WindBorne THINICE report

WindBorne Systems (contact: `joan@` and `john@windbornesystems.com`)

October 8, 2021

Table of contents

1 High level summary

WindBorne Systems participated in the 2021 THINICE data collection campaign, launching a total of 90 custom developed long-endurance sounding balloons in a campaign internally referred to as Medium-Scale Trial 2 (MST-2). Flights lasted from August 19 to September 13, with 50 units launched from Svalbard and 40 from Barrow, AK over that time. Each unit was actively controlled to maximize the value of collected data through a combination of frequent vertical transits to gather sounding data and navigation towards identified regions of interest. The following report outlines our preliminary analysis on the performance of the system and of the suite of meteorological sensors carried onboard.

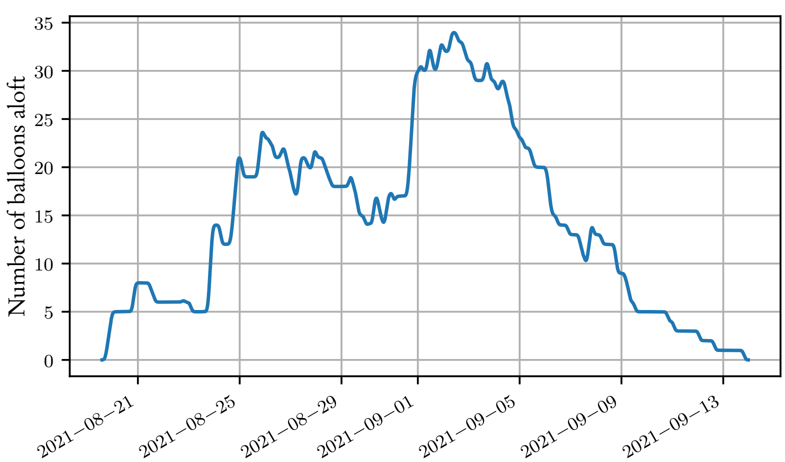

Figure 1: Progress of MST-2 over the course of the 25 days it lasted. The instantaneous peak was of 34 balloons aloft at the same time.

WindBorne accumulated 371 days of flight in total. Of those, 269 days were from balloons inside the domain (defined as the Arctic circle, $\phi > 66.5°$), resulting in 72.5% of the flight time in domain. The total horizontal distance traveled by the balloons was 432,106 km, equivalent to 1.12 times the mean distance to the Moon.

One metric of data collection that one can use is total vertical distance traveled by each balloon. This is a sum of the absolute value of the vertical distance traveled between each timestep of flight data, while ignoring small altitude fluctuations of less than 500 m. Including flights that terminated early due mechanical issues, our average vertical distance traveled per balloon was 196 km. Our flight with the maximum vertical distance traveled was 383 km. The total sum vertical distance traveled over all balloons is 17,672 km. The average vertical distance travelled per balloon equates to many multiples of the vertical distance traveled by routine radiosonde launches.

