Less is More: Course Redesign and the Development of an Atmospheric Science Process Skills Assessment

Dawn M. Kopacz[1] and Zachary J. Handlos[2]

1. University of Nebraska - Lincoln
2. Georgia Institute of Technology

Received: 23 June 2020; Accepted: 9 September 2021

General education science courses strive to promote scientific literacy and the development of scientific process skills. However, research shows that many general education courses are still designed to stress content mastery. In this study, the number of topics in five semester-long introductory atmospheric science courses was reduced to increase time for the development of scientific process skills, a critical component of scientific literacy. The Atmospheric and Climate Science Literacy Frameworks and a general science education skills rubric were used to guide the course redesign and development of course activities. Details of the course structure and sample course activities are described. A pre-post-test was developed to evaluate attainment of five scientific process skills and the efficacy of the course redesign. Preliminary validity and reliability studies suggest that the majority of the assessment questions are reliable, though further validation of the assessment is required.

INTRODUCTION

The majority of four-year U.S. colleges and universities have general education programs, and these often include a science requirement. Atmospheric science programs generally offer introductory courses that also serve as general education science courses. These are usually introductory-level courses that do not require students to have extensive mathematics and science backgrounds (Ulanski, 1993). Ideally, general education science courses should aim to teach students that science uses observations and experiments to study the structure and behavior of the physical and natural world rather than teaching them to memorize a collection of discipline-specific content knowledge (Aloi, Gardner, & Lusher, 2003; Cakir, 2008). In other words, courses should be designed to promote scientific literacy, which has been recognized as fundamental to an undergraduate science education (de Caprariis, 1997; Gormally, Brickman, & Lutz, 2012; Hazen & Trefil, 1991; Nuhfer et al., 2016; Surpless, Bushey, & Halx, 2014).

There are varying definitions of what constitutes scientific literacy. Most definitions include developing an ability to apply scientific knowledge to real world scenarios. Along with this, we should expect a scientifically literate citizen to have the ability to make informed decisions as members of society, particularly regarding the social and economic issues that will affect the quality of their lives and those of their children (American Association for the Advancement of Science [AAAS], 1993, 2010; National Research Council [NRC], 1996; United States Global Change Research Program [USGCRP], 2009; University Corporation for Atmospheric Research [UCAR], 2007; Zen, 1990).

To ensure that general education goals are met, many colleges and universities have developed rubrics or documents that describe the skills that students are expected to possess when they have finished the course, typically ranging from "deficient" to "exemplary." Despite the growing use of skills-based rubrics to assess general education programs, it has been shown that many general education courses are still designed for majors and stress content mastery rather than higher order thinking (de Caprariis, 1997; Nuhfer et al., 2016). For example, a study in 2008 showed that, while the teaching methods for introductory atmospheric science courses at one large U.S. university were sufficient for learning course content, they were insufficient for the course goal of application learning (Kahl, 2008). Furthermore, a survey of instructors at 122 U.S. colleges and universities found that while many introductory atmospheric science courses include experiential or laboratory activities, the majority do not include collaboration or group work as a frequent activity in their classes, which has been shown to promote the development of critical thinking and analytical skills (Blosser, 1993; Ulanski, 1993). These course structures are in direct conflict with teaching and learning theories; in fact, a substantial literature base demonstrates that student learning and long-term retention is enhanced when students are engaged socially as well as cognitively (McGuire, 2006; Moog & Spencer, 2008; Yilmaz, 2011). In other words, interaction between teachers and learners within the classroom is vital in helping students make sense of course material through fitting this newly acquired knowledge within their own experiences (Cakir, 2008; Stains et al., 2018).

Several studies have explored alternative course structures and activities within introductory atmospheric science classes. For example, student feedback indicates that the use of technology tools in the classroom increases student motivation for, and understanding of, course material (Charlevoix, Jackman, & Twine, 2006; Cutrim, Rudge, Kits, Mitchell, & Nogueira, 2006). It has been shown that the use of popular movies to teach weather and climate concepts to non-majors is effective at teaching atmospheric science knowledge and appreciation (Yow, 2014). Domack (1999) found that student engagement and comprehension of atmospheric science concepts can be achieved through semester-long observation-oriented projects that integrate course concepts with the real world.

While the above studies explored the impact of course design on student motivation, engagement, and knowledge acquisition, they do not directly address the *skills* required to apply newfound knowledge to real-life situations. Process Oriented Guided Inquiry Learning (POGIL) is a research-based teaching and learning strategy in which the instructor facilitates the development of process skills that allow students to apply what they learn

in new contexts. A POGIL activity is a guided inquiry exercise that students work on in groups, with the instructor serving as the facilitator. In this setting, the development of skills are embedded into the course activities, and skills are also developed as the instructor interacts with each group and the class as a whole (Moog & Spencer, 2008). Two recent studies have explored the impact of course structure on student skill acquisition in upper-level meteorology courses. One such study showed that using real-world simulations and a problem-based learning approach in an upper-level undergraduate/graduate-level course increased student engagement as well as students' ability to collect, analyze, and interpret data (Charlton-Perez, 2013). Davenport (2019) used "worked examples" within an atmospheric dynamics course to guide students through complex atmospheric science problems in a step-by-step manner while also asking students to actively engage with the example problem. Worked examples are paired with similar practice problems to enhance problem-solving ability.

Given that the above studies demonstrate that active learning environments motivate and engage students, and improve students' science process skills in upper-level atmospheric science courses, this study proposes a new approach to teaching *introductory* atmospheric science courses to encourage scientific skill development (NRC, 2012; Stains et al., 2018). Specifically, this study will describe a course redesign that is structured around replacing time typically used to lecture about course material with active learning exercises, which often requires that less material be covered in a given course (Roebber, 2005). The course was designed to facilitate the development of students' ability in applying a set of specific skills associated with the scientific method (e.g., students' ability to use scientific knowledge to pose questions, make predictions, interpret data, and evaluate conclusions). What follows is a description of the course redesign and development of the atmospheric science skills assessment. Preliminary statistics regarding the validity and reliability of the novel assessment tool are discussed. The last section of this study describes future research opportunities in the context of these results.

COURSE REDESIGN

Atmospheric science literacy requires the following: 1) knowledge of fundamental atmospheric and climate science concepts (UCAR, 2007), and 2) the ability to apply this understanding towards the evaluation of real-world science problems. A backward design approach was used to redesign five semester-long general education introductory atmospheric science courses at two large public Universities during the Fall 2018, Spring 2019, Fall 2019, and Spring 2020 semesters.

A key aspect of this course redesign was the reduction of the total number of topics covered to allow increased time for in-class active learning opportunities. The primary goal of all courses was to promote atmospheric science literacy, which is defined here as scientific skill development in concert with acquisition of content knowledge.

