averaging kernels (a) of the water vapour proxy states* wv_1 *and* wv_2 *(*$\boldsymbol{A}_{11}$ *and* $\boldsymbol{A}_{22}$*) and their cross-responses (*$\boldsymbol{A}_{12}$ *and* $\boldsymbol{A}_{21}$*) and (b) of the water vapour proxy states with respect to atmospheric temperature (*$\boldsymbol{A}_{x1}$ *and* $\boldsymbol{A}_{x2}$*)..*

Table 8 lists the matrix components of the averaging kernel matrix for the water vapour proxy state (variables starting with `musica_wvp_avk_<suffix>`) and of the averaging kernel matrix for the cross-response of the atmospheric temperature state to the water vapour proxy state (variables starting with `musica_wvp_xavkat_<suffix>`). The variables with the suffix `_lvec` refer to $\boldsymbol{U}$, those with `_rvec` to $\boldsymbol{V}$ and those with `_val` to $\boldsymbol{D}$. These variables are stored by considering the convention of using `musica_species_id`, if a variable refers to both water vapour proxy states. Thereby we reshape the dimension (2*`musica_nol`) to (`musica_species_id`, `musica_nol`), e.g. in case of $\boldsymbol{A}_{wv}$ the shape of $\boldsymbol{U}$ turns into (`musica_species_id`, `musica_nol`, `musica_wvp_avk_rank`). Please take note that the dimensions of these variables are permuted in the output netCDF files, e.g. the dimension of `musica_wvp_avk_lvec` is (`observation_id`, `musica_species_id`, `wv_avk_rank`, `atmospheric_levels`). To use this variable as $\boldsymbol{U}$ for a specific observation (e.g. `observation_id` = 0), its dimension `wv_avk_rank` has to be adjusted according to the corresponding rank given in `musica_wvp_avk_rank` and the dimension `atmospheric_levels` has to be adjusted according to the corresponding value given in `musica_nol`. A further dimension permutation achieves the desired shape of (2*`musica_nol`, `musica_wvp_avk_rank`). Listing 2 and Listing 3 provide code examples for this dimension permutation and for how these matrix components may be used to reconstruct the averaging kernel matrices $\mathbf{A}_{\mathrm{wv}}$ and $\mathbf{A}_{\mathrm{xwv}}$ for a single MUSICA IASI observation.

Table 8 *Variables describing the averaging kernel matrices of the water vapour proxy state.*

Name	Dimension	Specification

`musica_wvp_avk_lvec`	`observation_id,`
	`musica_species_id,`
	`wv_avk_rank,`
	`atmospheric_levels`
		Left leading singular vectors ($\boldsymbol{U}$) of averaging kernel matrix for water vapour proxy state

`musica_wvp_avk_rvec`	`observation_id,`
	`musica_species_id,`
	`wv_avk_rank,`
	`atmospheric_levels`
		Right leading singular vectors ($\boldsymbol{V}$) of averaging kernel matrix for water vapour proxy state

`musica_wvp_avk_rank`	`observation_id`	Number of leading singular values (rank) of averaging kernel matrix for water vapour proxy state
`musica_wvp_avk_val`	`observation_id,` `wv_avk_rank`	Leading singular values (diagonal values of $\boldsymbol{D}$) of averaging kernel matrix for water vapour proxy state

`musica_wvp_xavkat_lvec`	`observation_id,`
	`musica_species_id,`
	`wv_avk_rank,`
	`atmospheric_levels`
		Left leading singular vectors ($\boldsymbol{U}$) of cross averaging kernel matrix for water vapour proxy state w.r.t. atmospheric temperature

`musica_wvp_xavkat_rvec`	`observation_id,`
	`wv_avk_rank,`
	`atmospheric_levels`
		Right leading singular vectors ($\boldsymbol{V}$) of cross averaging kernel matrix for water vapour proxy state w.r.t. atmospheric temperature

`musica_wvp_xavkat_rank`	`observation_id`	Number of leading singular values (rank) of cross averaging kernel matrix for water vapour proxy state w.r.t. atmospheric temperature
`musica_wvp_xavkat_val`	`observation_id,` `wv_avk_rank`	Leading singular values (diagonal values of $\boldsymbol{D}$) of cross averaging kernel matrix for water vapour proxy state w.r.t. atmospheric temperature

Listing 2 *Code example in Python3.6 for reconstructing the averaging kernel matrix* $\mathbf{A}_{wv}$.

```
# select observation
obs = 0

# read metadata
nol  = ds['musica_nol'].values[obs]
rank = ds['musica_wvp_avk_rank'].values[obs]

# read averaging kernel arrays according to nol and rank
u   = ds['musica_wvp_avk_lvec'][obs,:,:rank,:nol]
sig = ds['musica_wvp_avk_val'][obs,:rank]
v   = ds['musica_wvp_avk_rvec'][obs,:,:rank,:nol]

# permute dimensions in u and v from shape (2, rank, nol) to (2, nol, rank)
u = u.transpose('musica_species_id', 'atmospheric_levels', 'wv_avk_rank')
v = v.transpose('musica_species_id', 'atmospheric_levels', 'wv_avk_rank')

# reshape dimensions in u and v from shape (2, nol, rank) to (2*nol, rank)
u = u.data.reshape(2*nol,rank)
v = v.data.reshape(2*nol,rank)

# create diagonal matrix from sig values
sig = np.diag(sig)

# reconstruct averaging kernel matrix for water vapour proxy states
A_wv = u @ sig @ v.T

# A_wv has shape (2*nol, 2*nol), such that:
A_11 = A_wv[:nol,:nol]
A_12 = A_wv[:nol,nol:]
A_21 = A_wv[nol:,:nol]
A 22 = A wv[nol:,nol:]
```

Listing 3 *Code example in Python3.6 for reconstructing the averaging kernel matrix* $\mathbf{A}_{xwv}$.

```
# read metadata
nol = ds['musica_nol'].values[obs]
rank = ds['musica_wvp_xavkat_rank'].values[obs]

# read averaging kernel arrays according to nol and rank
u   = ds['musica_wvp_xavkat_lvec'][obs,:,:rank,:nol]
sig = ds['musica_wvp_xavkat_val'][obs,:rank]
v   = ds['musica_wvp_xavkat_rvec'][obs,:rank,:nol]

# permute dimensions in u from shape (2, rank, nol) to (2*nol, rank)
# and in v from shape (rank, nol) to (nol, rank)
u = u.transpose('musica_species_id', 'atmospheric_levels', 'wv_avk_rank')
v = v.transpose('atmospheric_levels', 'wv_avk_rank')

u = u.data.reshape(2*nol,rank)
v = v.data.reshape(nol,rank)

# create diagonal matrix from sig values
sig = np.diag(sig)

# reconstruct averaging kernel matrix for water vapour proxy state
A_xwv = u @ sig @ v.T

# A_xwv has shape (2*nol, nol), such that:
A_x1 = A_wv[:nol,:nol]
A_x2 = A_wv[nol:,:nol]
```

2.3.6.2 Averaging kernel diagnostics

In addition to the averaging kernel matrices, various metrics describing the vertical characteristics of the averaging kernels for the water vapour proxy state are provided (see Table 9).