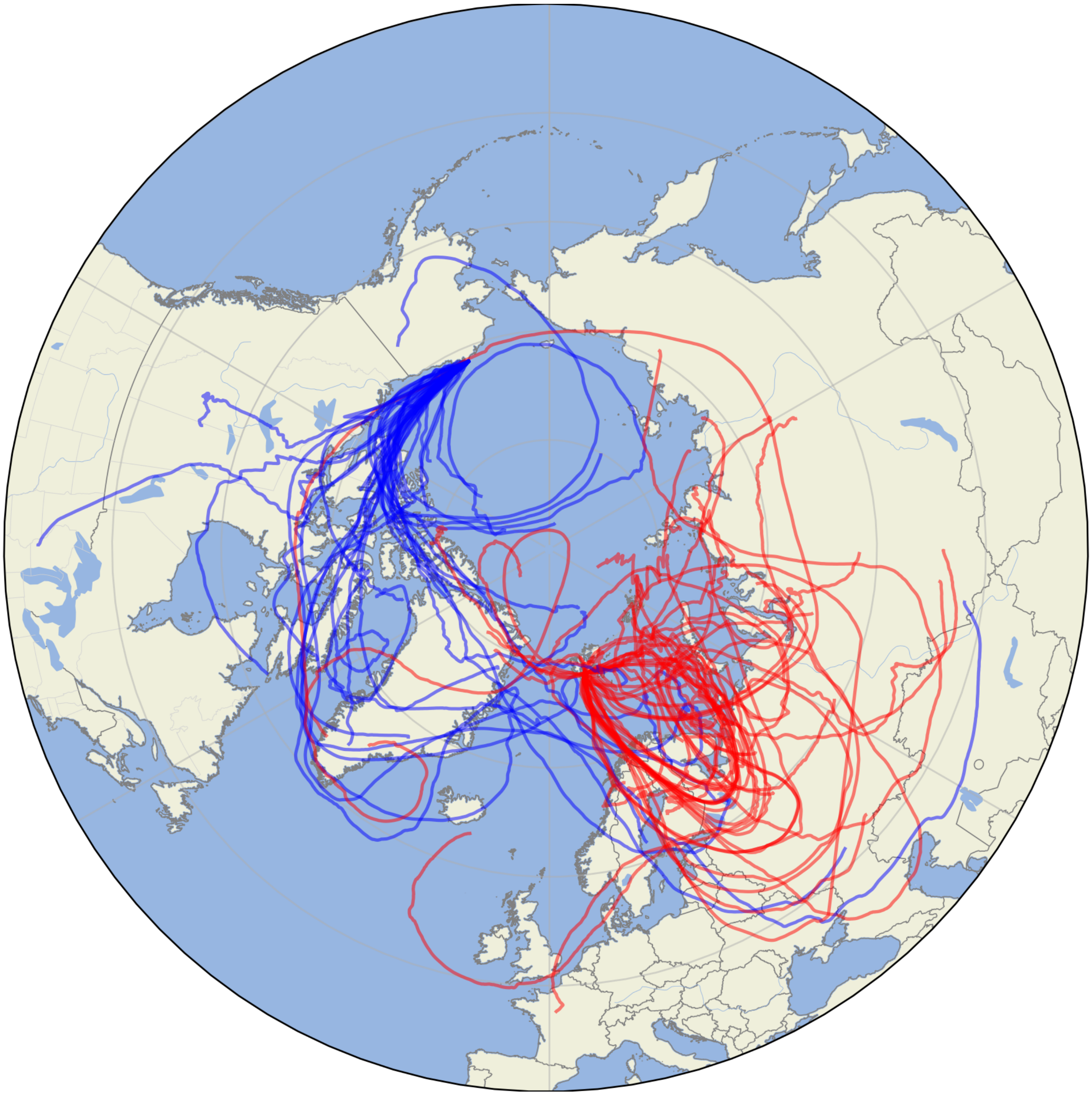

Figure 2: Map of all MST-2 launches. Launches from Alaska are shown in blue, while launches from Svalbard are shown in red.

2 Navigation

WindBorne's balloons are capable of a degree of navigation. There is often a range of wind speeds and directions throughout the range of altitudes that the balloon can fly at. By flying at a specific altitude, one can catch specific winds traveling in a desired direction to navigate to a target.

2.1 Methods

Balloon guidance

WindBorne navigates balloons relying on a combination of autonomous optimization methods, guided and tuned by a human directing the constellation. The software system that runs the autonomous balloon methods is called the "Planner." WindBorne's algorithms use wind forecasts to simulate the future trajectory of the balloon, and select a set of altitude waypoints for a given balloon to fly to in order to optimize a specified objective function. The given objective function is selected by a human who is directing the constellation of balloons, and can be a function of the location of the balloon, the location of other balloons, the balloon's altitude and data collection profiles, and a number of other simulated variables. One key factor that must be taken into account when guiding a balloon is that the more tightly a balloon's altitude is controlled, the more ballast it will use—and, as a result, the expected longevity of the flight will decrease. Additionally, flying at low altitudes in humid environments can cause icing, costing large amounts of ballast or taking the balloon out of the sky entirely unless corrective action is applied. How much the planner trades ballast as it tries to meet its navigational goals is called its "aggressiveness."

Launch selection

Selecting which launch sites to launch balloons from at a given time is a manual decision made by the person who is also directing the constellation of balloons. There will be many timing constraints based on contractor availability at the launch site, launch site weather, hardware availability, polar bears, etc. Launches are determined by looking at wind forecasts and trajectory simulations to see what times meet the site timing constraints and are best for balloon navigation. Proper selection of launch times plays a huge role in the coverage that one can achieve; if balloons were launched at uniform rates and only navigated after launch to meet desired targets, coverage would not be nearly as good.

2.2 Strategy

The general navigation strategy evolved as the trial went on. Early in the trial, the planner was set to be a bit too aggressive, and was optimizing to deliberately spread balloons out and target features of interest. This caused some of the early flights to not collect as much data as they otherwise might, due to early ballast exhaustion. Additionally, attempting to aggressively spread out balloons on a short time horizon actually results in less overall coverage; if they spread quickly but only last 3–4 days, that results in less area covered than slowly spreading out but flying for a period of 7–10 days.

Midway through the trial, the strategy was changed to make planning less aggressive. Rather than directly optimizing to spread out, after a group of new balloons were launched, half or more were put into "endurance mode." In this mode, the balloons would fly at higher altitudes (starting at 8–10 km), where they were safest, and optimizing just for minimizing ballast use. Balloons in endurance mode may not spread out as quickly, but over many days, they are likely to have spread out. Each endurance mode balloon would have reasonable odds of finding itself in a unique location 5 days in the future, with potential for it to be a region of interest. Balloons would eventually be taken out of endurance mode when they were in an interesting region.

