# v.1.0.1 User's Guide & Workflow Documentation

SR-lite: Automation and evaluation of surface reflectance estimates from multispectral VHR imagery

## Overview: What is SR-lite?

SR-lite (Montesano et al., in prep.) refers to an algorithm to estimate surface reflectance (SR) of multispectral very high resolution (VHR) input (eg, Maxar). The SR-lite workflow (Figure 1) is formalized in containerized code that returns SR estimates for VHR imagery (including WorldView-2/3/4 and GeoEye-1) at 2 m spatial resolution from input top-of-atmosphere (TOA) reflectance estimates and spatially and temporally coincident reference Landsat estimates of SR. Results are derived per-band using a linear model where SR = m(TOA) + b, and returned as Cloud Optimized GeoTIFFs. Version 1 of SR-lite uses the continuous change detection and classification (CCDC) algorithm (Zhu and Woodcock 2014) to generate modeled reference SR for the day corresponding to the input VHR dataset.

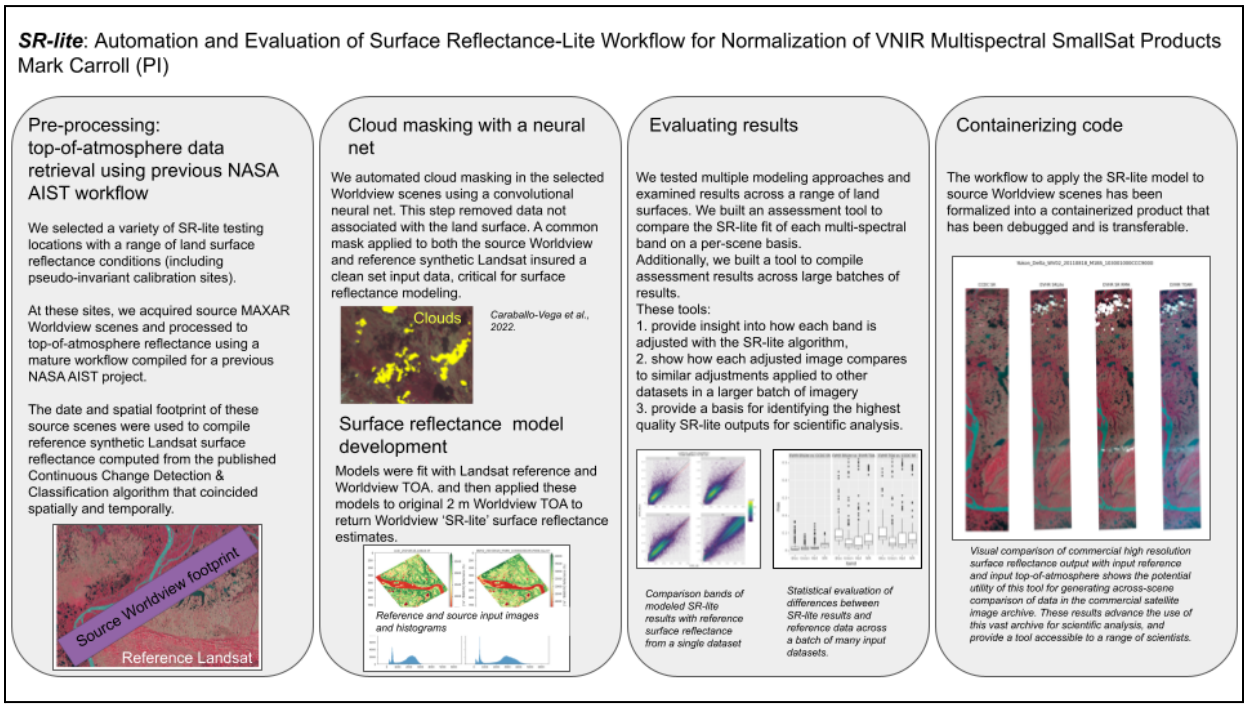


Figure 1. The SR-lite workflow models land surface reflectance using cloud-masked input top-of-atmosphere VHR reflectance and coincident reference surface reflectance estimates from Landsat. This workflow is run in containerized code to produce 2 m Cloud-Optimized GeoTIFFs.