`musica_wvp_dofs` consists of the degree of freedom for signal (DOFS), calculated as trace of the averaging kernel matrices $\boldsymbol{A}_{11}$ and $\boldsymbol{A}_{22}$ for both water vapour proxy states wv_1 and wv_2 (proxy states identified by the dimension `musica_species_id`). `musica_wvp_resolution` provides metrics for the vertical resolution of the averaging kernels in $\boldsymbol{A}_{11}$ and $\boldsymbol{A}_{22}$. It includes information about the position of the kernel centre relative to the nominal altitude (`resolution_parameter = 0`, see Table 9) as well as the resolving length of an averaging kernel (`resolution_parameter = 1`, Eq. 3.23 and 3.24 of Rodgers, 2000) and the layer width per DOFS (`resolution_parameter = 2`, Purser and Huang, 1993). These metrics are evaluated at each vertical grid level; therefore, `musica_wvp_resolution` has the dimension `atmospheric_levels`. The variable `musica_wvp_response` displays the measurement response of an observation, which is calculated as sum along an averaging kernel, i.e. the sum along the rows of the averaging kernel matrix. This again is performed for $\boldsymbol{A}_{11}$ and $\boldsymbol{A}_{22}$ at each vertical grid level.

The metrics describing the kernel centre, resolution and measurement response are used for defining quality flags and for performing the quality filtering of the MUSICA IASI {H₂O, δD} pair data (see Sections 2.3.8 and 2.4.1). As the MUSICA IASI L2 {H₂O, δD} pair post-processing achieves a harmonization of the averaging kernels for wv_1 and wv_2 (see Figure S1), also the kernel metrics from Table 9 lead to overall similar values for the two different water vapour proxy states. Therefore, the metric values from both entries `musica_species_id = 0` (referring to wv_1) and `musica_species_id = 1` (referring to wv_2) can be used as equivalent representative metrics for the {H₂O, δD} pairs. For instance, if a user is interested in the DOFS of a single {H₂O, δD} pair, both following options may be used with both leading to quasi identical results:

Listing 2 *Code example for reading the DOFS of a single {H₂O, δD} pair.*

```
# read observation
obs = 0

# read DOFS for {H2O,dD} pair with musica_species_id = 0
dofs_h2o_dD_pairs = ds['musica_wvp_dofs'].values[obs,0]

# read DOFS for {H2O,dD} pair with musica_species_id = 1
dofs_h2o_dD_pairs = ds['musica_wvp_dofs'].values[obs,1]
```

Table 9 *Variables describing the properties of the averaging kernel matrices for the water vapour proxy state.*

Name	Dimension	Specification
`musica_wvp_dofs`	`observation_id,` `musica_species_id`	Degree of freedom for signal (trace of averaging kernel matrix) for water vapour proxy state

`musica_wvp_resolution`	`observation_id,`
	`resolution_parameter,`
	`musica_species_id,`
	`atmospheric_levels`
		Metrics for vertical representativeness for water vapour proxy state (see dimensions in Table 1). The different resolution metrics are indicated by parameters of the dimension `resolution_parameter`:
		= 0: centre of averaging kernel (Eq. 3.24 of Rodgers, 2000)
		= 1: resolving length of averaging kernel (Eq. 3.23 and 3.24 of Rodgers, 2000)
		= 2: layer width per degree of freedom (Purser and Huang, 1993)

`musica_wvp_response`	`observation_id,`
	`musica_species_id,`
	`atmospheric_levels`
		Measurement response (sum of averaging kernel) for water vapour proxy state

2.3.7 Errors and uncertainties

Table 10 gives an overview of the error variables related to the water vapour states. The variables `musica_h2o_error` and `musica_deltad_error` refer to the errors of H_2O and δD and are calculated as sum of the measurement noise error and the temperature error. As reference, the individual error components are also provided in the water vapour proxy state base (`musica_wvp_error`), including the measurement noise error for `error_parameter = 0` and the temperature error for `error_parameter = 1`. Listing 3 shares an example of how `musica_wvp_error` can be transformed into errors of H_2O (`musica_h2o_error`) and δD (`musica_deltad_error`).

Additionally, a variable containing information about the quality of the spectral retrieval fit quality is included in Table 10 (`musica_fit_quality` according to the dimension `fit_quality_parameter`). This variable is used to set the flag variable `musica_fit_quality_flag`. Further details about this variable are given in Schneider et al. (2021a).

Table 10 *Variables including the errors of the different water vapour states.*

Name	Dimension	Specification
`musica_h2o_error`	`observation_id,` `atmospheric_levels`	Total H_2O error (sum of measurement noise error and temperature error), given in volume mixing ratios (ppmv)
`musica_deltad_error`	`observation_id,` `atmospheric_levels`	Total δD error, given in permille (‰)

`musica_wvp_error`	`observation_id,`
	`error_parameter,`
	`musica_species_id,`
	`atmospheric_levels`
		Relative errors of water vapour proxy states, given in natural logarithms of volume mixing ratios (ln(ppmv)). The different errors are indicated by the parameters of the dimension `error_parameter`:
		= 0: Uncertainty due to retrieval fit noise
		= 1: Uncertainty due to atmospheric temperature a priori constraint

`musica_fit_quality`	`observation_id,` `fit_quality_parameter`	RMS values of the residual of the spectral fit. Different spectral fit metrics are indicated by the parameters of the dimension `fit_quality_parameter`: = 0: RMS for the full residual =1: RMS for the systematic residual =2: RMS for the random residual