While balloons were not in endurance mode, using active navigation was still useful, but in a less aggressive setting. Non-aggressive navigation was especially useful for keeping balloons in the domain while collecting data. In some instances, balloons were deliberately put into an aggressive mode in order to get near a target of interest, and it paid off many times.

2.3 Coverage of interesting features

One way to analyze the targeting abilities of the platform is to look at how many balloons were in features of interest at a given time. Throughout the duration of the data collection campaign, the THINICE meteorology team gave us target features in the form of a PowerPoint presentation. After the trial, the data from the figures from these presentations were scraped via Python, to extract the exact location of the given targets. These locations, in conjunction with the location data from the balloons, were used to create a visual to show the targeting, hosted here [link]. One can use the arrow keys on the linked page to scroll through time. A still frame the plot is show in Fig. 3.

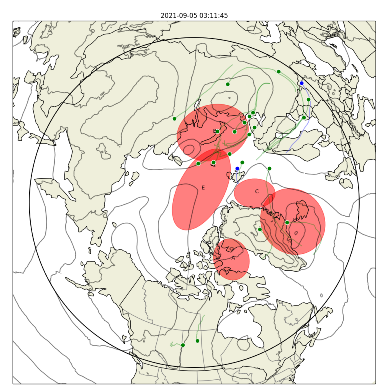

Figure 3: Frame showing balloon locations and target features. Each target feature is labeled with the letter that it was refereed to by during the observation campaign. The full figure can be seen here [link]. The left/right keyboard arrows can be used to navigate the frames as a function of time.

The locations of the features were then used to compute the number of balloons that were in a feature of interest at a a given time. Additionally, the converse was examined, the number of features that had balloons in them at a given time. These are shown below in Fig. 4.

A few things should be noted about this analysis. Before August 25th, targets were not given in written form, they were only conveyed verbally in a meeting, and thus there is no record for them. It was an important change for both high precision targeting in realtime, and for this post-analysis, to have targets be recorded. In addition, these targets were not the only areas of interest throughout the trial, as sensitivity analyses were done that helped

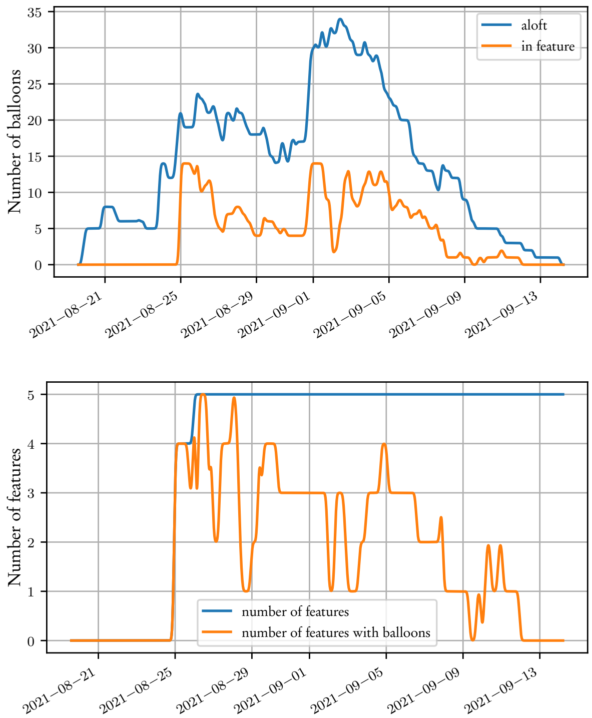

Figure 4: Number of balloons in a feature at each moment of the trial (top), and the number of features with balloons in them for each moment of the trial (bottom).

guide balloon trajectories, and there was a general goal of spacing out balloons to cover the entire domain. Still, these plots to give some indication that during the campaign, we were able to sample many of the areas of interest.

2.4 Specific targeting examples

While there were many more that could be discussed, below is a highlight of two targeting examples, one positive, and one negative.

Hurricane Larry

Near the end of the campaign, Hurricane Larry became a subject of interest, as it made it quite far north. This was late in the trial, and after we has already launched our last balloon, so it was difficult to plan to target. However, we were able to successfully target the storm and we got one balloon near the eye. This is shown in Fig. 5

Feature A

The large cyclone and TPV north of Alaska was a consistent challenge to target throughout the trial. The winds around the cyclone were quite consistent, with few opportunities to fly closer to the center. In addition, low altitude