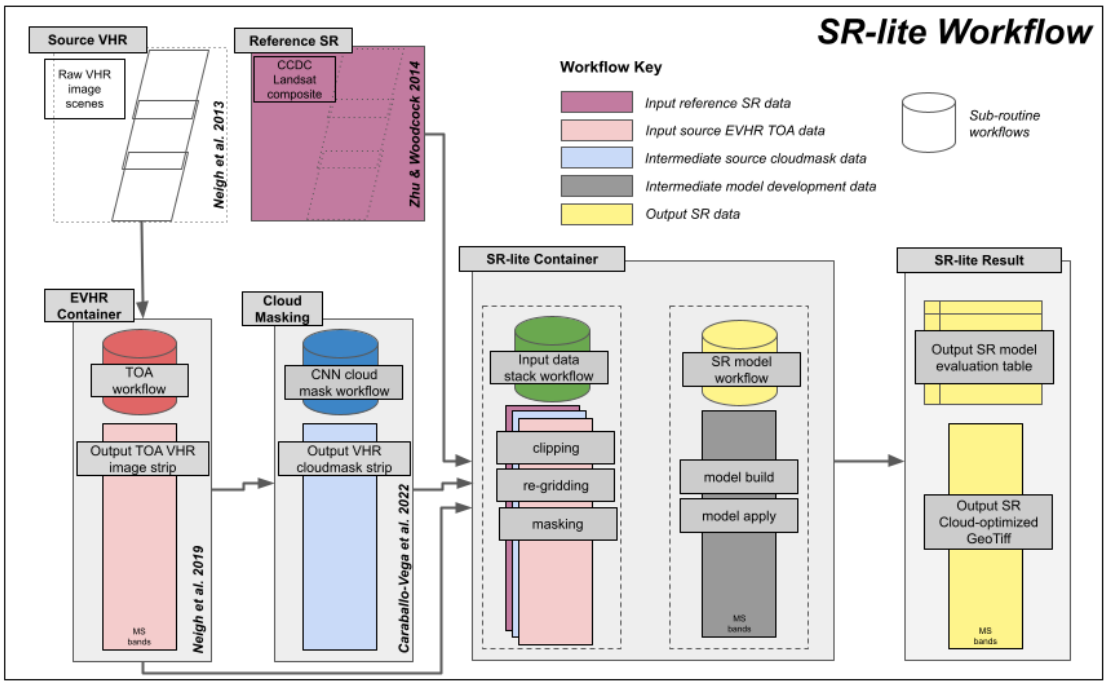


Figure 2. The SR-lite workflow.

# Quick reference for running SR-lite

SR-lite runs in a container, on a per-TOA basis, and requires 3 inputs:

1. VHR TOA geotiff
2. VHR TOA cloudmask geotiff
3. Reference surface reflectance multispectral geotiff

The SR-lite *Python* application is called using a *Singularity* container from a Linux terminal with the following general format:

```
$ singularity run -B <local path(s) to mount> <container name> python
<python application> <runtime parameters>
```

## Mandatory runtime parameters:

- -toa_dir - directory containing TOA 2m (default suffix = toa.tif)
- -target_dir - directory container model data (default suffix = ccdc.tif)
- -cloudmask_dir - directory containing cloudmasks (default suffix = toa.cloudmask.v1.2.tif)
- -bandpairs - list of band pairs to be processed [(model band name B, TOA band name B), (model band name R, TOA band name R) …]
- -output_dir - directory containing results for this specific invocation

