Anonymous Referee #1

Received and published: 2 May 2012

Review Dybkjaer et al (os-2012-30) “Arctic surface temperatures from METOP/AVHRR compared to in situ ocean and land data”

General comments
The authors have defined an operational method to derive Ice Surface Temperature (IST) from METOP/AVHRR over the Arctic region. They have gathered validation data from buoys and ship measurements and from two sites: in inland Greenland (air temperature measurements) and in a fjord in NE Greenland. This is an interesting paper, addressing a challenging task and which should be published after the authors have accounted for the remarks below. My main concern is the IST algorithm they use which has been developed for NOAA12 and is certainly not optimal for their purpose. I suggest that in a future work they use NWP model profiles, both for improving the retrieved IST accuracy and their cloud mask.

Specific comments
P.2, line 27. The authors should be more precise in their description of the existingproductsÂ˘ a: are they not available in real time? or available in real time but not reliable? or available in limited historical time series, etc.. A quick review of the methods used in the MODIS or the AVHRR case would be also quite useful.

We have specified the two essential IST data sets mentioned in the text and accounted for the unique contribution of this METOP IST product. Following is merged into the introduction:

“The Pathfinder data set is well suited for climatologically studies, but can not be used for recent or real time ice surface temperature analysis, due to irregular data set updates. Further, the Polar Pathfinder spatial resolution is 5km, which makes it less suitable for fine scale mapping and analysis. The MODIS IST product has very similar characteristics to the METOP IST product (see section 6), with product timeliness and sensor continuity as the main differences. Timeliness and data continuity are essential issues when the model communities setup data validation and assimilation schemes (Stammer et al., 2007). The MODIS sea ice products have a time lag of days, from observation to product availability, and the timeliness of present IST product is a couple of hours. The METOP AVHRR data stream, that is used for this IST production, is guarantied continuity and is scheduled until at least 2020, in contrast to the MODIS data stream that will end with the current Aqua and Terra missions.”

The reference (Stroeve et al., 2001) in the introduction is removed.

P.5, line 10. Thresholding T11 temperatures to determine ice free, marginal ice zoneand ice covered areas is only a gross rule that shows many exceptions. Ice areas in summer can be observed in summer with T11 around 0 C, due for instance to air temperature inversion. Did the authors envisage more efficient methods (Bayesian approachÂ˘ a?).

Yes, we are planning to replace this rather simple T11 thresholding approach with either thresholding of the corresponding ice concentration or a Bayesian approach based on statistics from ice concentration products.

However, a histogram of all ice concentration associated with T11 temperatures colder than -4.2C show an ice concentration mean of approximately 97% with a gaussian like shape (see Illustration 1) and thus indicating that at least the closed ice area is well defined by this means.

A few changes are applied to section 7 (Future works):

"Further development of MAST will be done in the OSISAF. This will including a future re-calibration of the IST algorithm, enhancement of cloud screening procedures, and tests of other surface type classification procudure to optimize algorithm decision making."

P.5, line 17. The IST algorithm is formally identical to a classical split window equation, could the authors briefly recall how the coefficients have been determined? The use of a IST algorithm defined for NOAA12/AVHRR and used for METOP/AVHRR is a very questionable solution. A better approach could have been to redefine an algorithm and its coefficients by using simulated brightness temperatures applied to arctic atmospheric profiles, as this is currently done for SST algorithms (see below). I understand the authors are pioneering a validation experiment, but adopting NOAA12 coefficients is clearly taking a risk.

We agree that it may not be optimal to adopt the NOAA12 calibration coefficients for the METOP IST algorithm. None the less, this solution seems to be equally good as other available solutions for the time being. We believe that the optimal calibration of the IST algorithm is achieved from an empirical relationship with in situ skin temperature data from positively cloud free conditions. Presently, we do not have sufficient in situ skin temperature data to perform such a calibration and we are left with the options of using either simulated top-of-atmosphere brightness temperatures for Arctic conditions or buoy and ship data for an empirical calibration.