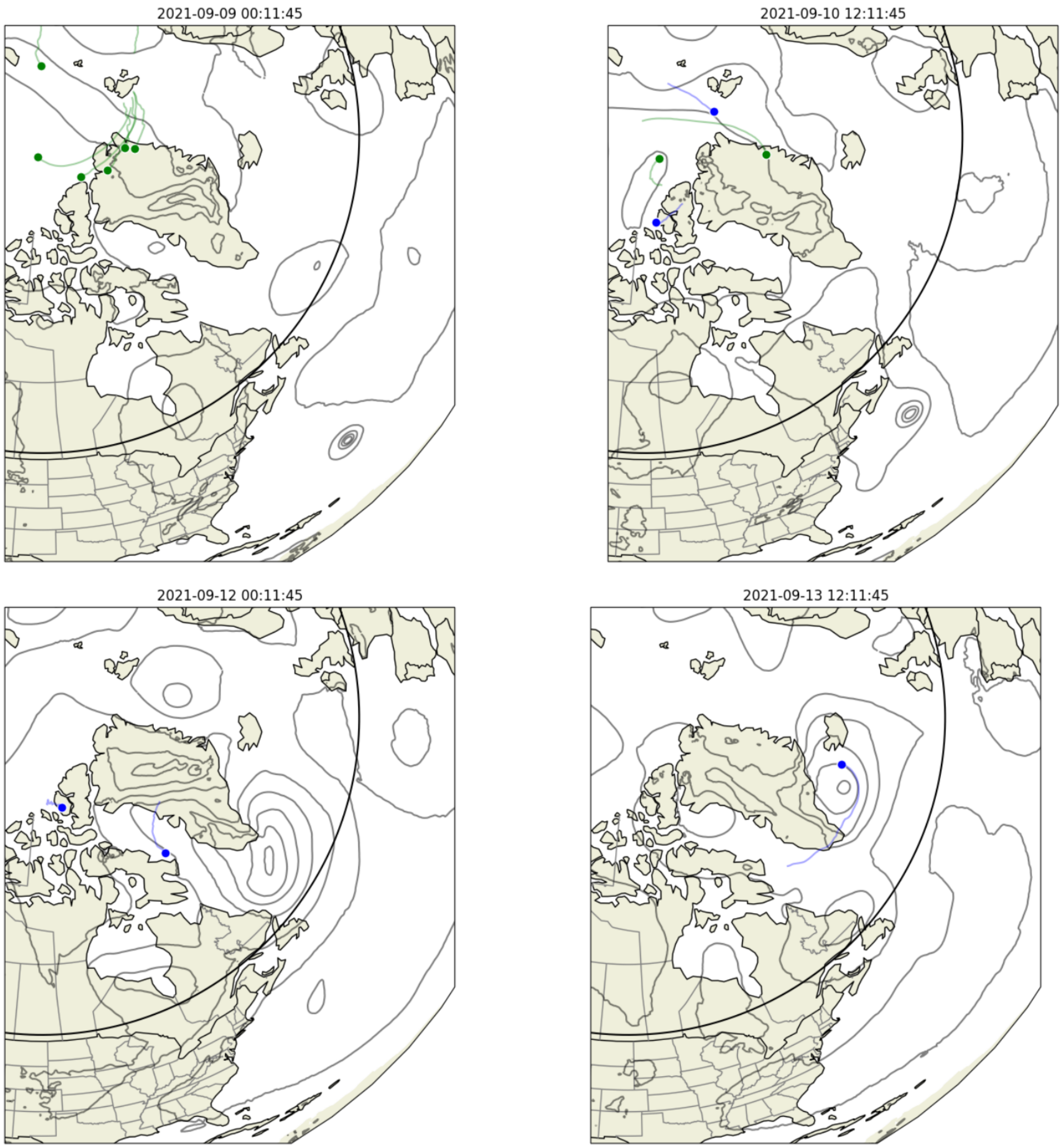

Figure 5: Targeting Hurricane Larry. We did not begin trying to target the storm until the time of the upper left image, after we had launched our final group of balloons. Over the next few days, we were able to target the balloon close to the center of the storm. We likely could have flown closer to the center than we ultimately did, but as this was our last balloon aloft, we didn't want to risk terminating the flight prematurely.

icing made low altitudes often not flyable. This targeting difficulty can be seen in the link visual from Fig. 3.

While we were not able to get close to feature A in this trial, we believe that there are strategies we could use in the future to navigate closer to systems like this. Had we started the launches earlier, we could have gotten endurance balloons into the region before the stable system with uniform winds formed. We saw something similar to this at the end of the campaign, where another large system was forming over the Barents Sea, and we were able to sample it as we had systems already in the area. Furthermore, had we been utilizing endurance mode from Svalbard earlier in the trial, we may have had more avenues to sample feature A.

2.5 Further navigation analysis

There are many questions left to ask about targeting abilities of this technology. It's currently not quantified to what degree targeting could be improved with better strategies and algorithms, versus what limits are fundamental. One may wonder, on average, how likely it is that we can hit a given target of interest. One may wonder how often we can hit a target when simulations show that we think we will. These are difficult questions to answer, as they are functions of many variables, and require expensive simulation and analysis to answer. Further exploration into these matters are left as an area of future research.

3 Sensors

The main sensor suite flown in MST-2 consists of a temperature sensor (glass bead thermistor), humidity sensor (heated E+E sensor), barometer, and GPS module (for winds and altitude). The first three are calibrated in house to rms of approximately 0.05 K, 1.4%RH, and 5 Pa; the GPS module does not require calibration. Of course, after deployment some integration errors can be expected, and when comparing them to either reanalysis or other radiosondes, representativeness, spacetime mismatch, and resolution all contribute additional variance. Additionally, a last minute addition was a radiation sensor that measures longwave and shortwave radiation fluxes, for both the upward and downward direction.