An NSF-funded project, ELIPSS (Enhancing Learning by Improving Process Skills in STEM), demonstrated the importance of aligning learning outcomes, course activities, and assessment tools to help the instructor understand the effect of course structure and design on student skill development. In line with this, the authors used three existing tools (i.e., two frameworks and one rubric) as a guide to determine course goals, develop the course structure and activities, and to create an assessment tool that could be used to assess atmospheric scientific literacy and science process skills (Reynders et al., 2020). The "Atmospheric and Science Literacy Framework" (https://scied.ucar.edu/atmospheric-science-literacy-framework) and "The Essential Principles of Climate Literacy" (https://www.climate.gov/teaching/essential-principles-climate-literacy/essential-principles-climate-literacy) frameworks aim to define the essential concepts of atmospheric and climate science that would enhance atmospheric and climate

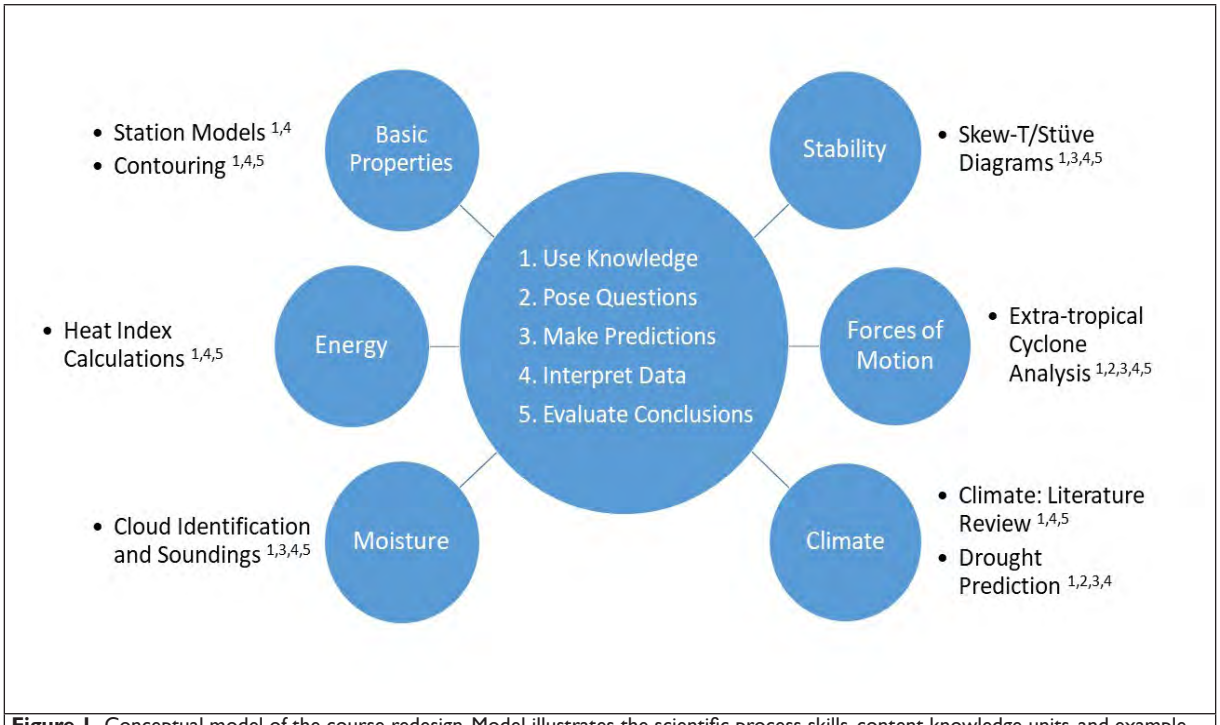


Figure 1. Conceptual model of the course redesign. Model illustrates the scientific process skills, content knowledge units, and example course activities utilized within the course redesign. Footnote numbers list scientific process skills emphasized during each activity, as assigned in the large center circle.

science literacy within the United States (UCAR, 2007). Each course was divided into six content units that were determined using these frameworks: Basic Properties of the Atmosphere, Energy, Moisture, Stability, Forces of Motion, and Climate (six smaller circles in Figure 1). Sample topics for each content unit are displayed in Table 1.

Table 1. Course topics by content area

Content Area	Topics Covered
Basic Properties	Weather variables, weather maps
Energy	Radar and satellite imagery, Earth's energy budget, seasons
Moisture	Air masses and fronts, clouds and precipitation, latent vs. sensible heat
Stability	Atmospheric stability, thunderstorms, tornadoes
Forces of motion	Atmospheric forces and winds, extratropical cyclones
Climate	Climate controls, climate change

The rubric used for the course redesign was the Achievement-Centered Education (ACE) 4 Rubric, which outlines five main goals for any University-wide general education science course (University of Nebraska - Lincoln [UNL], 2019; https://ace.unl.edu/ace-rubrics). These learning goals were adapted for the introductory atmospheric science courses investigated in this study (Table 2), and the instructors implemented various activities within each content unit to allow students to build the following set of skills: use scientific knowledge, pose questions, make predictions, interpret data, and evaluate whether conclusions are reasonable (large circle in Figure 1). These concepts and skills served as the guiding framework for redesigning the courses as well as designing an atmospheric science skills assessment that could be used to measure student knowledge and scientific skill gains.

Table 2. Description of the five skill categories

Skill Category	Description
Uses scientific knowledge	Demonstrates weather and climate knowledge
Poses questions	Poses questions about weather and climate that can be investigated.
Makes predictions	Make a weather or climate prediction based on existing knowledge.
Interprets data	Ability to analyze weather or climate data that is presented in a table, chart or map.
Evaluates whether conclusions are reasonable	Ability to determine if weather/climate prediction or solution to weather/climate problem is reasonable.

Note. Skill categories were adapted from UNL's ACE 4 rubric (UNL, 2019; https://ace.unl.edu/ace-rubrics).

COURSE ACTIVITIES

Four out of the five introductory atmospheric science courses analyzed were formally listed as general education science courses. Three of the five courses were required by meteorology undergraduate students but open to all undergraduate students to enroll. The courses were redesigned to allow students adequate time to strengthen their science process skills along with increasing their knowledge of atmospheric science concepts. Specifically, each class session included an activity that was designed to assess one or more of the scientific process skills (Figure 1). Activities were started (or assigned) in class and either completed during the same class session, finished as homework, or continued in the following class session. Figure 1 lists examples of activities implemented within each of the six course content units. Four of these activities are described in detail below.

Example Activity 1: Contouring Activity

With respect to the "Basic Properties" module, students at both Universities investigated in this study were asked to perform a weather map contouring activity as part of a take-home assignment. After discussing the basics of weather map contouring as a class, students contoured a surface weather map and answered a series of questions that required them to apply a few scientific skills (see Appendix A: *Example Activity 1: Contouring Activity*). This included a combination of assessing mastery of course content (i.e., "Use Knowledge"), meteorological feature identification (i.e., "Interpret Data") and error assessment of their contouring ability versus other classmates and a computer map analysis (i.e., "Evaluate Conclusions"). While weather map contouring exercises are common across introductory undergraduate atmospheric science courses, it is the assessment of the latter two scientific skills (i.e., "Interpret Data" and "Evaluate Conclusions") that shifts the focus within this redesigned course to an emphasis on both science process skills and content knowledge.

Example Activity 2: Cloud Identification Sounding Activity

In this activity (see Appendix B: *Example Activity 2: Cloud Identification Sounding Activity*), students learned about the basics of cloud identification and interpretation of sounding data on skew-T diagrams in class. Next, students were asked to take at least one picture of clouds locally and then compare and contrast their observation with the most recent sounding data available (using the sounding data collected closest to their observation location). From there, students were asked to identify cloud layers that they observed within the sounding data (i.e., "Use Knowledge" and "Interpret Data") and to discuss whether the identified cloud layers were consistent with the cloud layers they observed in their picture (i.e., "Evaluate Conclusions"). This activity requires students to utilize science process skills within the context of connecting cloud formation theory and sounding data interpretation to their own direct observations of clouds.