The NOAA12 calibration performed by Key et al. is retrieved from simulated brightness temperatures at the AVHRR infrared bands in Arctic conditions. This calibration is therefore applicable for the METOP algorithm, that is based on the same instrument as the NOAA12 AVHRR instrument, with identical band width and nearly identical spectral response functions. So, one can say that the calibration is spectrally specific, rather than specific for a certain satellite instrument. In present paper the Key calibration is proven to be competitive with a calibration to the full buoy data set. This is explained in section 4.5, where a clarefying comment is added (see below – comment to P8 L24):

However, we do agree that a new algorithm/calibration must be part of the future IST algorithm development. We have added following:

"The NOAA12 calibration coefficients are retrieved from RTM modelled brightness temperatures for the AVHRR infrared channels and related to model skin temperatures (Key et al., 1997). The channel centre, -width and spectral response function of the NOAA12 and METOP AVHRR instruments are nearly identical. We therefore considered the applied calibration equally valid for METOP AVHRR data than for NOAA AVHRR data."

P.5, line 26Â˘ a: The reference for the EUMETSAT OSI SAF AVHRR SST algorithm should beÂ˘ a: EUMETSAT (2010)Â˘aLow Earth Orbiter Sea Surface Temperature Product User ManualÂ˘ a, SAF/OSI/CDOP/M-F/TEC/MA/127Â˘aÂ˘a(http://www.osi-saf.org).

Done.

P. 6, line 12Â˘ a: Since the authors have access to NWP ice temperature (NWPsurface) as explained in section 4.2, why did they choose to compare MAST results with NWP 2m air temperature (NWP2mT)Â˘a? For non Arctic specialists, they should explain here (rather than at the end of the paper) the ice versus air temperature relationshipÂ˘a?

The 2mt field was convenient for plotting as the 2mt model field is global, in contrast to model IST and SST fields that are 2 separate parametres in the data stream from ECMWF. However, it is of cause a relevant point and we agree that the surface temperature and 2mt fields can be rather different for sea ice in particular. The NWP panel in the figure is changed to surface temperatures (See illustration 2). The discriptive text is adjusted accordingly.

P.8, line 24Â˘ a: The adjustment of coefficients on 4 days of data is very far from ideal and will lead to a very locally adapted algorithm.

The adjustment of coefficients, or “re-calibration tests” as they are called in the paper, are not performed to come up with new calibration coefficients, merely to test the performance of the applied algorithm. By 'recalibrating' to the entire buoy data set one find the best possible empirical calibration. This is now specified in the text, by adding this:

“Hence the re-calicration is not performed to establish new calibration coefficients, but to compare the best possible empirical calibration from the Arctic buoys and ISAR measurements to the operating algorithm. If the re-calibration tests do not improve the performance significantly, the dominant errors are associated with other issues than algorithm calibration.”

P.10, line 6Â˘ a: The acronym table should be introduced earlier in section1.

A reference to the acronyme table is given in section 1.

P.10, line 8Â˘ a: I do not understand what represents the number of cases (20000cases). The authors use a 4x4 pixel boxÂ˘a? and in this box each of this pixel is accounted for individuallyÂ ˘a? Are those individual pixel values compared independently to the in situ measurementsÂ˘a? Did the authors try any mean or median value in the boxÂ˘a? P.10, line 18Â˘ a: The MUsummit (air temperatureÂ˘a?) data are used in the validation experimentÂ˘ a: Using air temperature is surprising and should be justified (already mentioned above)Â˘a

Yes, each cloud free MIST data inside the 4x4km area is accounted for individually – giving approximately 20000 cases.

The error statistics for the MUisar data set was analysed using both mean of all induviduals, mean and median MIST values, without finding clear indications of which measure to use. The standard deviation of errors were practically equal and we decided to threat all MIST data individually. In figure 6 the MIST data are plotted with minimum, maximum and median values and the MISTnewcalibration is plotted as the average value. This is specified in section 5 with this comment:

“This validation strategy was based on experience from MUisar data. The MUisar error statistics was analysed using both mean of all induvidual data pairs, mean and median MIST values, without clear indications of a best measure. Thus, it was decided to threat all MIST-OBS data pairs individually.”