3.1 Comparison with reanalyses

Despite the limitations posed by representativenes and resolution, it is a useful exercise to compare the data collected to reanalysis such as ECMWF's ERA5 or the GFS-based FNL. Of course, the errors in this comparison themselves are a combination of all the sources of error: sonde errors, weather model biases, and interpolation and representativeness errors coming from the limited resolution of the models.

The general outlook is that the sensor data appears to be of high quality, comparable to existing radiosondes, despite a significantly more challenging sampling environment (with much lower aspiration velocities and much longer flight duration). The plots below show the median bias and median standard deviation as a function of altitude for the various sensors, as well as the same quantities computed for routine radiosonde launch stations in the Arctic circle in the same time period. It's very important to note that the radiosonde intercomparison is "cheating" here, as the data collected by those soundings was incorporated into the reanalysis itself, whereas WindBorne's data was obviously not assimilated into the reanalysis.

Fig. 6 shows the ERA5 intercomparison for winds, relative humidity, and pressure. Despite the fact that WindBorne's data was not assimilated into the reanalysis, the biases and variances are comparable to those of routine radiosonde launches from locations in the Arctic circle in the United States, Canada, and Greenland (sometimes even a bit smaller). Pressure was compared instead of geopotential height as a result of our internal infrastructure: for reference, 20 Pa at 8 km is approximately 4 gpm.

A lot of time was spent looking at a potential bias in the temperature sensor. When comparing against ERA5, there seems to be a very consistent bias of approximately $-0.25\,\mathrm{K}$ in the measurements (i.e. WindBorne's measurements read colder than ERA5). This was especially surprising given the lack of bias on the data collected in Feb–March 2021 in the eastern Pacific and western continental US, and the fact that the sensors were calibrated to within

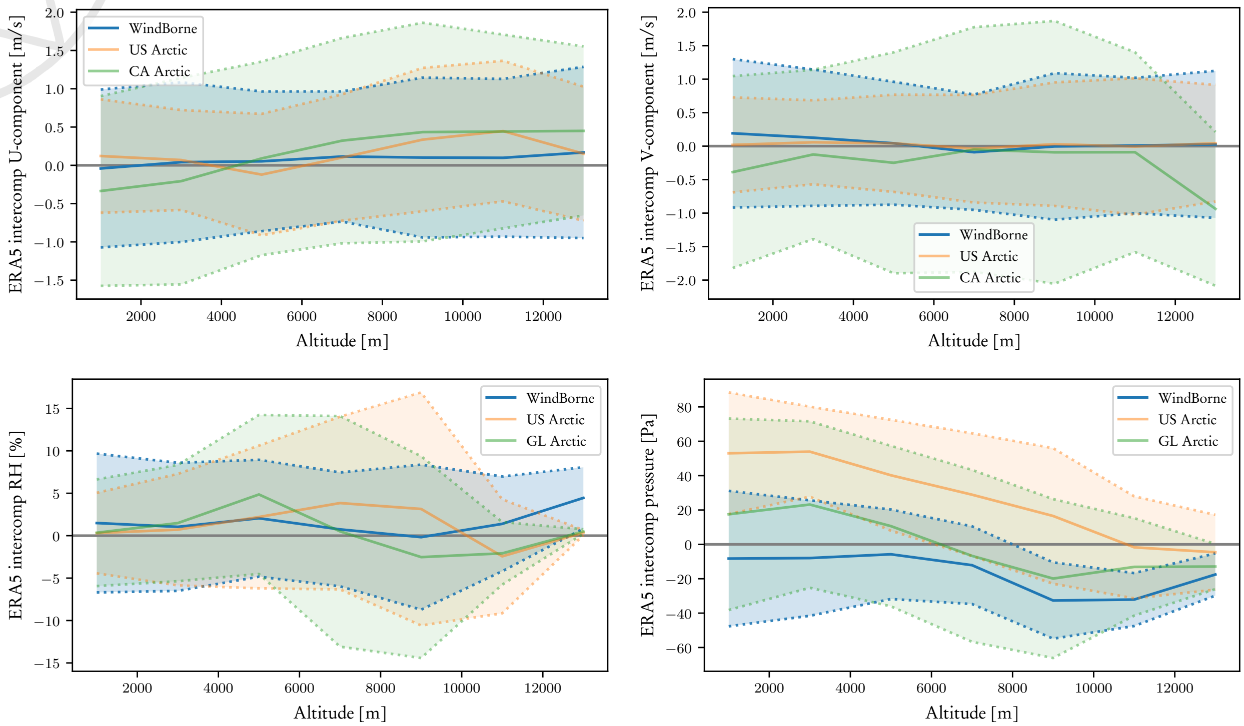

Figure 6: Intercomparison of soundings with ERA5 for winds, relative humidity, and pressure. The solid line represents the median bias; the shaded area represents an interval of $\pm 1\sigma_{\text{MAD}}$ (the median absolute deviation) for all the errors in each altitude bucket. Note that, while both WindBorne and routine radiosonde launches in the Arctic circle are all compared against ERA5, the routine launches were assimilated into ERA5, which significantly reduces their uncertainty.