## Optional runtime parameters:

- --regressor - Choose regression algorithm [‘rma’,'simple', 'robust']
- --cloudmask - Apply cloud mask values to common mask
- --csv - Generate comma-separated values (CSV) for output statistics
- --band8 - Create simulated bands for missing CCDC bands (C/Y/RE/N2)
- --xres - Specify target X resolution (default = 30.0).
- --yres - Specify target Y resolution (default = 30.0).
- --toa_suffix - Specify TOA file suffix (default = -toa.tif').
- --target_suffix - Specify TARGET file suffix (default = -ccdc.tif').
- --cloudmask_suffix - Specify CLOUDMASK file suffix (default = -toa.cloudmask.v1.2.tif').
- --clean - Overwrite previous output
- --qfmask - Apply quality flag values to common mask
- --qfmasklist - Choose quality flag values to mask [default='0,3,4']
- --thmask - Apply threshold mask values to common mask
- --thrange - Choose quality flag values to mask [default='-100, 2000']
- --pmask - suppress negative values from common mask
- --debug - [0=None, 1=trace]

Sample Invocation:

```
$ singularity run -B
/panfs/ccds02/nobackup/people/iluser/projects/srlite,/panfs/ccds02/nobackup/people/gtamkin,/home/gtamkin
/.conda/envs,/run,/explore/nobackup/people/gtamkin/dev,/explore/nobackup/projects/ilab/data/srlite/produ
cts /explore/nobackup/people/iluser/ilab_containers/srlite_1.0.1.sif python
/usr/local/ilab/srlite/srlite/view/SrliteWorkflowCommandLineView.py -toa_dir
/panfs/ccds02/nobackup/people/iluser/projects/srlite/test/input/baseline -target_dir
/panfs/ccds02/nobackup/people/iluser/projects/srlite/test/input/baseline -cloudmask_dir
/panfs/ccds02/nobackup/people/iluser/projects/srlite/test/input/baseline -bandpairs "[['blue_ccdc',
'BAND-B'], ['green_ccdc', 'BAND-G'], ['red_ccdc', 'BAND-R'], ['nir_ccdc', 'BAND-N'], ['blue_ccdc',
'BAND-C'], ['green_ccdc', 'BAND-Y'], ['red_ccdc', 'BAND-RE'], ['nir_ccdc', 'BAND-N2']]"
-output_dir
/explore/nobackup/projects/ilab/data/srlite/products/srlite_1.0.0-baseline/srlite_1.0.0-ADAPT-cli/Whites
ands/srlite-1.0.0-rma-baseline
--regressor rma --debug 1 --pmask --cloudmask --clean --csv --band8
```

The output data is delivered to a directory specified in the program call (`-output_dir`). There are two outputs:

1. The image data is output in COG format with the following naming convention: <SENSOR>_<YYYYMMDD>_<CATID>-sr-02m.tif.
2. The regression results are output as *.csv* files (see Table 1). They contain linear model (slope and intercept) coefficients along with SR-Lite performance statistics. .

# Methodology

The SR-lite workflow consists of the methodology detailed below. In summary, it stacks all input layers into a common modeling grid (eg, 30 m; coarser than the input TOA) grid for the input TOA spatial extent, derives a mask from invalid data in each of the 3 input datasets, applies that mask to the input TOA and reference SR to remove all invalid pixel, builds an SR model at 30 m resolution, and applies that model to the original VHR TOA at 2m spatial resolution.

## Computing TOA reflectance for VHR

The Enhanced Very High Resolution (EVHR, Neigh et al. 2019) workflow is used to produce TOA geotiffs for any VHR scene or a mosaic of a sequential collection of scenes (strip). A multi-spectral stacked geotiff is returned that is georeferenced to the local Universal Transverse Mercator coordinate system in a grid with a resolution native to the input TOA from the VHR imagery (2 m).