Listing 3 *Code example for calculating H2O and dD errors from water vapour proxy state errors.*

```
# select observation and altitude
obs = 0
alt = 10

# read variables
h2o    = ds['musica_h2o'].values[obs,alt]
wv2    = ds['musica_wvp'].values[obs,1,alt]
err_wv = ds['musica_wvp_error'].values[obs]

# noise errors
err_wv_n_wv1 = err_wv[0,0,alt]
err_wv_n_wv2 = err_wv[0,1,alt]

# temperature errors
err_wv_t_wv1 = err_wv[1,0,alt]
err_wv_t_wv2 = err_wv[1,1,alt]

# calculate errors for h2o and dD
err_h2o = (err_wv_n_wv1 + err_wv_t_wv1) * h2o
err_dD  = (err_wv_n_wv2 + err_wv_t_wv2) * np.exp(wv2) * 1000
```

2.3.8 Flag variables

For providing a user-friendly and intuitive overview of the quality of the MUSICA IASI {H_2O, δD} pair data, different flag variables are derived and provided.

The flags `musica_deltad_error_flag` and `musica_wvp_kernel_flag` are derived during the L2 {H_2O, δD} pair post-processing. The former is defined according to the total error of δD (`musica_deltad_error`), whereas the latter is based on the averaging kernel matrices for the water vapour proxy states (see Section 2.3.6). Thereby, both flag variables are evaluated individually at each vertical grid level, consequently they have the dimension `atmospheric_levels`. Users interested in the definitions of the individual flags are referred to Diekmann et al. (2021a).

The flag `musica_fit_quality_flag` is defined according to the variable `musica_fit_quality` and provides information about the quality of the spectral retrieval fit quality. Details about the flags `musica_fit_quality_flag` and `surface_emissivity_flag` are given in Schneider et al. (2021a).

Table 11 *Flag variables for indicating the quality of the MUSICA IASI {H_2O, δD} pair data.*

Name	Dimension	Specification
`musica_deltad_error_flag`	`observation_id,` `atmospheric_levels`	Flag variable for the total δD error: = 0: error above 40 ‰ = 1: error below 40 ‰
`musica_wvp_kernel_flag`	`observation_id,` `atmospheric_levels`	Flag variable for vertical representativeness for the water vapour proxy state: = 0: limited vertical representativeness = 1: high vertical representativeness
`musica_fit_quality_flag`	`observation_id`	Flag variable indicating the quality of the spectral fit during the retrieval: = 0: poor quality = 1: restricted quality = 2: fair quality = 3: good quality
`surface_emissivity_flag`	`observation_id`	Flag variable indicating the source of surface emissivity data: = -2: Cloud (no retrieval) = -1: Masuda et al. (1988) over sea water for wind speed 5m/s = 0: IREMIS, no MOD11 data = 1: IREMIS baseline fit method = 2: IREMIS averaged from the 2 adjacent months

		= 3: IREMIS annual average = 4: IREMIS average over the annual average for lat < -80° = 5: ASTER spectral library version 2.0 sea ice emissivity = 6: EUMETSAT IASI L2 emissivity

2.3.9 Further retrieval parameters

This section shares details about variables that provide extended information about the water vapour retrieval.

Table 12 includes the variables necessary for reconstructing the a priori assumed uncertainty covariance matrix $\boldsymbol{S}_a$ according to Eqn. (7) from Schneider et al. (2021a). Here, `musica_apriori_cl` represents the altitude dependent vertical correlation length (σ_{cl}), whereas `musica_wvp_apriori_amp` and `musica_at_apriori_amp` are the altitude dependent amplitudes of the assumed variability (v_{amp}) for the water vapour proxy states and the atmospheric temperature. Based on these variables, Listing 5 provides a code example of how Eqn. (7) from Schneider et al. (2021a) can be implemented to reconstruct the uncertainty covariance matrices $\boldsymbol{S}_a$ for the water vapour proxy state and the atmospheric temperature.

Additionally, Table 12 lists those parameters required for calculating the retrieval constraint matrices $\boldsymbol{R}$ (used during MUSICA IASI retrieval, see Eqn. (8) in Schneider et al., 2021a, and Eqn. (13) in Diekmann et al., 2021a) and $\boldsymbol{R}_d$ (used during L2 {H_2O, δD} pair post-processing, see Eqn. (14) in Diekmann e al., 2021a). These matrices serve to constrain the variability of the water vapour proxy states wv_1 and wv_2 with respect to the a priori states. They are calculated as approximated inverse of the uncertainty covariance matrix $\boldsymbol{S}_a$. Following Eqn. (8)-(11) from Schneider et al. (2021a), this is achieved by constructing the constraint matrices as a sum of the diagonal constraint and first and second order Tikhonov-type regularisation matrices. The parameters to rebuild these three matrix components are stored in `musica_wvp_reg` and are identified by the dimension `regularisation_parameter`. Listing 4 displays an example for using these parameter to calculate the constraint matrices $\boldsymbol{R}_{H2O}$ and $\boldsymbol{R}_{d,H2O}$ for the first water vapour proxy state wv_1 (`musica_species_id = 0`; by considering `musica_species = 1` the constraint matrices can be calculated analogously for wv_2, and the full constraint matrices can then be reconstructed following Eqn. (5) from Schneider et al., 2021a).

Table 12 *Variables providing extended information about the water vapour retrieval.*

Name	Dimension	Specification
`musica_apriori_cl`	`observation_id,` `atmospheric_levels`	A priori assumed vertical correlation length for atmospheric variability, given in meters (m)
`musica_at_apriori_amp`	`observation_id,` `atmospheric_levels`	A priori assumed amplitude of atmospheric temperature variability

`musica_wvp_apriori_amp`	`observation_id,`
	`musica_species_id,`
	`atmospheric_levels`
		A priori assumed amplitude of water vapour proxy state variability

`regularisation_parameter`	-	Dimension variable indicating the parameter used during the retrieval regularisation (see Sect. 2.3.9): = 0: L0 coefficients = 1: L1 coefficients = 2: L2 coefficients

`musica_wvp_reg`	`observation_id,`
	`regularisation_parameter,`
	`musica_species_id,`
	`atmospheric_levels`
		Regularisation parameter for water vapour proxy state (see dimensions in Table 1)

`musica_iterations`	`observation_id`	Number of iterations, after which convergence was achieved during retrieval process