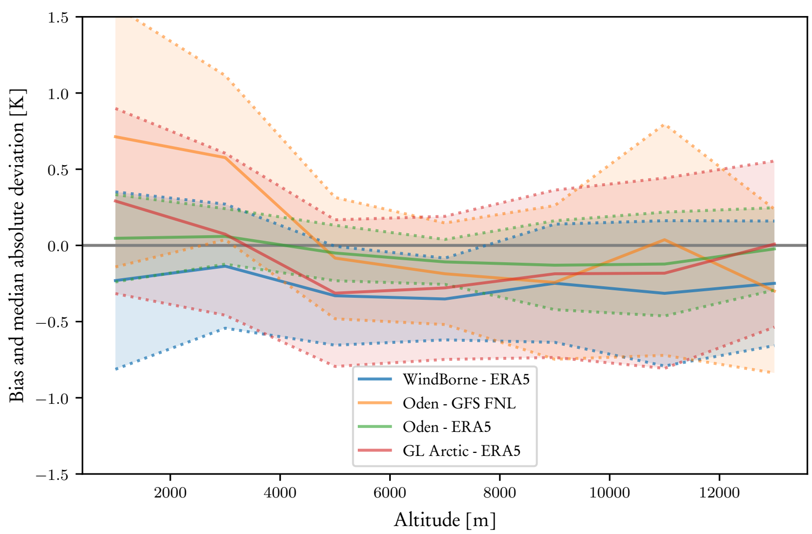

Figure 7: Same plot but with temperature, and including data from Oden. Note that FNL did *not* assimilate Oden data, and the errors are a lot bigger than ERA5's. Despite not being assimilated, WindBorne's bias and variances are comparable to FNL-Oden and Greenland routine temperature soundings.

0.05 K. A number of potential physical sources of error were investigated (for instance, potential longwave emission effects, for which the additional sensors proved useful) but were found to be negligible. The bias also does not have a strong ascent rate dependence, as one would expect for a physical bias (such as the solar bias during daytime, which is already corrected for based on historical data). It's further surprising as the direct intercomparison with Barrow radiosondes in the upcoming section did not seem to show a bias. Limiting the intercomparison to points where the ERA5 lapse rate was large (to avoid comparing inversions, which models tend to underestimate) improved it slightly, but did not completely eliminate it; limiting to the valid times or levels did not have much of an effect either.

The exported data does not modify temperature values based on that bias; while it might be worth experimenting with an offset at some point, it's not even clear it's a real bias. Indeed, Fig. 7 shows various examples of intercomparisons against ERA5 and FNL, using the unique dataset provided by Oden. Oden's radiosonde launches were not assimilated by GFS (or FNL after that). The figure shows that as a result ERA5's comparison has a much diminished variance and bias compared to FNL's comparison (because it assimilated the observations into the reanalysis). The bias and variance displayed by WindBorne's data overall are roughly comparable (if sometimes smaller) than those displayed by Oden's data when compared to FNL, suggesting that model error might be a reasonable explanation for this effect. Yet another comparison is Arctic routine radiosonde launches from Greenland to ERA5 which, despite being assimilated, still have a nontrivial leftover bias and variance.

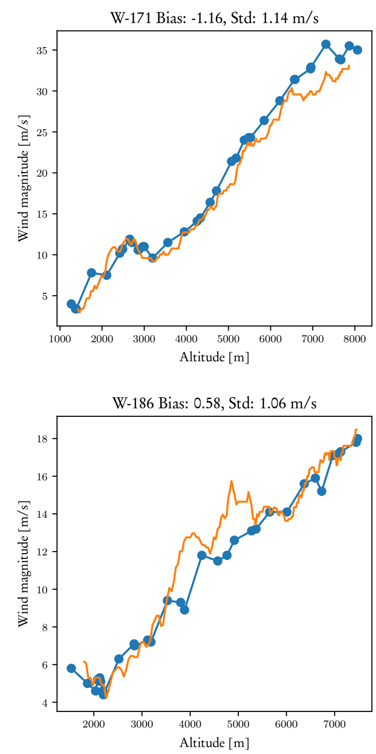

Figure 8: Actual differences between wind magnitude profiles for two launches from Barrow. Besides slightly higher frequency (we will downsample by averaging later) it looks like the differences are simply local changes in the winds (the measurements between WindBorne and NWS were taken 1.6 and 1.5 hours apart, respectively).