Example Activity 3: Extra-tropical Cyclone Analysis Activity

During the "Forces of Motion" module, a case study of an extra-tropical cyclone was used to assess students' ability to apply science process skills towards interpretation of the life-cycle of a significant weather event (see Appendix C: Example Activity 3: Extra-tropical Cyclone Analysis Activity). Students were provided with a series of surface and upper level weather maps and asked to interpret the state of the atmosphere (i.e., surface pressure distribution, magnitude of the pressure gradient force, wind direction and speed, etc.; "Use Knowledge", Interpret Data"). After analyzing the current surface conditions, students were presented with upper level weather maps and asked to make a prediction about the evolution of the surface low with respect to time (i.e., "Makes Predictions"). Finally, students were provided with the surface weather map for a later time and were asked to evaluate their forecast relative to the observations (i.e., "Evaluate Conclusions").

Example Activity 4: Semester Project

Two of the courses required students to develop a final project in place of a final exam. Since the project was comprehensive, it was not assigned to a particular content module exhibited in Figure 1 (see Appendix D: *Example Activity 4: Semester Project*). Students began work on this project approximately midway through the semester. The project was designed to assess student understand-

ing of the material and development of the set of skills outlined in Figure 1 (large center circle). Students worked in groups of five to seven to analyze a significant weather event, prepare a technical summary describing the physical mechanisms responsible for that event, and to present their findings to the class. Students were given several work days during class to develop their research questions (i.e., "Poses Questions") and to begin to gather and analyze data (i.e., "Interpret Data" and "Evaluate Whether Conclusions are Reasonable").

Again, the activities described above were designed to help students develop the set of science process skills outlined in Figure 1. Implementation of the activities required class time which was factored into the course redesign. To incentivize participation in these activities, a substantial portion of the course grade (i.e., 10-20%) was determined by their performance on and/or participation in these activities.

DEVELOPMENT OF AN ATMOSPHERIC SCIENCE PROCESS SKILLS ASSESSMENT

Concept Inventories (CI), which typically consist of a multiple choice test used to measure student understanding of course concepts and identify misconceptions, have been used to assess students in general education courses. The physics community was the first to develop a CI, and several other disciplines developed their own within the decades that followed (Halloun & Hestenes, 1985; Hestenes, Wells, & Swackhamer, 1992).

A CI for the atmospheric science community was developed over the past few years and has recently been made available to the community (Davenport, Wohlwend & Koheler, 2015; Davenport & French, 2020). While important, assessing student content understanding is distinct from determining their level of ability with respect to content applications, problem-solving, and, in the sciences, understanding the scientific process, the latter of which this project aims to assess.

The broader academic community has developed a number of different instruments to assess scientific literacy (Benjamin et al., 2017; Hanson, 2016). These have ranged from a general critical thinking and reasoning skills assessment to a recent study that utilized a concept inventory to assess the reasoning component of scientific literacy. The latter of the two aimed to develop a multidisciplinary assessment tool by incorporating ideas from biology, chemistry, environmental science, geoscience, and physics (Nuhfer et al., 2016). Recognizing the importance of testing for scientific literacy, but finding the existing tools inadequate for their assessment needs, some have developed their own discipline-specific instruments (Gormally et al., 2012; Davenport, Wohlwend & Koheler, 2015; Davenport & French, 2020).

Despite the emergence of CI's within STEM disciplines, a concept inventory or non-discipline-specific assessment is insufficient for assessing the development of *scientific process skills* within the context of atmospheric science. Given the lack of a pre-existing scientific skills assessment (as well as the need for such a discipline-specific tool), the authors developed such an assessment and pilot-tested this within the redesigned courses discussed above. The authors used classical test theory as a guide to develop the assessment, using course objectives as a guide for the construction of the test items (Engelhardt, 2009). Following Engelhardt's recommendations, multiple questions for each skill category were developed with the assumption that not all test questions would be strong discriminators. The assessment was designed with five questions for each skill category (large circle in Figure 1) and one for each of the six content units (smaller circles in Figure 1), for a total of 30 questions.[1] Each question was formatted according to the guidelines for writing multiple choice test items set forth by Haladyna et al. (2002). For example, questions were designed to assess a particular skill rather than simply testing for recall of facts, and the central idea and directions were clearly written and contained within the stem. The distractors (choices) were written in a homogeneous format, were equal in length, and whenever possible, typical student errors were used as distractors.

Validity of Atmospheric Science Literacy and Skills Assessment

It is important to ensure that the assessment will measure what it is intended to, and that conclusions regarding the effectiveness of the course redesign or other classroom interventions can be drawn from student scores. This is known as validity and it is measured in several different ways, including content, construct and criterion validity. Content validity requires that an independent panel of experts review the individual items to ensure they match the objectives and to examine the accuracy, formatting, and grammar of the test items. Evidence of construct validity should also be assessed to ensure that the assessment measures skill acquisition. Examples of construct validity include, but are not limited to, administering the assessment before and after instruction to show that student performance increases after instruction (intervention study), and administering the assessment to different student groups (i.e., atmospheric science majors versus non-majors), as we would expect the atmospheric science majors to perform differently than non-majors (differential population study). A statistical analysis can also be performed to identify the underlying structure of the test items that account for observed variations in student performance (factor analysis; Engelhardt, 2009). Evidence of criterion validity is determined by comparing student assessment scores with another measure of proficiency, such as coursework.

With the test items created for the atmospheric science process skills assessment, eight University-level atmospheric science educators were contacted to anonymously review the test questions to ensure that they matched the objectives (i.e., that the content and skill categories assigned to each question were appropriate). The grammar, formatting, accuracy and clarity of each question was also reviewed (Engelhardt, 2009). Only one of the individuals completed the content validity review. This review, combined with the authors' independent content validity reviews of the test items, resulted in a total of three evaluations. The content validity review resulted in minor revisions to the formatting and clarity of some questions.

The content validity check for the Atmospheric Science Literacy and Skills Assessment revealed 100% agreement on the skill categories for 18 of the 30 questions. These 18 questions were retained and the remaining questions were discarded. The revised assessment consisted of five "Uses scientific knowledge", two "Poses questions", three "Makes predictions", six "Interpret data", and two "Evaluates conclusions" questions. At this point the assessment was field tested in order to conduct a reliability check of the test items.

After the content validity check is performed, the objectives or individual test items might be re-evaluated, or a reliability check may be performed. The reliability check may reveal the need to revise existing test items or create new ones, which would require another content validity check. Thus, the development of an assessment and its evaluation is often a cyclic process. Once a reliable version of the test is created, an evaluation of construct and criterion validity is performed (Engelhardt, 2009).

Reliability of the Atmospheric Science Process Skills Assessment

The pre-post-test assessment was field-tested during the Fall 2018, Spring 2019, Fall 2019, and Spring 2020 semesters to assess its reliability. It was administered to student participants at the beginning (i.e., pre-test) and end of the course (i.e., post-test).[2] Given that participants completed an 18-question assessment, this sample size falls within the range of 90-180 participants recommended by Engelhardt (2009) regarding an appropriate sample size (i.e., sample size of 5-10 times that of the number of test items). Approximately 35.3% of the participants identified as male and 61.7% were female. The majority of the students (80.9%) were between 18 and 20 years old and were first or second-year students (86.7%). Approximately 54% of the students had taken a prior science course, and 20.6% were taking another science course at the time of the study.