Listing 5 *Code example for reconstructing the a priori assumed covariance matrices for the water vapour proxy state and the atmospheric temperature*

```
# function for Sa reconstruction (Eqn. 7 from Schneider et al., 2021a)
def construct_Sa(alt, amp, sigma):
    Sa = amp[:,np.newaxis] * amp[np.newaxis,:] \
        * np.exp( -(alt[:,np.newaxis] - alt[np.newaxis,:])**2 / \
                   (2 * sigma[:,np.newaxis] * sigma[np.newaxis,:]) )
    return Sa

# select observation
obs = 0

# read data
nol = ds['musica_nol'].values[obs]
alt = ds['musica_altitude_levels'].values[obs,:nol]
wv1_amp = ds['musica_wvp_apriori_amp'].values[obs,0,:nol]
wv2_amp = ds['musica_wvp_apriori_amp'].values[obs,1,:nol]
t_amp    = ds['musica_at_apriori_amp'].values[obs,:nol]
corrlength = ds['musica_apriori_cl'].values[obs,:nol]

# Sa for water vapour proxy states
Sa_wvp = np.zeros([2*nol, 2*nol])

Sa_wvp[:nol,:nol] = construct_Sa(alt, wv1_amp, corrlength)
Sa_wvp[nol:,nol:] = construct_Sa(alt, wv2_amp, corrlength)

# Sa for atmospheric temperature
Sa_t = construct_Sa(alt, t_amp, corrlength)
```

Listing 4 *Code example for reconstructing the retrieval constraint for the water vapour proxy state* wv_1.

```
# constraint reconstruction (here shown for musica_species_id = 0)
# see Eqn. (5) and Section 4.6 in Schneider et al. (2021a)
nol = ds['musica_nol'].values[obs]

L0coef = ds['musica_wvp_reg'].values[obs,0,0,:nol]
L1coef = ds['musica_wvp_reg'].values[obs,1,0,:nol]
L2coef = ds['musica_wvp_reg'].values[obs,2,0,:nol]

# constraint R0 for absolute variability
a_L0 = np.diag(L0coef)
R0   = a_L0.T @ a_L0

# R1 for variability of first vertical derivatives of the profiles
L1coef = L1coef[:nol-1]
a_L1   = np.diag(-L1coef, k=1)[:nol-1]
np.fill_diagonal(a_L1, L1coef)
R1 = a_L1.T @ a_L1

# R2 for variability of second vertical derivates of the profiles
L2coef = L2coef[:nol-2]
a_L2   = np.diag(L2coef, k=2)[:nol-2]
np.fill_diagonal( a_L2[:,1:], -2 * L2coef)
np.fill_diagonal( a_L2, L2coef )
R2 = a_L2.T @ a_L2

# build constraint according to Eqn. 13 from Diekmann et al. (2021a)
R_wv1  = R0 + R1 + R2

# build reduced constraint according to Eqn. 14 from Diekmann et al. (2021a)
Rd_wv1 = R1 + R2
```

2.3.10 EUMETSAT variables

Table 13 lists all flag variables provided by EUMETSAT as part of the IASI L1C and L2 files. This consists of flags for indicating cloud contaminations, EUMETSAT retrieval properties and surface types. A comprehensive description for all listed flags is given in the IASI L2: Product Guide (https://www.eumetsat.int/media/45982, accessed on 31.08.2021).

The variable `eumetsat_cloud_summary_flag` is used during the MUSICA IASI retrieval in order to select observations with no significant cloud contaminations (see Section 2.4.1). Furthermore, `eumetsat_surface_type_flag` provides information about the surface type of each observation. The terms “land low” and “land high” refer to regions where the variability of the surface topography is low or high, respectively. The same holds for the specifications “land water low” and “land water high”, which both point to inland waters. The remaining flag variables are included mainly for reference.

Table 13 *Auxiliary flag variables provided by EUMETSAT. In addition to the variable names used for the MUSICA IASI output, the names given in the parentheses indicate the corresponding variable names of the EUMETSAT IASI L2 files,*

Name	Dimension	Specification
`eumetsat_cloud_area_fraction` `(FRACTIONAL_CLOUD_COVER)`	`observation_id`	EUMETSAT L2 fractional cloud cover
`eumetsat_cloud_formation_flag` `(FLG_CLDFRM)`	`observation_id`	EUMETSAT L2 cloud formations origin
`eumetsat_cloud_summary_flag` `(FLAG_CLDNES)`	`observation_id`	EUMETSAT L2 cloudiness assessment summary: = 1: IFOV is clear = 2: IFOV is processed as cloud-free, but small cloud contaminations possible = 3: IFOV is partially covered by clouds = 4: High or full cloud coverage
`eumetsat_dust_flag` `(FLG_DUSTCLD)`	`observation_id`	EUMETSAT L2 dust cloud indicator
`eumetsat_first_guess_flag` `(FLG_INITIA)`	`observation_id`	EUMETSAT L2 PPF first guess usage
`eumetsat_parameter_flag` `(FLG_FGCHECK)`	`observation_id`	EUMETSAT L2 geophysical parameter check
`eumetsat_surface_type_flag` `(FLG_LANSEA)`	`observation_id`	EUMETSAT surface flag: = 0: water = 1: land low = 2: land high = 3: land water low = 4: land water high = 5: sea ice
`eumetsat_cloud_test_flag` `(FLG_CLDTST)`	`observation_id`	EUMETSAT L2 cloud test (Only for EUMETSAT L2 PPF versions 4 and 5. Only kept as placeholder, as currently version 6 is used.)
`eumetsat_residual_flag` `(FLG_RESID)`	`observation_id`	EUMETSAT L2 residual check (Only for EUMETSAT PPF versions 4 and 5. Only kept as placeholder, as currently version 6 is used.)

2.4 Recommendations for data usage

This section shares some recommendations when using the post-processed L2 {H_2O, δD} pair data. This includes a list of recommended filter conditions for obtaining {H_2O, δD} pairs with high data quality (Section 2.4.1). Section 2.4.2 discusses different ways for comparing the MUSICA IASI L2 L2 {H_2O, δD} pair data with simulations from isotope-enabled models.

2.4.1 Quality filtering of {H_2O, δD} pair data

The post-processed L2 {H_2O, δD} pair data are provided together with flag variables for a user-friendly quality filtering of the {H_2O, δD} pair data. The flags as well as their recommended values for achieving {H_2O, δD} pairs with high quality are given in Table 14.