3.2 Comparison with radiosondes

Another comparison opportunity is to use routine radiosonde launches in various locations around the Arctic circle twice daily (at 11Z and 23Z, usually). The limitation of such comparisons is that if there isn't very close agreement in space and time (less than a hundred meters and an hour or two) the variation in the upper atmosphere will be significant enough to make the comparisons meaningless. As a result the only comparisons shown here are against routine Barrow radiosonde launches for launches that were concidentally launched close in time to the radiosonde launches. Even filtering for location and time, slight differences in time, as well as much different drifts (due to different ascent rates) results in imperfect comparisons; Fig. 8, for instance, shows the actual profiles for wind magnitude for two launches, comparing the NWS sounding and WindBorne data. All measurements can be taken to have negligible error, and yet the difference between them is nontrivial.

Fig. 9 shows individual flight comparisons and the overall resulting bias between the two. While there's non-trivial error bars due to the limited number of soundings in the intercomparison, they overall look unbiased, and without an obvious potential $-0.25\,\mathrm{K}$ as discussed in the previous section.

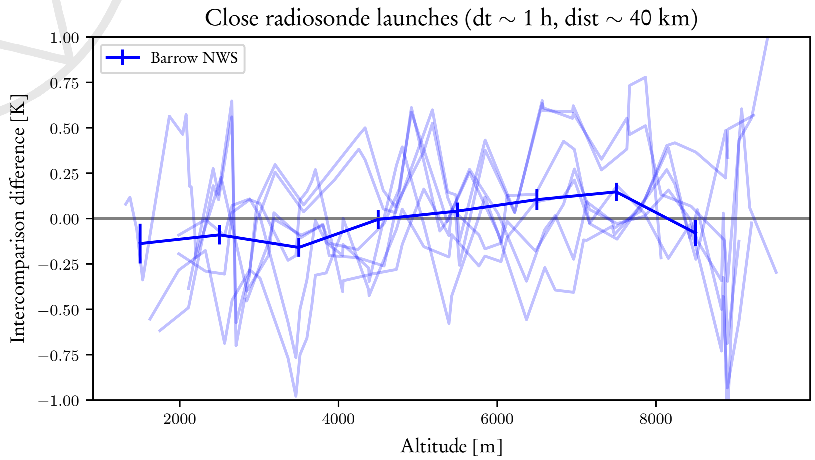

Figure 9: Intercomparison of temperatures measured by WindBorne and routine Barrow radiosonde launches. Note that this includes both the individual errors from each sounding system and a (probably dominant) component of spatio-temporal variation. Even then, they agree fairly closely.

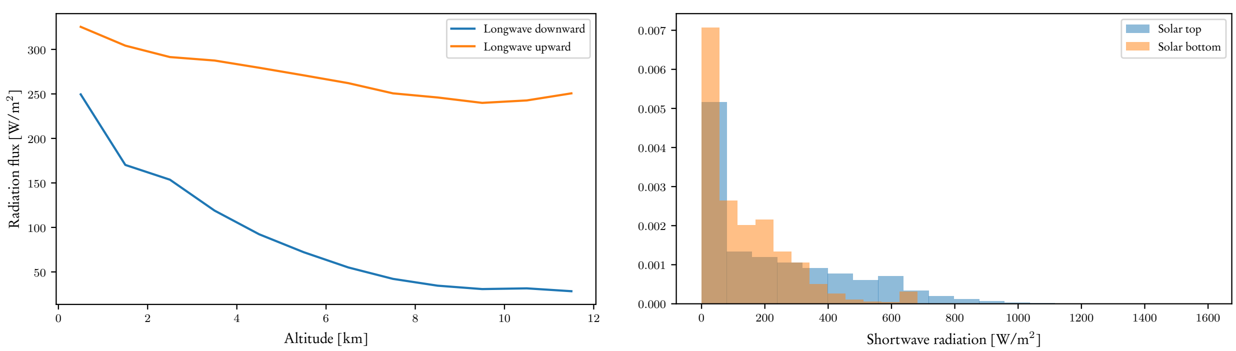

Figure 10: Longwave radiation fluxes measured as a function of altitude, and a histogram of the measured shortwave radiation fluxes.

3.3 Shortwave and longwave radiation data

Finally, we also debuted a radiation sensor that provides the upward and downward fluxes in both shortwave (400–1000 nm) and longwave (5.5–14 µm). This was not originally planned and represented a last-minute addition, so the yield was not ideal due to manufacturing limitations and limited integration testing (of the 90 flights, 49 had longwave data, and 55 had shortwave data). They still provided a good amount of data that was useful for looking into potential biases in the temperature sensor. They could also prove useful for looking into ice in the Arctic or modelling radiation in weather models, since as far as we know there's limited data on radiation fluxes as a function of altitude (usually only provided at the top of atmosphere and modelled at the ground as well).