## Identifying valid pixels by masking cloud cover in VHR TOA

To identify valid surface pixels, we applied a convolutional neural net (CNN) algorithm to mask cloudcover for each Worldview VHR dataset (Caraballo et al 2022). The mask is returned as a binary map that separates cloud from non-cloud pixels. The development of this algorithm is on-going. In some cases, transparent cirrus clouds are not identified as such, while some very bright non-cloud surfaces (smooth snow cover or bright lichen ground cover extents) may be mis-classified as clouds.

## Compiling reference surface reflectance

To compile a reference of surface reflectance we derived Landsat-derived SR estimates using the CCDC algorithm. CCDC model parameters are generated from all available Landsat 4/5/7/8/9 Tier 1 Level 2 surface reflectance observations (masked to exclude cloud, cloud shadows, snow, pixels that are saturated in any band, and gaps). Then, we use the projection, bounding box and acquisition date of the input VHR TOA scene to generate a synthetic (modeled) map of estimated SR based on the CCDC model parameters. The CCDC parameters are based on the Landsat time-series inputs and incorporate seasonality,trends, and disturbances in the Landsat surface reflectance record.

## Constructing a modeling data stack: re-gridding and masking

We compile an input filename list of the 3 datasets (reference SR, input TOA, and cloud cover mask of input TOA) that will be warped, aggregated, and masked before model building.

### Regridding

The input files are regridded by warping each to the projection of the input TOA and re-gridding to a coarsened (30 m) pixel resolution. We warp the reference data to match the projection of the input TOA and re-grid using mean for the reflectance bands and mode for the cloud mask. Regridding the 2 m inputs to the coarser spatial resolution of the 30 m reference surface reflectance at this stage makes for more efficient model building.

### Masking

After regridding, we build a common mask that includes all “no data” collected from both reference and input TOA datasets. This mask is thus a union of all input “nodata”, and includes as “nodata” the masked cloudcover pixels from the input TOA.  Both reference and input TOA are masked with this common mask so that the same set of pixels are present in each dataset.

At this point, each input TOA pixel has a corresponding reference pixel, and the data is ready for model building.

### Building and applying the SR-lite model with the data stack

We use the re-gridded and mask data stack to build a linear model to describe the relationship of the input TOA (dependent) to the reference SR (independent). The models are applied bandwise for the blue, green, red, and near-infrared bands (Band 7, or NIR1, in the case of Worldview-2/3/4), where a given TOA band is matched to the closest corresponding reference band based on the central wavelength. We provide a choice of 3 linear models that are applied to each bandwise pairing (eg. $TOA_{blue}$ ~ f($SR\_ref_{blue}$) of dependent and independent data in this 30 m data stack. These choices are:

1. *simple:* ordinary least squares regression from *sklearn* using the *LinearRegressor* module.
2. *rma* : Python's *pylr2* implementation of reduced major axis (RMA) regression to each
3. *robust*: Python's *sklearn* implementation using the *HuberRegressor* module.

The bandwise model fit from that is returned based on the model choice is then applied back to corresponding input VHR TOA band (2 m). This is done for each band of the input TOA to return the multi-band SR-lite surface reflectance estimates at 2 m spatial resolution. For the special case of an 8-band input TOA, the model coefficients for the 4 extra bands (that are not present in the reference Landsat-derived SR input) are weighted coefficients derived from (in the case where the focal input TOA band falls between 2 reference bands) the two nearest bands (yellow and red-edge), or the nearest band (coastal blue and NIR2).

## Description of Output

Output images are returned as multi-band Cloud Optimized GeoTIFFs (COGs) with a corresponding linear model result table (CSVs). These table includes:

1. a summary of the correction model coefficients applied bandwise to each pixel of the TOA,
2. statistical results of the SR-lite comparison with input SR reference data and input TOA

The image products inherit the projection, extent, and cell sizeof the EVHR TOA outputs, typically the local UTM coordinate system

Table 1. An example of the output CSV table for each SR-lite result. This table contains correction model coefficients, which result from the linear regression between VHR TOA 30m and the chosen reference (e.g., CCDC) 30m, that are then applied to the input VHR TOA 2m to create the output grid (COG format). The performance statistics compare SR-Lite outputs to corresponding reference values. Note: in this example output there are no performance statistics for input TOA bands that don't have corresponding reference bands (eg. RedEdge is present in Worldview 2 input TOA's, but reference CCDC does not have a corresponding band).