The flags `musica_deltad_error_flag` and `musica_wvp_kernel_flag` from Table 11 allow for a height-depending data filtering with respect to the quality of the {H_2O, δD} pairs (see Diekmann et al., 2021a), whereas `musica_fit_quality_flag` indicates observations, where the spectral MUSICA IASI retrieval fit achieved high agreement between the measured and simulated spectrum (see Schneider et al., 2021a). The variable `eumetsat_cloud_summary_flag` is used for filtering out observations with cloud contaminations. Only data shall be considered, which show either no cloud contamination at all (`eumetsat_cloud_summary_flag = 1`) or only minor cloud contaminations (`eumetsat_cloud_summary_flag = 2`). The reason for also allowing minor clouds to appear in the observed scene is that high and thin ice clouds are expected to have only little impact on the results of the mid-tropospheric water vapour retrieval. However, this cloud filtering condition is already applied during the MUSICA IASI retrieval, therefore all post-processed L2 {H_2O, δD} pairs fulfil the cloud condition implicitly.

Table 14 *Flag variables used for quality fitering of the {H_2O, δD} pair data.*

Name	Recommended values	Specification
`musica_deltad_error_flag`	1	Considering only data with low error in δD (see Section 2.3.7 and Table 11); to be evaluated individually for each vertical grid level
`musica_wvp_kernel_flag`	1	Considering only data with reasonable averaging kernels of the water vapour proxy states (see Section 2.3.6 and Table 11); to be evaluated individually for each vertical grid level
`musica_fit_quality_flag`	2 and 3	Considering only data with good spectral retrieval fit quality (see Table 11); no height-dependence
`eumetsat_cloud_summary_flag`	1 and 2	Considering only data with no or minor cloud contaminations (see Table 13); no height-dependence; already satisfied for all L2 {H_2O, δD} pairs

2.4.2 Comparing with data from isotope-enabled models

The comprehensive data coverage and availability of the MUSICA IASI {H_2O, δD} pairs make this dataset highly valuable for cross-comparisons and -evaluations with simulations from isotope-enabled models (i.e. atmospheric models that incorporate the fractionation physics of water vapour and its isotopes). However, due to fundamental differences in the data characteristics of these two different data products, a comparison of modelled against remotely sensed distributions of H_2O and δD should be performed and interpreted very carefully.

As a first step, some kind of temporal collocation criteria should be applied to the model data with respect to the local overpass times of the Metop satellites. For instance, the local overpasses can be approximated with 09.30 a.m. (morning overpasses) and 09.30 p.m. (evening overpasses) local time, what can be used to identify the corresponding model data from each day.

Furthermore, it is strongly recommended to consider some kind of cloud filtering for the model data, which should be in line with the cloud filtering characteristics of the MUSICA IASI data. An example would be to select those model data, where the total column does not contain any liquid clouds, but might contain minor ice cloud abundances (see Section 2.4.1).

As discussed in Section 2.3.6, the vertical structure of the MUSICA IASI {H_2O, δD} pair data is characterized by means of the averaging kernel matrix. In contrast, data from isotope-enabled models are typically provided as discrete values on fixed vertical grid levels. Therefore, a direct satellite-to-model comparison is a non-trivial task and requires a careful treatment.

A rough way of comparing the MUSICA IASI data to model output, in particular for rather climatological analyses, is to compare the modelled H_2O and δD distributions directly to the quality-filtered MUSICA IASI {H_2O, δD} pairs. This might be justified with the fact that the quality filters from Section 2.4.1 aim to select those MUSICA IASI {H_2O, δD} pair data, which are largely representative for the targeted nominal altitudes (no significant kernel centre displacement, with reasonable vertical resolution and high measurement response). However, it should be kept in mind that with such a comparison without considering any averaging kernel effects still discrepancies in the vertical representativeness of the model and satellite data might remain.

More sophisticated comparisons of simulations from isotope-enabled models with the MUSICA IASI {H_2O, δD} pair data should adjust the modelled profiles of H_2O and δD according to the vertical characteristics of the remotely sensed {H_2O, δD} pair product. This should be achieved by considering Eqn. (5) and treating the modelled state $\boldsymbol{x}_{mod}$ as “truth”:

$$\widehat{\boldsymbol{x}}_{mod} = \boldsymbol{A}(\boldsymbol{x}_{mod} - \boldsymbol{x}_{\boldsymbol{a}}) + \boldsymbol{x}_{\boldsymbol{a}} \tag{9}$$

By considering the averaging kernel matrix $\boldsymbol{A}$ and the a priori assumed profiles $\boldsymbol{x}_{\boldsymbol{a}}$, this equation modifies the modelled state $\boldsymbol{x}_{mod}$ according to the vertical characteristics of the remote sensing product and creates the new adjusted state $\widehat{\boldsymbol{x}}_{mod}$. In other words, it mimics how the IASI sensor would observe the atmosphere of the isotope-enabled model and thereby simulates a theoretical IASI observation $\widehat{\boldsymbol{x}}_{mod}$ based on the modelled atmospheric state.

The performance of Eqn. (9) largely depends on the choice of the averaging kernel matrix $\boldsymbol{A}$. The simplest approach would be to consider $\boldsymbol{A}$ and $\boldsymbol{x}_{\boldsymbol{a}}$ from a collocated IASI observation and to apply it to a vertical profile $\boldsymbol{x}_{mod}$ from the model. But as tropospheric distributions of water vapour can be highly variable, it is not guaranteed that in that case $\boldsymbol{A}$ and $\boldsymbol{x}_{mod}$ refer to the exact same atmospheric state. Applying averaging kernels to a “wrong” profile can thereby generate misleading results. Therefore, the optimal approach would be to simulate the MUSICA IASI averaging kernels for the atmospheric state given by the model. In this way, it is ensured that $\boldsymbol{A}$ and $\boldsymbol{x}_{mod}$ refer to the same atmospheric state. A first example of how this can be achieved with respect to the MUSICA IASI {H_2O, δD} pair product is presented and discussed in Schneider et al. (2017), where a simple radiative transfer model is assumed and applied on the model data. A more sophisticated but computationally intensive approach that is based on the full radiative transfer assumptions of the MUSICA IASI retrieval can be found in Diekmann (2021).