The reliability of the assessment examines whether the test consistently produces similar results. The three categories of reliability are the following: stability, equivalency, and internal consistency. Stability examines how consistent the student test scores are over time and is typically analyzed using a test-retest method. Equivalency examines student performance on two alternate versions of the assessment; it is expected that student performance on two tests assessing the same concepts will show little difference. Internal consistency examines the similarity of the test items using statistical measures between different items after a single test session.

The stability of a test can be examined using a modified test-retest method which is used when it cannot be assumed that the ability of the test-taker has remained unchanged. This involves administering the assessment to two similar populations (i.e., two sections of a particular course; Engelhardt, 2009; Arthurs et al., 2015; Davenport & French, 2020). In this study, a modified test-retest method was used to examine the stability of the assessment. The pre-test data from the two sections of the same course (Fall 2018 and Fall 2019 semesters) were used to compute correlation coefficient, r, which is computed as follows:

$$r = \frac{\sum_{i=1}^{N} x_i \, y_i}{N \, \sigma_x \, \sigma_y}$$

where x_i and y_i represent a student's deviation score for each course section, respectively, represents the number of students within each section (N = 44 for both sections), and σ_x and σ_y represent the standard deviation of the scores for each course section, respectively. The correlation coefficient, r, for the atmospheric science process skills assessment was 0.80, which indicates that the differences in student scores on the assessment are more likely to be associated with a true measure of skill, with any variation in student scores likely due to random error (Table 3). These errors may result from a lack of concentration during test-taking, students responding randomly to the test items, or it may be due to difficulty comprehending the test items. While these results are encouraging, additional analysis with a larger sample size is warranted to improve the robustness of the results, particularly since the intent is to use the assessment for pre-post analysis (Engelhardt, 2009; Arthurs et al., 2015; Ingram, 2018; Davenport & French, 2020).

Five important reliability statistics to consider when evaluating the internal consistency of the exam are the following: i) difficulty of the test items, ii) the discrimination index (i.e., measures how well the test items distinguish between students that know the correct answer and those that don't), iii) Ferguson's Delta (i.e., examines discriminatory power of the test as a whole), iv) the discrimination index and point-biserial correlation (which can be used to determine the discriminatory power of each assessment item), and v) the Kuder-Richardson 20 (KR-20) value (i.e., examines internal consistency of the test as a whole; Engelhardt 2009). Equivalency was not measured for this assessment since only one version of the test exists.

The internal consistency of the individual items on the assessment was examined for all course sections using the difficulty index, the discrimination index, and the point biserial correlation. The difficulty of the individual test items is determined by dividing the number of students answering a test item correctly, by the total number of students taking the test. The difficulty index ranges from 0 to 1, with a 0 indicating no students answered correctly (i.e., a difficult question) and a value of 1 indicating that all students answered correctly (i.e., an easy question). The difficulty index for a test question is ideally equal to 0.5, but values from 0.3 to 0.9 are considered acceptable. The difficulty index for the pre-post-test questions ranged from 0.11 to 0.67, with an average difficulty of 0.47. Three of the 18 questions (Q4, Q5, and Q15; see Table 3 and Figure 2) had a difficulty index below 0.3 indicating they were difficult questions and likely need revision.

Table 3. Reliability and stability statistics for the atmospheric science process skills assessment.

Question #	Difficulty Index	Discrimination Index	Point Biserial Correlation
Q1	***0.56***	***0.35***	***0.29***
Q2	***0.60***	***0.46***	***0.39***
Q3	***0.57***	***0.62***	***0.47***
Q4	0.11	0.04	0.02
Q5	0.23	0.26	**0.27**
Q6	***0.41***	***0.40***	***0.28***
Q7	***0.47***	***0.62***	***0.49***
Q8	***0.57***	***0.41***	***0.27***
Q9	***0.53***	***0.60***	***0.42***
Q10	***0.43***	***0.52***	***0.31***
Q11	***0.61***	***0.77***	***0.58***
Q12	***0.38***	***0.68***	***0.53***
Q13	***0.47***	***0.66***	***0.53***
Q14	***0.61***	***0.73***	***0.50***
Q15	0.26	0.29	**0.24**
Q16	***0.66***	***0.46***	***0.38***
Q17	**0.67**	0.29	**0.27**
Q18	***0.36***	***0.69***	***0.53***
Averages	0.47	0.49	0.38
	r	Ferguson's Delta	KR-20
	0.80	0.97	0.76

Note: Values in **bold** meet the acceptable threshold for each statistic.
Italics are used to denote questions with acceptable values for all statistics

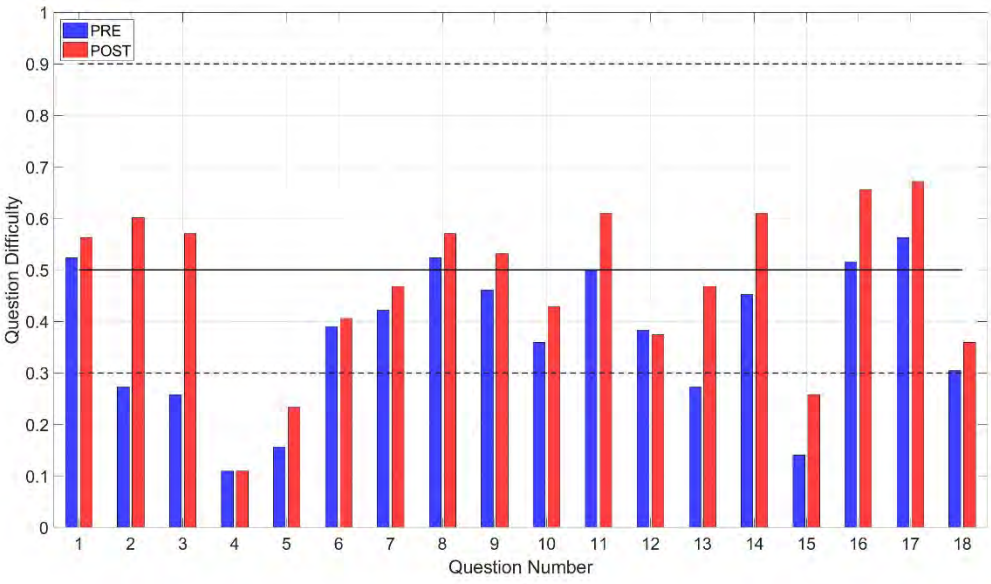


Figure 2. Item difficulty on the pre-test (blue) and post-test (red) for the atmospheric science process skills assessment. The solid black line denotes the idealized difficulty level, while the dashed black line indicates the minimum acceptable difficulty index. Figure format based on Figure 5 from Davenport & French (2020).