The motivation to match MIST with summit air temperatures is that NWP 2mt and the deduced skin temperatures can be tenth of degrees off the measured 2mt and that the observation network is extremely sparse. The correlation between 2mt and skin temperature is due to level and homogeneous surface considered high on Summit, in contrast to 2mt and skin temperature on the sea ice, where even small fractures in the sea ice can change the heat flux between sea and air drastically. Therefore we consider differences between 2mt and skin temperature on Summit to be a bias issue mainly, and that the surface and 2m temperature correlation is high. This is commented in the discussion (section 6):

“An intercomparison of surface temperature observations, NWP skin temperatures and MIST would be ideal, but the only available long term temperature observations on Summit are 2m temperature and surface pressure. However, 2m and surface temperatures are comparable at Summit, due to level and homogeneous surface conditions, which results in a very high correlation between Summit 2m and surface temperatures (Hall et al., 2004a)”

P.10, line 22Â˘ a: The disappointing results of the re-calibrated algorithm is due to the fact that it is narrowly specialized for the location and the time period of the ISAR experiment.

We believe this is a misunderstanding based on the re-calibration issue answered above. As mentioned above, the re-calibration is used to estimate the error contribution from a possible poor NOAA12 calibration. However, the result show that we practically gain nothing from re-calibrating the algorithm to the full OBSarctic data set, thus indicating that the NOAA12 calibration is working well – or at least, contribute much less that other error sources. A clearifying comment is added in the text in section 5 (Results):

“... i.e. indicating that the adopted NOAA 12 calibration coefficients work well for the METOP AVHRR instrument and that erroneous cloud screening is a dominating source of error.”

P.10, line 29: An improvement of the agreements with the MUisar datasets is no surprise, for the same reason.

The significantly better error statistics obtained from the re-calibrated MUisar data set is obtained because errors from clouds, time lag and surface homogenity are minimized, thus leaving much more weight to the remaining error sources. So, we agree it is not surprising.

P.11, line 10: The figures obtained with ISAR measurement are quite interesting. Do the authors think they are representative of what could be obtained with an adequate algorithm, a good cloud mask and a reliable in situ. In other words, is this the potential accuracy of a TIR based IST method?

As mentioned above, the OBSisar data are recorded under spatially homogeneous conditions, with practically no time-lag between OBS and MIST and in near clear sky conditions. So – yes – we consider this to result to be 'very' upper limit of satellite based IST performance.

P.20, line 21: The Diurnal cycle shown in figure 3 and 6 is impressive and surprising, at least for non arctic specialists. Can the authors give an indication of the amplitude of this diurnal warming, since it is not easy to infer from the figures.

This information is written in the respective captions.

P.11, line 26. It is difficult to understand how the same split window algorithm can provide atmospheric correction at sea surface level and at 3200m altitude, where the atmospheric absorption should be much lower, could the authors comment on that?

The algorithm is calibrated to Arctic sea surface level, where the atmospheric water content is high relative to the atmospheric conditions at Summit. This may result in a too large weight (coefficient) on the atmospheric correction term of the IST algorithm and consequently in a cold bias when applied in a dryer atmosphere than that for which the algorithm is tuned. This is also indicated by the validation results. A comment on this is added to the discussion in section 6:

"However, part of this bias difference may be caused by the calibration. MIST is calibrated to Arctic sea surface conditions, where the atmospheric water content is high relative to the conditions at Summit. This may result in a too large weight (coefficient) on the atmospheric correction term of the IST algorithm and consequently in a cold bias when applied in dryer atmospheric conditions."

P12, line 3: The comparison of OBSsummit and NWP2mT is not quite clear: the bias is small but the standard deviation large; how do this fit with the author's explanation of OBSsummit being assimilated in the model?

A data assimilation scheme will pull a model towards the observation, thus towards zero bias. This becomes more pronounced with fewer observations.