When working with Eqn. (9), it is important to note that the kernel matrix $\boldsymbol{A}$ is provided for the water vapour proxy states wv_1 and wv_2 (see Section 2.3.6). This implies that the variable `musica_wvp_apriori` (see Table 6) should be used as a priori states $\boldsymbol{x}_{\boldsymbol{a}}$. Consequently, the modelled profiles of H_2O and δD have to be transformed into the water vapour proxy state base following Eqn. (2), in order to be used as $\boldsymbol{x}_{mod}$. The adjusted model profiles $\widehat{\boldsymbol{x}}_{mod}$ are then also given in the water vapour proxy state base and can be transformed back to profiles of H_2O and δD by using Eqn. (3).

3 Level-3 {H_2O, δD} Pair Data

This section provides a documentation of the Level-3 (L3) dataset of the post-processed MUSICA IASI {H_2O, δD} pair product. This dataset is derived using the output of the Level-2 (L2) {H_2O, δD} pair post-processing and is presented in Diekmann et al. (2021a). In the following, an overview will be given about the output filenames (Section 3.1), the main product (Section 3.2) and the structure of the output files (Section 3.3).

3.1 Output filenames

Analogous to the L2 files of the {H_2O, δD} pairs (Section 2.1), also the L3 files are sorted into global and daily files, with considering a further separation into morning (local overpass around 09.30 am) and evening (local overpass around 21.30 pm) files for each day. The file naming convention is as follows:

`'IASI<multipleS>_MUSICA_<V>_L3pp_H2Oiso_<VPP>_<YYYYMMDD>_<OP>_global.nc'`

- `<multipleS>`: information about the considered sensors
    - Files between 2007-07-10 and 2013-02-19: `'A'`
    - Files between 2013-02-20 and 2019-10-23: `'AB'`
    - Files from 2019-10-24: `'ABC'`
- `<V>`: MUSICA retrieval processor version
    - Files until 2019-06-30: `'030201'`
    - Files from 2019-07-01: `'030300'`
- `<VPP>`: MUSICA IASI {H_2O, δD} pair post-processing version
    - All files: `'v2'`
- `<YYYYMMDD>`: day of observation (universal time format)
- `<OP>`: overpass information
    - Files with data with local overpass times around 09.30 am: `'morning'`
    - Files with data with local overpass times around 21.30 am: `'evening'`

3.2 Product specification

The main product of the L3 dataset is the {H_2O, δD} pair product from Section 2.2, but with only considering data quality-filtered according to Section 2.4.1, interpolated to the fixed altitude levels 2.95, 4.22 and 6.38 km (see Table 4) and re-gridded individually for each altitude level on a regular 1° x 1° grid. Details about the re-gridding methods considered for data of the different output variables within the individual grid boxes are documented in the following section.

3.3 Dataset documentation

This section serves as technical documentation of the output files from the L3 datasets of the {H_2O, δD} pairs.

3.3.1 Dimensions

Table 15 lists all dimensions included in the output files of the L3 dataset of the post-processed MUSICA IASI {H_2O, δD} pair product.

Table 15 *Dimensions included in the L3 output files.*

Name	Specification
`lat`	Dimension indicating the number of geographical latitudes
`lon`	Dimension indicating the number of geographical longitudes
`altitude_levels`	Dimension indicating the number of atmospheric altitude levels
`surface_type`	Dimension indicating the specific surface types (see Sect. 3.3.6)

3.3.2 Spatio-temporal information

Table 16 lists the variables describing the spatio-temporal properties of the L3 data.

The horizontal coordinates are `lat` and `lon` and the vertical information is indicated by the dimension `altitude_levels`. The variables `time` and `time_local_solar` indicate the UTC time and the local time (according to Section 2.3.3) and are averaged for all data points within the individual grid boxes at each chosen altitude level. Consequently, these time variables consist of the three spatial coordinates. The variable `nobs` informs about the number of observations within each grid box, which are used for averaging.

Table 16 *Dimensions and variables displaying the spatio-temporal information of the L3 data.*

Name	Dimension	Specification
`lat`	`lat`	Dimension variable indicating the geographical latitude, relative to equator; positive (negative) values indicate North (South)
`lon`	`lon`	Dimension variable indicating the geographical longitude, relative to prime meridian; positive (negative) values indicate East (West)
`altitude_levels`	`altitude_levels`	Dimension variable indicating the number of atmospheric altitude levels
`time`	`altitude_levels, lat, lon`	UTC time given in seconds since 2000-01-01 00:00:00
`time_local_solar`	`altitude_levels, lat, lon`	local solar time (considering eccentricity of the orbit of the Earth) given in seconds relative to solar noon
`nobs`	`altitude_levels, lat, lon`	number of quality-filtered observations per grid box used for averaging

3.3.3 Water vapour state variables

The {H_2O, δD} pairs of the L3 dataset are given by the variables `musica_h2o` and `musica_deltad` (see Table 17).

The re-gridding on the regular 1° x 1° grid is achieved for H_2O by linear averaging all values within each individual grid box. For δD, this is done by first averaging H_2O and HDO (derived from δD) individually and then, second, by calculating the averaged δD based on the averaged H_2O and HDO following Eqn. (1).

Table 17 *Variables indicating the L3 water vapour state data.*

Name	Dimension	Specification
`musica_h2o`	`altitude_levels, lat, lon`	Retrieved H_2O state, given in volume mixing ratios (ppmv)
`musica_deltad`	`altitude_levels, lat, lon`	Retrieved δD state, given in permille (‰)

3.3.4 Atmospheric state variables

Table 18 lists the variables that describe the atmospheric state and are included in the L3 dataset. This comprises the atmospheric temperature (`musica_at`) and pressure (`musica_pressure_levels`). Analogous to H_2O, both variables are re-gridded on the regular 1° x 1° grid through linear averaging all values within the individual grid boxes at the different altitude levels.

Table 18 *Variables indicating the atmospheric state variables included in the L3 dataset.*

Name	Dimension	Specification
`musica_at`	`altitude_levels, lat, lon`	Retrieved atmospheric temperature, given in Kelvin (K)
`musica_pressure_levels`	`altitude_levels, lat, lon`	Atmospheric pressure levels, given in Pascal (Pa)

3.3.5 Errors and uncertainties

Table 19 lists the variables that characterize the uncertainty of the L3 {H_2O, δD} pair data.