The discrimination index is calculated to determine how well each test question distinguishes between high-scoring and low-scoring students (e.g., students scoring in the top and bottom 27%, respectively). The discrimination index can be calculated as follows:

$$D = U - L$$

where U is the proportion of students with scores in the upper 27% that answered the question correctly, and L is the proportion of students with scores in the lower 27% that answered the question correctly. The discrimination index varies between -1 and 1. When more students in the top 27% answer a question correctly, the index is positive, and a negative index means that more students in the lower 27% answered a question correctly. Typically, a discrimination index above 0.30 is considered acceptable, with an index greater than or equal to 0.4 considered excellent. It should be noted that the difficulty of each item must also be considered when interpreting the discrimination index (Engelhardt, 2009). The majority (13 out of 18) of the assessment questions were found to be excellent discriminators, one question was considered good, and three were found to be acceptable. One question had a poor discrimination index indicating the need for revision (Q4; Figure 3).

The point biserial correlation is another statistical measure that correlates an assessment item with the overall assessment score. This correlation index falls between -1 and 1, and a value close to +1 indicates that students with a high score on the assessment are more likely to answer a test item correctly. It is important for the test items to be strongly correlated with the overall assessment score, thus the point biserial correlation should be greater than 0.2 (Engelhardt, 2009). The point biserial correlation is determined by:

$$r_{pbs} = \left(\frac{\overline{X}_{correct} - \overline{X}_{whole\ test}}{\sigma_{whole\ test}}\right)\left(\sqrt{\frac{p_i}{1-p_i}}\right)$$

where $\overline{X}_{correct}$ is the average assessment score for the students that answered the assessment item correctly, $\overline{X}_{whole\ test}$ is the average assessment score for all participants, $\sigma_{whole\ test}$ is the standard deviation of the assessment score for all participants, and p_i is the difficulty index for each item.

The point biserial correlation for one of the questions is near zero indicating that this question should be revised or removed from the assessment (Q4; Figure 4). However, the point biserial correlation for the remaining questions exceeds 0.2, indicating that nearly all of the test items are positively correlated with performance on the test as a whole.

The discriminatory power of the entire assessment and its internal consistency were measured using Ferguson's Delta and the Kuder-Richardson 20 statistic, respectively. Ferguson's Delta is a statistical measure that describes how well the test as a whole discriminates between students. Ferguson's delta is given by:

$$\delta = \frac{N^2 - \sum f_i^2}{N^2 - \left(\frac{N^2}{K+1}\right)}$$

where N represents how many students are in the sample, K is the number of test items, and f_i is the count for each score. The value varies between 0 and 1, with a value greater than 0.9 considered acceptable. The Ferguson's Delta value for the pre-post-test was 0.97 indicating that the pre-post-test as a whole does a good job of differentiating between students (Table 3).

The Kuder-Richardson 20 (KR-20) statistic looks at the covariance of the test questions and can be used as a measure of internal consistency of the entire multiple choice test (used for tests with dichotomously scored items). The KR-20 value is a correlation that is computed as follows:

$$KR-20 = \left(\frac{k}{k-1}\right)\left(1 - \frac{\sum_{i=1}^{k} p_i(1-p_i)}{\sigma_i^2}\right)$$

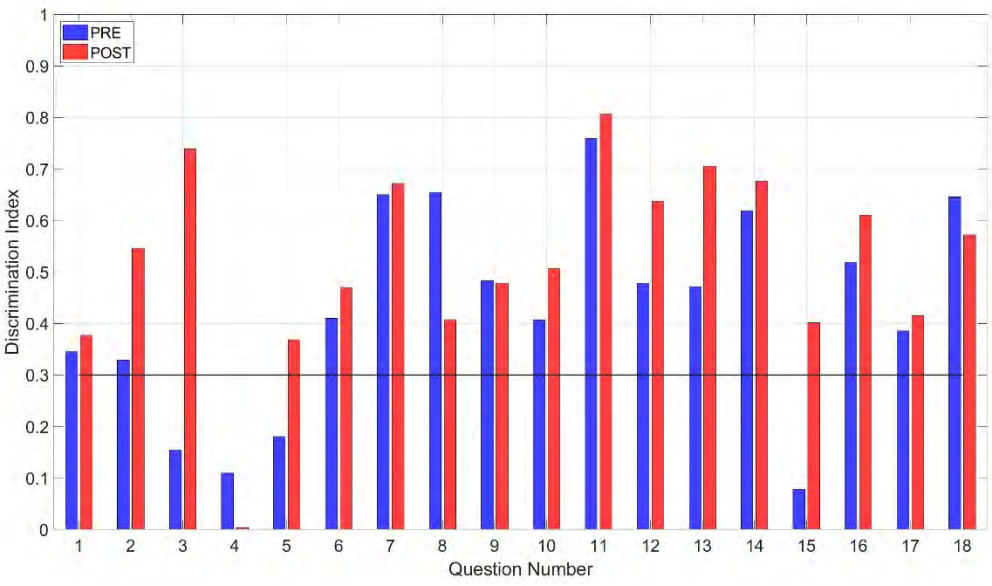


Figure 3. Discrimination index values on the pre-test (blue) and post-test (red) for the atmospheric science process skills assessment. The solid black line denotes the idealized item discrimination. Figure format based on Figure 6 from Davenport & French (2020).

where the number of assessment items is represented by k, the variance for the assessment as a whole is denoted by σ_i^2, and p_i is the difficulty index for each item. A KR-20 value above 0.7 indicates a test is statistically consistent. The post-assessment KR-20 value was 0.76 demonstrating internal consistency for the atmospheric science process skills assessment (Table 3).

In summary, the atmospheric science literacy and skill assessment exhibited validity and reliability when pilot-tested within the context of the redesigned courses described within this study. This assessment demonstrates promise with respect to quantifying student science process skills within the context of an introductory atmospheric science course. The next section will discuss suggestions for improvement to the assessment (e.g., addressing four questions that did not meet one or more of the minimum reliability threshold values), and plans for additional validity and reliability studies.

SUMMARY AND FUTURE WORK

Five semester-long introductory atmospheric science courses were redesigned to reduce the number of course topics to allow more time for students to develop scientific process skills in addition to building their content knowledge. An assessment tool was developed using the Atmospheric and Climate Science Literacy Frameworks and a general science education skills rubric. This

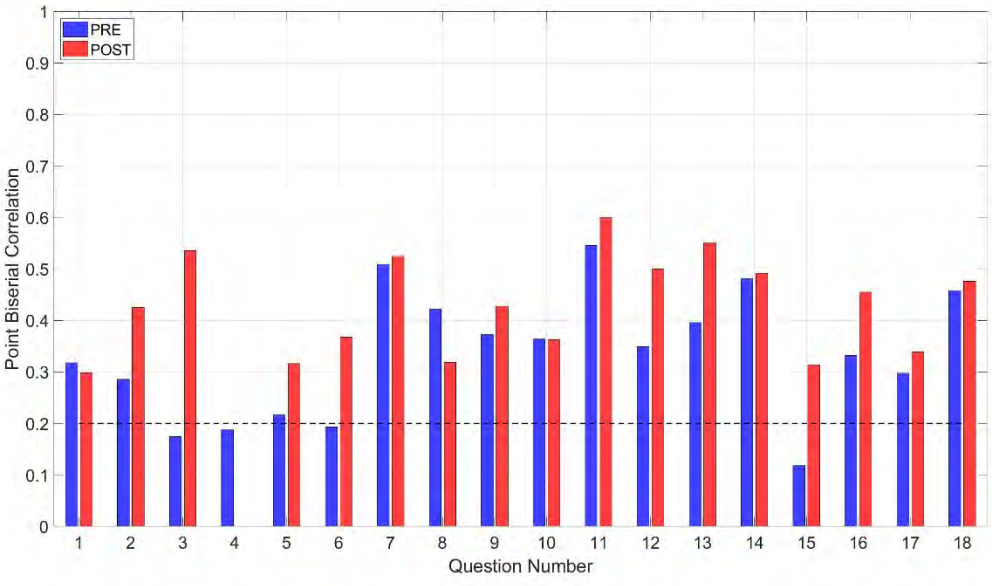


Figure 4. Point biserial correlation values for the pre-test (blue) and post-test (red) for the atmospheric science process skills assessment. The dashed line represents the minimum threshold desired for point biserial correlation. Figure format based on Figure 7 from Davenport & French 2020.

course redesign allowed the opportunity to ingest active learning exercises that ranged from assignments requiring students to connect course content to atmospheric science data and "real-world" experiences to a term project investigating a significant weather event of interest.