P13, line 27: There must be errors also linked to the algorithm itself, even though I agree the error trend with satellite zenith angle is encouraging. This error is illustrated by the fact that the original algorithm showed a negative bias of –1.81 K against ISAR measurements according to table 2. This algorithm linked negative bias probably contributes the negative biases recorded in table 3.

We wish not to go into a discussion about algorithm errors, but we have added some reservations regarding assumed algorithm errors in the discussion.

P14, line 1: This discussion should have been introduced earlier (in section 4.1 for instance)

Following sentense is added to section 4.1:

"Different in situ data sources can result in rather dubious validation results as surface and air temperatures can differ by several degrees. This is discussed in section 6."

P.14, line 14 and p.16 line 16: I am surprised that the authors envisage to “recalibrate” their algorithm with in situ measurements, knowing how scarce good matchups conditions are in the Antarctic. In my opinion this is clearly a weakness in the authors’ approach. My recommendation would be to use a NWP model based approach to determine an optimal retrieval algorithm. To do that the authors could use either radiosonde profiles or NWP model profiles, and build up a simulated BTs by using a fast radiative transfer model such as RTTOV and realistic surface temperatures (see Francois et al. RSE 2002 or Merchant and LeBorgne JAOT 2004) . Similarly they could introduce model profiles in their operational retrieval scheme; this would guarantee a better adaptation of the atmospheric correction to actual atmospheric conditions (including altitude effects). they could either adopt a full Optimal Estimation technique (OE, Merchant et al. RSE 2008), or a Bias Correction (BC) method (Le Borgne et al RSE 2011, Petrenko et al. RSE 2011). These methods are now currently used in SST retrieval and I do not see why they cannot be adapted to IST retrievals, providing a correct ice emissivity model is available. Since improving the cloud mask is an other challenge, comparing the true IR and the simulated IR values gives a good indication on how cloud contaminated is the pixel (if not using a full Bayesian method) Interactive comment on Ocean Sci. Discuss., 9, 1009, 2012.

These comments are very inspirering and usefull and they will be considered in the future work of this product.

OTHER CHANGES:

Reference removed from text - p2:

Stroeve et al. (2001)

Reference removed from reference list:

Stroeve, J. C., Box, J. E., Fowler, C., Haran, T., and Key, J.: Intercomparison Between in Situ and AVHRR Polar Pathfinder-derived Surface Albedo Over Greenland. Remote Sens. Environ., 75, 360–374, 1998.