As documented in Diekmann et al. (2021a), the calculation of the averaged values of the total errors of H_2O and δD is performed following the simple assumption is that the temperature and measurement noise errors (which together form the total error of the L2 {H₂O, δD} pair product) equally consist of 50% systematic error components and 50% random error components. Thus, the averaged errors of H₂O and δD are calculated as follows:

$$x_{rand} = \frac{\sqrt{x_{t,1}^2 + x_{t,2}^2 + x_{t,3}^2 + \cdots + x_{n,1}^2 + x_{n,2}^2 + x_{n,3}^2 + \cdots}}{nobs} \tag{10}$$

$$x_{syst} = \frac{\sqrt{(x_{t,1} + x_{t,2} + x_{t,3} + \cdots)^2 + (x_{n,1} + x_{n,2} + x_{n,3} + \cdots)^2}}{nobs} \tag{11}$$

$$x_{tot} = \frac{x_{rand} + x_{syst}}{2} \tag{12}$$

with $x_{t,i}$ as temperature error for the observation i, with $x_{n,i}$ as measurement noise error for the observation i, with x_{rand} as random error of all temperature and noise error values within a grid box, with x_{syst} as systematic error of all temperature and noise error values within a grid box and with x_{tot} as derived total error of all data within a grid box.

The representativeness of the averaged L3 data for H₂O and δD is documented by `musica_h2o_rms` and `musica_deltad_rms`. This metric is derived as RMS of the differences of the individual H₂O and δD data within a single grid box to their averaged value. For H₂O, the respective calculation is applied on the logarithmic scale.

Table 19 *Variables indicating the errors and uncertainty of the L3 water vapour product.*

Name	Dimension	Specification
`musica_h2o_error`	`altitude_levels, lat, lon`	Total H₂O error (sum of measurement noise error and temperature error), given in volume mixing ratios (ppmv)
`musica_h2o_rms`	`altitude_levels, lat, lon`	RMS of the differences of all ln(H₂O) values to the ln(H2O) average within each grid box
`musica_deltad_error`	`altitude_levels, lat, lon`	Total δD error (sum of measurement noise error and temperature error), given in permille (‰)
`musica_deltad_rms`	`altitude_levels, lat, lon`	RMS of the differences of all δD values to the δD average within each grid box

3.3.6 Surface information

The variable `surface_type_frac` serves as indicator for the surface type and is derived from the L2 variable `eumetsat_surface_type_flag` from Table 13. It displays the relative fraction of each considered surface type (indicated by the dimension `surface_type`) for the data points within the individual grid boxes, evaluated at the different chosen altitude levels. For this purpose, we consider the surface types given in Table 13 and discussed in Section 2.3.10, with the only change that, due to the reduced horizontal resolution of 1°, observations with the types “land water low” and “land water high” are now treated as “land low” and “land high”, respectively.

Table 20 *Variables providing information about the type of surface.*

Name	Dimension	Specification
`surface_type_frac`	`altitude_levels, lat, lon, surface_type`	Surface type, providing the fraction (given in percent, %) of the different surface types according to the parameters of the dimension variable `surface_type`: = 0: water = 1: land low = 2: land high = 3: sea ice

4 References

Barthlott, S., Schneider, M., Hase, F., Blumenstock, T., Kiel, M., Dubravica, D., García, O. E., Sepúlveda, E., Mengistu Tsidu, G., Takele Kenea, S., Grutter, M., Plaza-Medina, E. F., Stremme, W., Strong, K., Weaver, D., Palm, M., Warneke, T., Notholt, J., Mahieu, E., Servais, C., Jones, N., Griffith, D. W. T., Smale, D., and Robinson, J.: Tropospheric water vapour isotopologue data (H216O, H218O, and HD16O) as obtained from NDACC/FTIR solar absorption spectra, Earth Syst. Sci. Data, 9, 15–29, https://doi.org/10.5194/essd-9-15-2017, **2017**

Blumstein, D., Chalon, G., Carlier, T., Buil, C., Hebert, P., Maciaszek, T., Ponce, G., Phulpin, T., Tournier, B., Simeoni, D., Astruc, P., Clauss, A., Kayal, G., R. Jegou, "IASI instrument: technical overview and measured performances," Proc. SPIE 5543, Infrared Spaceborne Remote Sensing XII; https://doi.org/10.1117/12.560907, 2004

Borger, C., Schneider, M., Ertl, B., Hase, F., García, O. E., Sommer, M., Höpfner, M., Tjemkes, S. A., and Calbet, X.: Evaluation of MUSICA IASI tropospheric water vapour profiles using theoretical error assessments and comparisons to GRUAN Vaisala RS92 measurements, Atmos. Meas. Tech., 11, 4981–5006, https://doi.org/10.5194/amt-11-4981-2018, **2018**

Clerbaux, C., Boynard, A., Clarisse, L., George, M., Hadji-Lazaro, J., Herbin, H., Hurtmans, D., Pommier, M., Razavi, A., Turquety, S., Wespes, C., and Coheur, P.-F.: Monitoring of atmospheric composition using the thermal infrared IASI/MetOp sounder, Atmos. Chem. Phys., 9, 6041–6054, https://doi.org/10.5194/acp-9-6041-2009, **2009**

Diekmann C. J.: Analysis of stable water isotopes in tropospheric moisture during the West African Monsoon, PhD Thesis, Institute of Meteorology and Climate Research – Atmospheric Trace Gases and Remote Sensing (IMK-ASF), Karlsruhe Institute of Technology https://doi.org/10.5445/IR/1000134744, **2021**

Diekmann, C. J., Schneider, M., Ertl, B., Hase, F., García, O., Khosrawi, F., Sepúlveda, E., Knippertz, P., and Braesicke, P.: The MUSICA IASI {H2O, δD} pair product, Earth Syst. Sci. Data Discuss. [preprint], https://doi.org/10.5194/essd-2021-87, in review, **2021a**

Diekmann, C. J. Schneider, M. and Ertl, B.: MUSICA IASI water isotopologue pair product (a posteriori processing version 2). Institute of Meteorology and Climate Research, Atmospheric Trace Gases and Remote Sensing (IMK-ASF), Karlsruhe Institute of Technology (KIT). https://doi.org/10.35097/415, **2021b**