Special care was taken in the course redesign to align learning outcomes and course activities. No suitable assessment tool existed to evaluate the efficacy of the course redesign at promoting the development of atmospheric science process skills, and thus the authors designed an assessment tool for this purpose. Preliminary statistics regarding the validity and reliability of the novel assessment tool were analyzed after field testing it across five sections of introductory atmospheric science courses, accounting for a total of 128 students. The item difficulty, discrimination index, and point biserial correlation measures revealed that the majority of the assessment items fell within the desired thresholds for reliability while also helping to identify areas for improvement. The Kuder-Richardson-20 statistic indicates that the assessment as a whole is internally consistent, and the modified test-retest correlation coefficient demonstrates that once fully validated the assessment will likely be able to produce consistently stable and reliable results.

While the content validity and reliability check show promising results, they also revealed that four of the test questions need revisions and reevaluation. These questions will be revised and feedback will be solicited from students and experts to ensure clarity and accuracy. Additional validity and reliability studies will need to be conducted on the revised assessment. In addition, in order to use the assessment to evaluate the efficacy of the course redesign at promoting the development of scientific process skills, an evaluation of construct and criterion validity must also be performed. Future work includes plans to expand the IRB protocol to collect coursework information in order to gather evidence of criterion validity through a comparison of student scores on the assessment with other measures of proficiency such as coursework and course evaluations. Additionally, intervention and differential population studies are needed. This would involve administering the assessment, once fully validated and reliable, to different groups (e.g., across institutions and demographic sub-groups such as atmospheric science majors and non-majors) before and after instruction to determine if student performance increases following lessons and activities designed to promote scientific skill development. Refining the assessment questions and conducting additional validation and reliability studies will allow for the assessment to be applied universally to assess student attainment of atmospheric science process skills in any atmospheric science course.

In addition to evaluating the validity and reliability of the assessment, additional studies may also include student reflection through interviews as well as a survey to learn more about the student perspective on course structure and activities. Results of a study conducted on a broader range of courses could provide guidance for improving introductory atmospheric science courses to ensure student success in scientific skill development. This work is important from a general education perspective, as development of these scientific process skills are crucial as they are widely applicable throughout a variety of careers and to life in general (i.e., scientific literacy). From the perspective of atmospheric science programs, improving our courses and strengthening skill development within introductory courses may help improve retention in our programs and ultimately lead to a more skilled workforce.

ACKNOWLEDGEMENTS

The authors wish to thank the anonymous atmospheric science community member that served as the third person vetting our pre-post assessment tool. We would also like to thank the anonymous reviewers whose comments and suggestions helped improve and clarify this manuscript.

NOTES

1. The test questions are not published here. Interested parties can contact the authors for the test questions.
2. The collection of pre-post-assessment data within the courses investigated within this study was approved by University of Nebraska-Lincoln IRB 0180918669 EX and Georgia Institute of Technology IRB H18311.

REFERENCES

Aloi, S. L., Gardner, W. S., & Lusher, A. L. (2003). A Framework for Assessing General Education Outcomes Within the Majors. *Journal of General Education*, *52*, 237-252.

American Association for the Advancement of Science. (1993). Benchmarks On-line. *Project 2061*. Retrieved from http://www.project2061.org/publications/bsl/online/index.php

American Association for the Advancement of Science. (2010). Vision & Change in Undergraduate Biology Education: A Call to Action. Retrieved from https://live-visionandchange.pantheonsite.io/wp-content/uploads/2011/03/Revised-Vision-and-Change-Final-Report.pdf

Arthurs, L., Hsia, J. F., & Schweinle, W. (2015). The oceanography concept inventory: A semicustomizable assessment for measuring student understanding of oceanography. *Journal of Geoscience Education, 63:*4, 310 – 322. https://doi.org/10.5408/14-061.1

Benjamin, T. E., Marks, B., Demetrikopoulos, M. K., Rose, J., Pollard, E, Thomas, A., & Muldrow, L. L. (2017). Development and Validation of Scientific Literacy Scale for College Preparedness in STEM with Freshmen from Diverse Institutions, *International Journal of Science and Math Education, 15,* 607-623, https://doi.org/10.1007/s10763-015-9710-x

Blosser, P. E. (1993). Using Cooperative Learning in Science Education. *The Science Outlook*, 1-11.

de Caprariis, P. P. (1997). Impediments to Providing Scientific Literacy to Students in Introductory Survey Courses, *Journal of Geoscience Education*, *45*, 207-210. https://doi.org/10.5408/1089-9995-45.3.207

Cakir, M. (2008). Constructivist Approaches to Learning in Science and Their Implications for Science Pedagogy: A Literature Review. *International Journal of Environmental & Science Education*, *3*, 193 - 206.

Charlton-Perez, A. J. (2013). Problem-Based Learning Approaches in Meteorology. *Journal of Geoscience Education*, *61*, 12-19. https://doi.org/10.5408/11-281.1

Davenport, C. E., Wohlwend, C. S., & Koheler, T. L. (2015). Motivation for and Development of a Standardized Introductory Meteorology Assessment Exam, *Bulletin of the American Meteorological Society, 96*, 3062-3086.

Davenport, C.E. (2019). Using Worked Examples to Improve Student Understanding of Atmospheric Dynamics. *Bulletin of the American Meteorological Society, 100*, 1653-1664. https://doi.org/10.1175/BAMS-D-18-0226.1

Davenport, C. E. & French, A. J. (2020). The Fundamentals in Meteorology Inventory: Validation of a Tool Assessing Basic Meteorological Conceptual Understanding. *Journal of Geoscience Education*. *68*:2, 152 – 167, https://doi.org/10.1080/10899995.2019.1629193

Domack, C.W. (1999). Teaching Introductory Meteorology Through an Observation-Oriented Term Project. *Journal of Geoscience Education*, *47*, 260-265. https://doi.org/10.5408/1089-9995-47.3.260

Engelhardt, P.V. (2009). An Introduction to Classical Test Theory as Applied to Conceptual Multiple-Choice Tests. *Getting Started in PER*, *2*, 1-40.

Gormally, C., Brickman, P., & Lutz, M. (2012). Developing a Test of Scientific Literacy Skills (TOSLS): Measuring Undergraduates Evaluation of Scientific Information and Arguments, *CBE: Life Sciences Education, 11*, 364 - 377.