Therefore, the output coefficients are the result of the weights used to calculate a correction from the nearest existing reference band.

| band_name | model | intercept | slope | r2_score | explained_variar | mbe | mae | mape | medae | mse | rmse | mean_ccdc_sr | mean_evhr_srlite | mae_norm | rmse_norm |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Blue | rma | -0.088426766 | 1.128632282 | 0.947089612 | 0.947089718 | 5.06E-05 | 0.025088347 | 0.071376474 | 0.0177 | 0.00127554 | 0.0357147 | 0.356482825 | 0.356432245 | 0.070377435 | 0.100186312 |
| Green | rma | -0.03953199 | 1.105702518 | 0.975006745 | 0.975006845 | 5.07E-05 | 0.018795305 | 0.048212171 | 0.0138 | 0.000645313 | 0.025403009 | 0.419271043 | 0.419220304 | 0.044828532 | 0.060588513 |
| Red | rma | -0.026764598 | 1.055592341 | 0.960893363 | 0.960893444 | 4.98E-05 | 0.024189597 | 0.053166193 | 0.0172 | 0.001197043 | 0.034598303 | 0.481573152 | 0.481523381 | 0.05023037 | 0.071844336 |
| NIR | rma | 0.010801264 | 1.028431628 | 0.969368953 | 0.969369056 | 5.01E-05 | 0.019911501 | 0.039614134 | 0.0144 | 0.000745825 | 0.027309796 | 0.534534573 | 0.534484518 | 0.037250165 | 0.051090795 |
| Coastal | rma | -0.088426766 | 1.128632282 | NA | NA | NA | NA | NA | NA | NA | NA | NA | NA | NA | NA |
| Yellow | rma | -0.032803575 | 1.079294455 | NA | NA | NA | NA | NA | NA | NA | NA | NA | NA | NA | NA |
| RedEdge | rma | -0.012527137 | 1.045298431 | NA | NA | NA | NA | NA | NA | NA | NA | NA | NA | NA | NA |
| NIR2 | rma | 0.010801264 | 1.028431628 | NA | NA | NA | NA | NA | NA | NA | NA | NA | NA | NA | NA |

*Table 2. Description of output linear model results table column names.*

| ***Column Name*** | ***Description*** |
|---|---|
| *band_names:* | *the name of the input TOA bands* |
| *model* | *he name of the linear model choice* |
| *intercept, slope* | *the values of the coefficients for each model relating input TOA to reference SR* |
| *r2_score* | *coefficient of determination regression score, representing the proportion of variance of the dependent variable explained by the independent variables. https://scikit-learn.org/stable/modules/model_evaluation.html#r2-score* |
| *explained_variance* | *the "r-squared value" representing the proportion of variance between the dependent and independent variables explained by the linear model. Similar to r2_score but does not account for systematic offsets in the prediction.* |
| *mae* | *mean absolute error in the prediction* |
| *mbe* | *mean bias in the prediction* |
| *mape* | *mean absolute percentage error from the model residuals* |
| *medea* | *median absolute error (https://scikit-learn.org/stable/modules/generated/sklearn.metrics.median_absolute_error.html#sklearn.metrics.median_absolute_error)* |
| *mse* | *mean squared error from the model residuals (https://scikit-learn.org/stable/modules/model_evaluation.html#mean-squared-error)* |
| *rmse* | *root mean squared error from the model residuals (y)* |
| *mean_ccdc_sr* | *mean surface reflectance of the reflectance target (CCDC) in the final (fully masked) model array.* |
| *mean_evhr_srlite* | *mean surface reflectance of the SR-Lite product in the final (fully masked) model array.* |
| *mae_norm* | *normalized mean absolute error, normalized by dividing the mae by the mean CCDC reflectance.* |
| *rmse_norm* | *normalized root mean square error, normalized by dividing the rmse by the mean CCDC reflectance.* |

## Evaluating results

SR-lite results can be evaluated in 2 ways, on an individual image basis, and by batch for a set of many images for a particular area.