García, O. E., Schneider, M., Ertl, B., Sepúlveda, E., Borger, C., Diekmann, C., Wiegele, A., Hase, F., Barthlott, S., Blumenstock, T., Raffalski, U., Gómez-Peláez, A., Steinbacher, M., Ries, L., and de Frutos, A. M.: The MUSICA IASI CH4 and N2O products and their comparison to HIPPO, GAW and NDACC FTIR references, Atmos. Meas. Tech., 11, 4171–4215, https://doi.org/10.5194/amt-11-4171-2018, **2018**

Hase, F., J. W. Hannigan, M. T. Coffey, A. Goldman, M. Höpfner, N. B. Jones, C. P. Rinsland, and S. W. Wood: Intercomparison of retrieval codes used for the analysis of high-resolution, ground-based FTIR measurements. Journal of Quantitative Spectroscopy and Radiative Transfer, 87(1):25–52. ISSN 00224073. doi: 10.1016/j.jqsrt. 2003.12.008, **2004**

Marsh, D. R. M. J. Mills, D. E. Kinnison, J. F. Lamarque, N. Calvo, and L. M. Polvani. Climate change from 1850 to 2005 simulated in CESM1(WACCM). Journal of Climate, 26(19):7372–7391, ISSN 08948755. doi: 10.1175/JCLI-D-12-00558.1, **2013.**

Milne, R. M.: Note on the Equation, of Time. The Mathematical Gazette, vol. 10, no. 155, pp. 372 – 375, **1921**

Purser R. J. and H. L. Huang. Estimating effective data density in a satellite retrieval or an objective analysis. Journal of Applied Meteorology, 32(6):1092–1107. ISSN 08948763. doi: 10.1175/1520-0450(1993)032<1092:EEDDIA>2.0.CO;2, **1993**

Rodgers C. D.. Inverse methods for atmospheric sounding: Theory and Practice, Series on Atmospheric, Oceanic and Planetary Physics–Vol. 2,. Series on Atmospheric, Oceanic and Planetary Physics–Vol. 2, 2:256, **2000**

Schneider, M. and Hase, F.: Improving spectroscopic line parameters by means of atmospheric spectra: Theory and example for water vapor and solar absorption spectra. Journal of Quantitative Spectroscopy and Radiative Transfer, 110(17):1825–1839, ISSN 00224073. doi: 10.1016/j.jqsrt.2009.04.011, **2009**

Schneider, M. and Hase, F.: Optimal estimation of tropospheric H2O and δD with IASI/METOP, Atmos. Chem. Phys., 11, 11207–11220, https://doi.org/10.5194/acp-11-11207-2011, **2011**

Schneider, M., Barthlott, S., Hase, F., González, Y., Yoshimura, K., García, O. E., Sepúlveda, E., Gomez-Pelaez, A., Gisi, M., Kohlhepp, R., Dohe, S., Blumenstock, T., Wiegele, A., Christner, E., Strong, K., Weaver, D., Palm, M., Deutscher, N. M., Warneke, T., Notholt, J., Lejeune, B., Demoulin, P., Jones, N., Griffith, D. W. T., Smale, D., and Robinson, J.: Ground-based remote sensing of tropospheric water vapour isotopologues within the project MUSICA, Atmos. Meas. Tech., 5, 3007–3027, https://doi.org/10.5194/amt-5-3007-2012, **2012**

Schneider, M., Wiegele, A., Barthlott, S., González, Y., Christner, E., Dyroff, C., García, O. E., Hase, F., Blumenstock, T., Sepúlveda, E., Mengistu Tsidu, G., Takele Kenea, S., Rodríguez, S., and Andrey, J.: Accomplishments of the MUSICA project to provide accurate, long-term, global and high-resolution observations of tropospheric {H2O,δD} pairs – a review, Atmos. Meas. Tech., 9, 2845–2875, https://doi.org/10.5194/amt-9-2845-2016, **2016**

Schneider, M., Borger, C., Wiegele, A., Hase, F., García, O. E., Sepúlveda, E., and Werner, M.: MUSICA MetOp/IASI {H2O, δD} pair retrieval simulations for validating tropospheric moisture pathways in atmospheric models, Atmos. Meas. Tech., 10, 507–525, https://doi.org/10.5194/amt-10-507-2017, **2017**

Schneider, M., Ertl, B., Diekmann, C. J., Khosrawi, F., Weber, A., Hase, F., Höpfner, M., García, O. E., Sepúlveda, E., and Kinnison, D.: Design and description of the MUSICA IASI full retrieval product, Earth Syst. Sci. Data Discuss. [preprint], https://doi.org/10.5194/essd-2021-75, in review, **2021a**

Schneider, M., Ertl, B., Diekmann, C. J., Khosrawi, F., Röhling, A. N., Hase, F., Dubravica, D., García, O. E., Sepúlveda, E., Borsdorff, T., Landgraf, J., Lorente, A., Chen, H., Kivi, R., Laemmel, T., Ramonet, M., Crevoisier, C., Pernin, J., Steinbacher, M., Meinhardt, F., Deutscher, N. M., Griffith, D. W. T., Velazco, V. A., and Pollard, D. F.: Synergetic use of IASI and TROPOMI space borne sensors for generating a tropospheric methane profile product, Atmos. Meas. Tech. Discuss. [preprint], https://doi.org/10.5194/amt-2021-31, in review, **2021**

Schneider, M., Ertl, B. and Diekmann, C. J.: MUSICA IASI full retrieval product standard output (processing version 3.2.1). Institute of Meteorology and Climate Reasearch, Atmospheric Trace Gases and Remote Sensing (IMK-ASF), Karlsruhe Institute of Technology (KIT). https://doi.org/10.35097/408, **2021c**

Schneider, M., Ertl, B. and Diekmann, C. J.: MUSICA IASI full retrieval product extended output (processing version 3.2.1). Institute of Meteorology and Climate Reasearch, Atmospheric Trace Gases and Remote Sensing (IMK-ASF), Karlsruhe Institute of Technology (KIT). https://doi.org/10.35097/412, **2021d**

Wiegele, A., Schneider, M., Hase, F., Barthlott, S., García, O. E., Sepúlveda, E., González, Y., Blumenstock, T., Raffalski, U., Gisi, M., and Kohlhepp, R.: The MUSICA MetOp/IASI H2O and δD products: characterisation and long-term comparison to NDACC/FTIR data, Atmos. Meas. Tech., 7, 2719–2732, https://doi.org/10.5194/amt-7-2719-2014, **2014**