4 Data availability and format

The data generated by these balloons is somewhat unique; the goal of this section is to explain where to find the data, what to make of it, and, perhaps most importantly, stress the fact if anything is unclear or you would prefer a different format you should contact Joan at `joan@windbornesystems.com`, and I'll be more than happy to help; this represents my best guess at what might be useful, but I'm happy to answer any and all questions about both sensor data or manipulating the raw data into a more useful format. If you end up downloading the data I'd appreciate if you could send me a quick email so I can track who to update whenever a new version of the data is released if applicable.

The URL is: `https://a.windbornesystems.com/syzygy/mst2/`

By default, the data is sampled at very high resolution (10 seconds, where the average ascent rate is a bit under 1 m/s). This often causes issues in data assimilation systems, so the data uploaded at the URL above (both BUFR and NETCDF) has been superobbed by averaging observations within 100 m or 20 min of each other (equivalent to thinning by over $10\times$). For reference, the raw count of data rows is 3.1 million; after superobbing, there are approximately 178,000 rows of data (of which 3/4 include valid temperature).

The data uploaded also does not include shortwave and longwave fluxes, as these are not standard radiosonde-style observations. If anyone is interested in either the full resolution dataset or fluxes, let Joan know; a different format (such as NumPy) might be more convenient in that case.

4.1 BUFR considerations

This type of observations does not map directly to anything in the WMO tables, given that it represents radiosonde-like balloons that fly for many days and drift over long distances. Additionally, since we carry independent GPS sensors and a pressure sensor[1] (and the pressure sensor had issues on some flights, and doesn't work below 20 kPa) the preferred coordinate for assimilation is geopotential height, which we always measure (converted from GPS height assuming an oblate sphere as per WMO-8). As a result, the BUFR files are neither `TEMP` nor `ACARS` data; I selected individual codes. The most notable difference is that I picked `007009` (geopotential height) and `010004` (non-coordinate pressure) for the vertical levels.

[1] Note that this differs from most current radiosonde designs, which integrate out pressure based on GPS height and the vertical profile. Since we fly for many days at a time and reach remote areas, we carry an onboard barometer calibrated to 5 Pa or so.

Another important fact is that I was unable to find a way to mark missing values using the ECMWF ecCodes library to encode BUFR's (some flights only have missing segments of temperature of pressure, for instance). For those, I used the largest possible value in the encoding, which is obviously invalid (163830 Pa, 655.35 K, and 309.5%RH).

4.2 General considerations

A somewhat-subjective characterization of the data is that, when it comes to measurement error, geopotential altitude and wind components can be taken to have negligible error (besides, of course, representativeness errors). After that, pressure has small error when available (for its operating range of 20–100 kPa, errors should be below 20 Pa). Temperature should be pretty solid as well and not far from established radiosondes, with perhaps slightly higher variances during daytime (as the solar correction is not perfect—I can provide estimated standard deviation if that would be useful). However, looking at bulk statistics, there *might* be a bias of -0.25 K against ERA5 (i.e. adding 0.25 K to the data makes it appear unbiased) as discussed in Sec. 3. While I have lost my sanity chasing potential sources of errors as small as 0.01 K, I couldn't find a root cause for such a bias, so it could potentially simply be the sparsity of observations in the reanalysis. Finally, relative humidity is fairly unknown in terms of performance, as it's hard to compare against reanalysis—while we calibrate it to about 1.4%, errors in flight could end up bigger than that (based on Fig. 6, roughly comparable to the errors of existing routine radiosondes). It is expressed with respect to a liquid surface.

In other words, some possibly interesting experiments, if resources and time allow (obviously feel free to override with better judgement, as I'm far from an expert in this field), are:

- The entire dataset excluding relative humidity (I believe ECMWF default).
- The entire dataset excluding relative humidity, and adding 0.25 K to air temperature. This bias may or may not be real, but it might be worth investigating in this way.
- Including relative humidity to the experiments.
- The entire dataset excluding both temperature and relative humidity (in other words, assimilating only geopotential height, pressure, and winds). This one can be taken to have negligible measurement error.

One final consideration is that, even though each individual flight is superobbed to have lower resolution, sometimes balloons were launched in clusters, leading to groups of 5 balloons that collect data very close to one another until they spread out. As a result, if the data assimilation system can have trouble with observations close to another, a higher level of superobbing might be desirable.

5 Looking forward

Regarding this trial we just completed, the steps going forward are going through the data, both for applications in tropopause polar vortices, and data assimilation. We're looking forward to what results come out, or any papers that end up being written; as discussed in Sec. 4, we are more than happy to help get the data in the way that's most useful.

Going forward to next year, we're hoping to introduce a fair number of improvements to the units. These include doubling the endurance of each balloon, which, combined with the already planned increase in number of units (from 90 this past summer, to something like 170 next year), should result in a lot more data. We also hope to have made improvements to navigation by then to collect and target more effectively, doing proactive rather than reactive navigation. In terms of hardware, if the data looks useful we will be operationalizing the shortwave and longwave radiation sensors (which seem like they could be useful given the ice environment, and modelling clouds), and some more experimental potential developments if time permits (such as dropsondes, or being able to take measurements over ice or the ocean for weeks after landing).