Haladyna, T. M., Downing, S. M., & Rodriguez, M. C. (2002). A Review of Multiple-Choice Item-Writing Guidelines for Classroom Assessment. *Applied Measurement in Education, 15*, 309 - 334. https://doi.org/10.1207/S15324818AME1503_5

Halloun, I. A., & Hestenes, D. (1985). The Initial Knowledge State of College Physics Students, *American Journal of Physics, 53*, 1043-1055, https://doi.org/10.1119/1.14030.

Hanson, S. (2016). *The Assessment Of Scientific Reasoning Skills Of High School Science Students: A Standardized Assessment Instrument* (Master's Thesis, Illinois State University, Normal, Illinois). Retrieved from http://doi.org/10.30707/ETD2016.Hanson.S

Hazen, R. M. & Trefil, J. S., (1991). Achieving Geological Literacy, *Journal of Geological Education, 39*, 28-30, https://doi.org/10.5408/0022-1368-39.1.28.

Hestenes, D, Wells, M., & Swackhamer, G. (1992). Force Concept Inventory, *The Physics Teacher, 30,* 141-158, https://doi.org/10.1119/1.2343497.

Ingram, P. B., & Ternes, M. S. (2018). The SAGE Encyclopedia of Educational Research, Measurement, and Evaluation. "Test–Retest Reliability", 1689-1691. Retrieved from https://methods.sagepub.com/base/download/ReferenceEntry/the-sage-encyclopedia-of-educational-research-measurement-and-evaluation/i20932.xml

Kahl, J. D.W. (2017). Automatic, Multiple Assessment Options in Undergraduate Meteorology Education. *Assessment & Evaluation in Higher Education*, *42*, 1319-1325. https://doi.org/10.1080/02602938.2016.1249337

McGuire, S.Y. (2006). The Impact of Supplemental Instruction on Teaching Students *How* to Learn. *New Directions for Teaching and Learning*, *106*, 3-10. https://doi.org/10.1002/tl.228

Moog, R. S. & Spencer, J. N. (2008). Process Oriented Guided Inquiry Learning (POGIL). American Chemical Society Symposium Series, 994, 1 - 13. https://doi.org/10.1021/bk-2008-0994.ch001

National Research Council. (1996). National Science Education Standards. *National Academies Press*. Retrieved from https://www.nap.edu/read/4962/chapter/1

Nuhfer, E. B., Cogan, C. B., Kloock, C., Wood, G. G., Goodman, A., Delgado, N. Z., & Wheeler, C.W. (2016). Using a Concept Inventory to Assess the Reasoning Component of Citizen-Level Science Literacy: Results from a 17,000-Student Study. *Journal of Microbiology & Biology Education*, *17*, 143-155. http://dx.doi.org/10.1128/jmbe.v17i1.1036

Reynders, G., Lantz, J., Ruder, S. M., Stanford, C. L., & Cole, R. S. (2020). Rubrics to assess critical thinking and information processing in undergraduate STEM courses. *IJ STEM Ed*, 7. https://doi.org/10.1186/s40594-020-00208-5

Roebber, P. J. (2005). Bridging the Gap Between Theory and Applications: An Inquiry into Atmospheric Science Teaching. *Bulletin of the American Meteorological Society*, *86*, 507 – 517. https://doi.org/10.1175/BAMS-86-4-507

Stains, M., Harshman, J., Barker, M. K., Chasteen, S.V., Cole, R., DeChenne-Peters, S. E.,... Young, A. M. (2018). Anatomy of STEM Teaching in North American Universities. *Science*, *359*, 1468-1470. https://doi.org/10.1126/science.aap8892

Surpless, B., Bushey, M., & Halx, M. (2014). Developing Scientific Literacy in Introductory Laboratory Courses: A Model for Course Design and Assessment, *Journal of Geoscience Education, 62*, 244-263. https://doi.org/10.5408/13-073.1

University Corporation for Atmospheric Research. (2007). Atmospheric Science Literacy Framework. *UCAR Center for Science Education*. Retrieved from https://scied.ucar.edu/atmospheric-science-literacy-framework

Ulanski, S. L. (1993). An Analysis of the Liberal Arts in Introductory Meteorology Courses, *Bulletin of the American Meteorological Society*, *74,* 2203 – 2209.

University of Nebraska - Lincoln. (2019). ACE 4 Rubric. *ACE Rubrics*. Retrieved from https://ace.unl.edu/ace-rubrics

U.S. Global Change Research Program. (2009). Climate Literacy: The Essential Principles of Climate Sciences. Retrieved from https://scied.ucar.edu/sites/default/files/images/long-content-page/climate_literacy_brochure.pdf.

Yilmaz, K. (2011). The Cognitive Perspective on Learning: Its Theoretical Underpinnings and Implications for Classroom Practices. *The Clearing House: A Journal of Educational Strategies, Issues and Ideas*, *84*, 204-212. https://doi.org/10.1080/00098655.2011.568989

Yow, D. M. (2014). Teaching Introductory Weather and Climate Using Popular Movies. *Journal of Geoscience Education*, *62*, 118-125. https://doi.org/10.5408/13-014.1.

Zen, E-an. (1990). Science Literacy and Why it is Important. *Journal of Geological Education, 38,* 463-464. https://doi.org/10.5408/0022-1368-38.5.463

APPENDIX A

Surface Weather Map Contouring Exercise

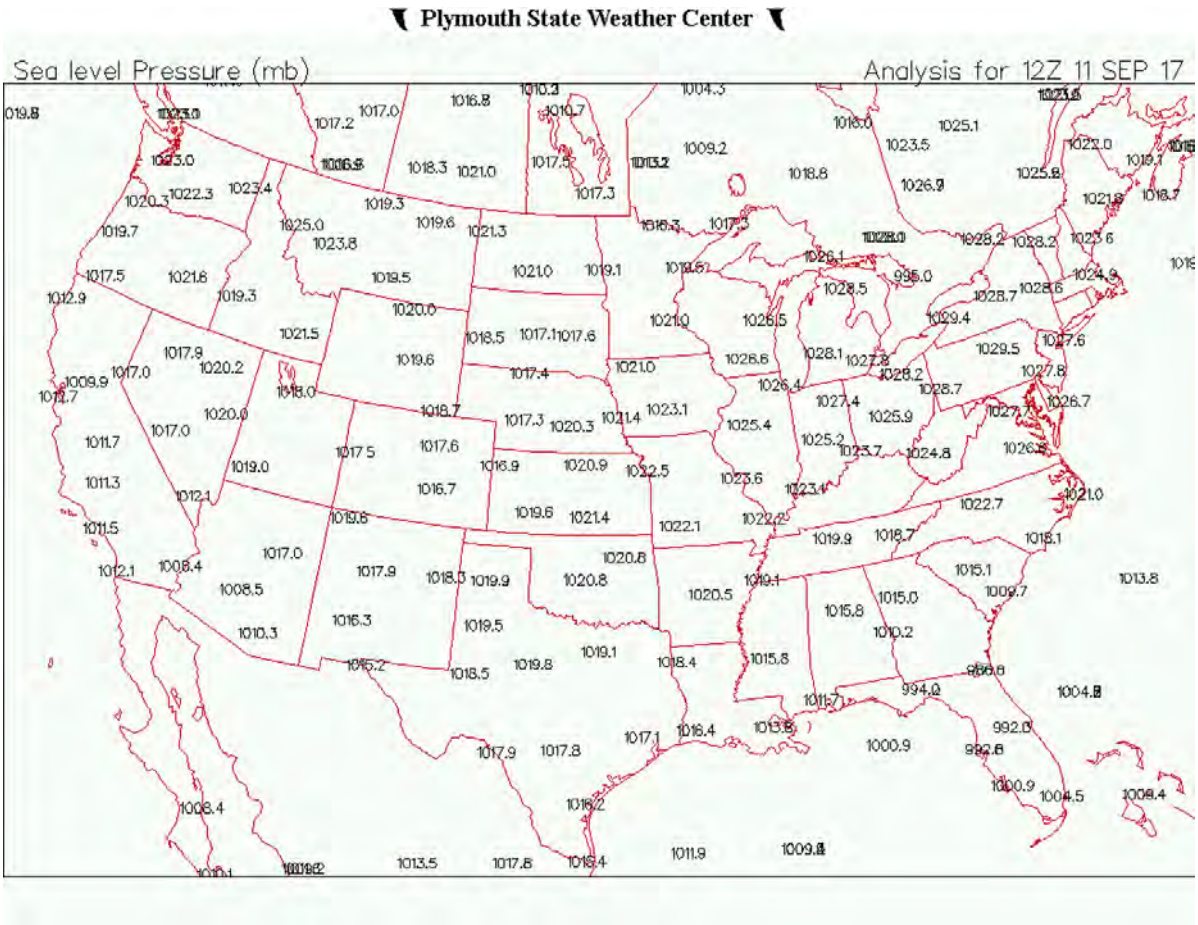
Given the surface map with only mean sea level pressure values provided (1200 UTC 11 September 2017), do the following:

a. Draw isobars ranging from 972 mb to 1024 mb every 4 mb. Draw smooth and continuous contours that are labeled appropriately. Refer to the attached appendix for specific details about how to best contour.

b. After completing the contouring exercise above, consider the following:
   1. Label the center of any low pressure regions with a red "L" symbol. What is the approximate minimum sea level pressure value associated with this cyclone?
   2. What type of feature are you observing where you labeled your map with the red "L"?
   3. Where do you expect precipitation to occur on your surface map? Explain why you chose the region(s) that you did.
   4. Where do you expect clear skies to occur on your surface map? Explain why you chose the region(s) that you did.

c. Compare/contrast your map with:
   - One of your classmate's hand-analyzed surface maps
   - The analyzed surface map from the Weather Prediction Center

      1. How does your analysis compare/contrast with your classmate's analysis? The Weather Prediction Center?
      2. Why do you think differences exist between the various analyses? Is more correct than the other? Why/why not?
      3. What advantages are there to hand-analyzing a surface weather map? Disadvantages?
      4. What advantages are there to using a computer to analyze a surface weather map? Disadvantages?



APPENDIX B

Example Activity 2: Cloud Identification Sounding Activity

a. Grab a camera, go outside, and take a picture of the clouds you see in the sky. Make sure there are clouds in your picture!

b. Record the date/time that you took the picture. Print out your picture to turn in with this assignment.

c. Go onto the following website and find the atmospheric sounding for your location. Use the time that most closely corresponds to when you took your picture: http://weather.uwyo.edu/upperair/sounding.html

   Make sure you select the following settings on the webpage:

   Region: North America
   Type of plot: GIF: Skew-T
   Year, Month: select appropriate year and month based on when you took your picture
   From, To: Pick the date/time that is closest to when you took your picture
   Station Number: Don't change

d. Print the sounding. You will need it for the rest of this problem.

   1. Circle all layers in which you observe clouds on your sounding. Based on our discussion of cloud types and soundings, as well as what you observe in your picture, label the cloud type that corresponds to each cloud layer.

   2. What differences and/or errors do you observe when comparing/contrasting your cloud picture with the sounding data? Explain.

APPENDIX C

Example Activity 3: Extra-tropical Cyclone Analysis Activity

Mini-Case Study – Surface Cyclone on 01 November 2018

You have been provided with the following forecast maps for 1800 UTC 01 November 2018:

- 300 mb geopotential height (black solid contours; units dam) and wind speed (wind barbs and fill pattern; units knots)
- Mean sea level pressure (black solid contours; units mb), thickness (dashed contours; units dam) and precipitation type/rate (fill pattern; see legends at bottom of figure)
- Mean sea level pressure (black solid contours; units mb), 2 m above ground level temperature (fill pattern; units °F) and 10 m above ground level wind (wind barbs; units knots) – note that the temperature and wind barbs represent surface temperature and wind properties

Given these figures, answer the following below:

a. On all maps, label the center of the mid-latitude surface cyclone with an "L".

b. On all maps, draw the cold and warm fronts associated with the surface cyclone in their appropriate locations.

c. In 1-2 paragraphs, given the 300 mb map provided, predict whether the surface cyclone over the next 12 hours will strengthen, weaken, or maintain the same minimum sea level pressure value. Defend your prediction in the context of the following:

   - Comment on the role that any jet streaks play that are nearby the surface cyclone.
   - Comment on the role that any trough/ridge flow pattern plays nearby the surface cyclone.
   - Your response should demonstrate how upper level processes induce upward/downward vertical motions and the role that these vertical motions play in affecting the strength of the surface cyclone.
   - You will be graded based on your ability to accurately analyze this data and describe the role that jet streaks/troughs/ridges play in the cyclone and its evolution over the next 12 hours.

APPENDIX D

Example Activity 4: Semester Project

We will spend the last few weeks of class working on a final project that will be presented to the class during final exams week. The final project will be used to assess your understanding of the material and application skills.

For this project you will work in groups of 5 – 7 to investigate a particular topic related to atmospheric sciences (e.g., investigate an interesting weather event of the past (e.g., tornado outbreak, blizzard, flooding event, etc.), or analyze a current weather event). While you will complete the data gathering and analysis as a group, each group member will be expected to compile a summary of the event (e.g., radar, satellite, etc.), and discuss the key meteorological features that contributed to the event (e.g., surface and upper level maps, soundings, etc.). This information will be compiled in a technical summary. As a group, you will also create a short presentation (Powerpoint, movie, etc.) that summarizes your results for the class. The grade for the final project will account for 20% of your overall course grade and will be based on the following:

Final project topic proposal:

Your group will create and submit a topic proposal. The proposal can be in paragraph or outline form, but it must provide a clear synopsis of the plans for your project. From your proposal, it should be clear to me what event type you plan to investigate. If you are researching a particular event, please specify the date and location of the event. Please lay out your plans for data/figures related to your analysis (e.g., radar, satellite, soundings, surface observations, upper level weather maps, etc.). The proposal is an opportunity for you to receive instructor feedback on your project plans.

Final project work days:

Two final project work days will take place in class; the entire class period will be used for groups to meet, gather and analyze data, create and plan oral presentations, and consult with your instructor.

In-class presentation:

Your group must create an oral presentation that will allow you to present your results to the class. You can use any format you wish, but it is expected that you will describe the significance of your case study, share and describe at least three images, and summarize your group's findings. Your presentation should be approximately 5 – 7 minutes.

Technical summary:

You will prepare a write-up of your findings that is 1 – 2 pages in length (excluding figures and references). More details on how to construct your technical summary will be provided later in the semester.