### Individual image evaluation

For each individual SR-lite image output we compare for each MS band:

1. the image's SR-lite vs. input TOA reflectance values (30 m).
2. the image's SR-lite vs. input reference reflectance values (30 m).

A suite of regression metrics for describing the strength of the model used to derive the output SR-lite from the input TOA are returned to a CSV file.

### Global evaluation

For a batch of SR-lite image output, typically consisting of imagery with a variety of sun-sensor-geometry image acquisition characteristics, we compare model results of the image's SR-lite vs TOA reflectance values. Plots showing all individual linear model results show which images were likely not well corrected (slopes significantly less than 1). This can often be a way for users to easily identify images of particularly poor quality relative to the rest of their data.

## Future considerations

SR-lite is not yet designed to handle the following circumstances:

1. **Snow**: Snow is masked from CCDC so reference image during partial snow season generally will depict snow-free reflectance. We should be masking snow from EVHR before applying any regression with CCDC reference, however we do not currently have a snow mask to use.
2. **Cloud shadow**: Artificially dark areas in EVHR that are not masked can greatly affect regression
3. **Water**: Especially with high sediment loads or ice, not a stable reflectance target
4. **Saturation**: More common in QB2/GE1 images. Often associated with snow. Regression not appropriate with saturated pixels. Saturated pixels are excluded from the CCDC model.
5. **Out of bounds values**: Reflectance < 0 or > 1 are sometimes far outside the bound of the valid 0-1 range.

SR-lite output in an upcoming version will include:

1. VHR acquisition geometry plotted on polar coordinates, useful for understanding some of the variation within a batch of SR-lite output.
2. SR-lite output assessment notebooks.

SR-lite adjustments to consider:

1. Geographic shift: CCDC (generated on a 30m grid) gets an extra arbitrary resampling step to match one corner of the TOA image. This results in a subpixel shift, possibly accompanied by a smoothing if NN resampling is not used.
2. Algorithm speed: mode resampling for cloud mask could be switched to average as long as pixels with a decimal value are then classified as cloud (conservative - more cloud

area than the input). Mode may be an order of magnitude or more slower and is one of the slowest steps in the algorithm.

## References


Caraballo-Vega, J.A., ML Carroll, CSR Neigh, M Wooten, B Lee, A Weis, M Aronne, WG Alemu, Z Williams. 2022. Optimizing WorldView-2,-3 cloud masking using machine learning approaches. Remote Sensing of Environment, 284, https://doi.org/10.1016/j.rse.2022.113332

Montesano, P. M., Carrol, M., Neigh, C. S. R., Macander, M. J., Caraballo-Vega, J. A., Frost, G. J., Tamkin, G. 2023. Producing a science-ready commercial data archive: a workflow for estimating surface reflectance for high resolution multispectral imagery. (in prep.)

Neigh, C. S. R., Carrol, M.L., Montesano, P. M., *et al*. 2019. An API for Spaceborne Sub-Meter Resolution Products for Earth Science. *IGARSS 2019 - 2019 IEEE International Geoscience and Remote Sensing Symposium*, 2019, pp. 5397-5400, doi: 10.1109/IGARSS.2019.8898358.

Zhu, Z., and Woodcock, C. 2014. Continuous change detection and classification of land cover using all available Landsat data. Remote Sensing of Environment, 144, 152-171. https://doi.org/10.1016/j.rse.2014.01.011