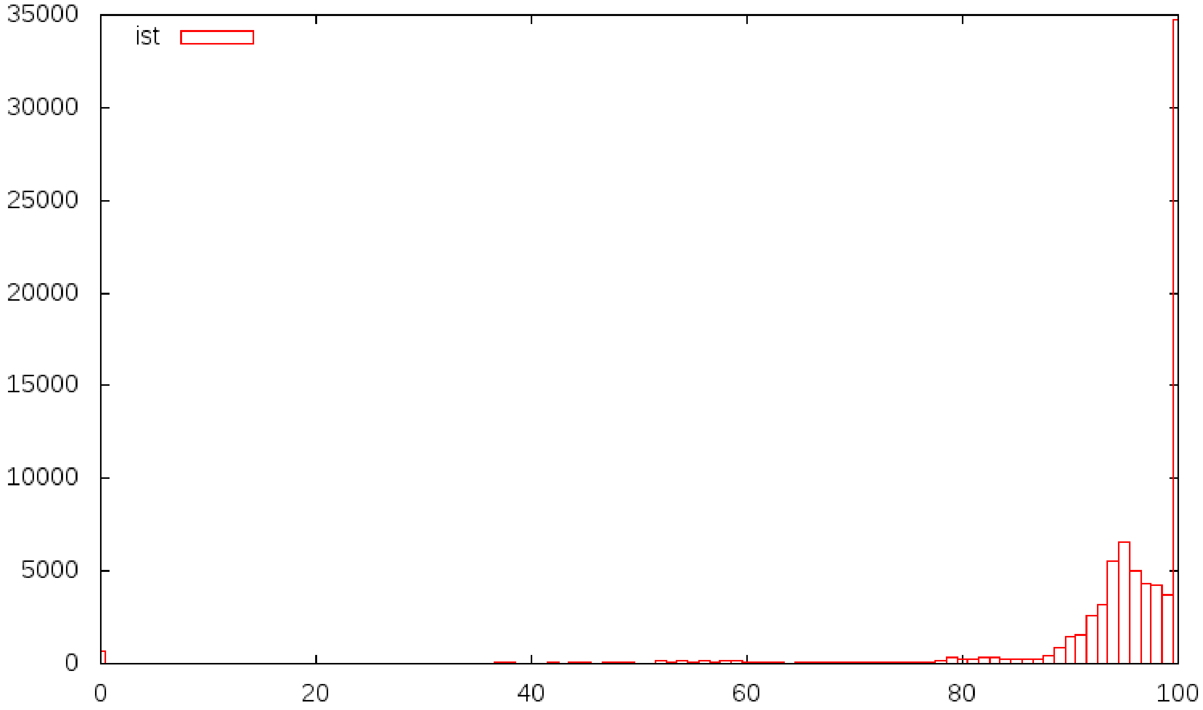

Illustration 1: Frequency distribution of ice concentration for all IST data colder than -4.2C. The gausian like distribution has mean ice concentration around 97%. The high frequency of 100% ice cover is caused by aggregating ice concentration equal or higher than 100%

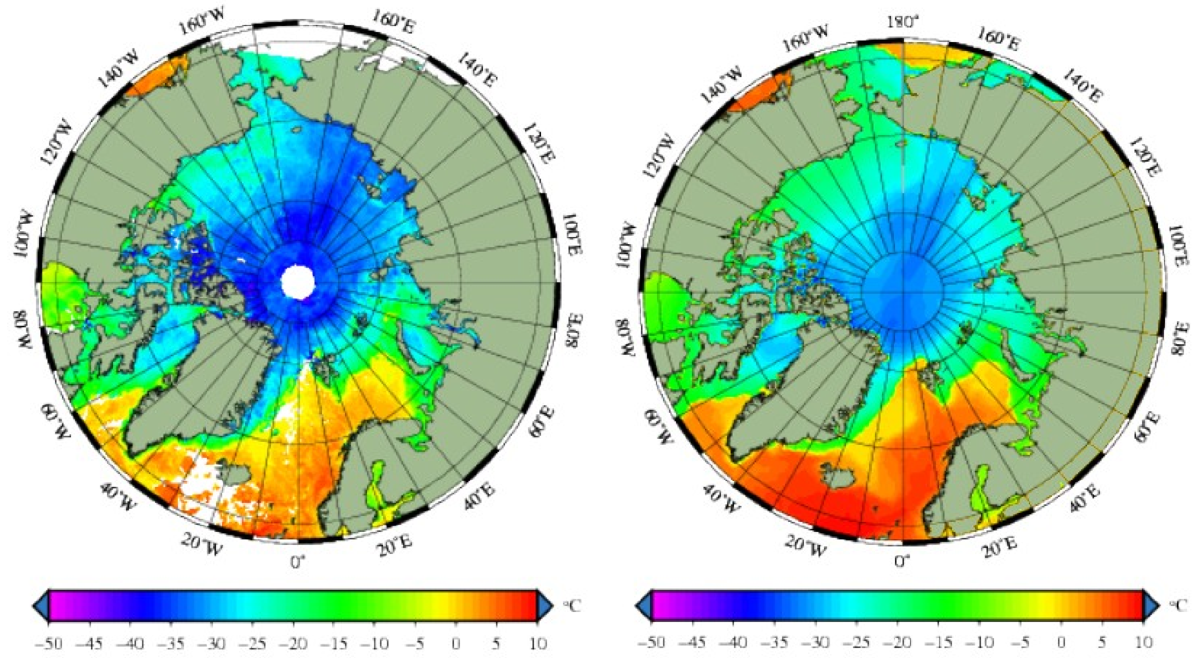

Illustration 2: New figure 4 with NWP surface temperatures in the right panel